Quantitative quality control and background correction for two-colour microarray data

Matthew Edward Ritchie

Submitted in total fulfilment of the requirements of the degree of Doctor of Philosophy

December 2004

Department of Medical Biology
The University of Melbourne

The Walter and Eliza Hall Institute of Medical Research
Abstract

Two-colour microarrays are a popular tool for measuring relative gene expression between RNA populations for thousands of genes simultaneously. This thesis develops methods for assessing the quality and variability of data from such experiments and for incorporating these assessments into algorithms for discovering differential expression. The variability of microarray data depends not only on the quality of the arrays, but also on how they are processed and normalised. The intimate relationship between variability of expression log-ratios and the method used for background correcting the expression values is specifically explored. The performance of different estimators of the background level and various model-based processing methods, including a novel normal-exponential convolution model are compared in search of a ‘best’ alternative. The results indicate that the choice of method should be guided by the specific question of interest; the model-based methods give gene expression measures with low bias, and do very well at choosing differentially expressed genes, while subtracting low background estimates, or not background correcting the data produces low variance estimates which are the most biased, however perform best at choosing DE genes. All of these alternatives give better results than those obtained by the standard approach of subtracting high local background estimates from the foreground signal, which is not recommended. Another important low level analysis issue is that of data quality. Filtering microarray data according to different criteria to remove suspect arrays and/or spots from further analysis is widely used. This thesis investigates a more graduated approach to quality involving the use of array and spot quality weights in the linear model analysis of expression log-ratios. For array weights, a heteroscedastic linear model is fitted using a novel gene-by-gene update algorithm to calculate the array variance parameters. These variances are used as inverse weights in re-fits of the mean
model to gain more precise estimates of the gene expression coefficients. To assign spot weights, a predictive approach based on observed relationships between precision measured for genes from a large data set and various spot quality measures is taken. From the training data set, generalised weight functions are derived, and used to assign spot specific weights to data from other experiments. These weights are also used in the linear model analysis of the gene expression measures and can be used in conjunction with array weights, both of which are demonstrated to improve our power to detect differentially expressed genes in various data sets. The methods developed in this thesis aim to give fewer false positive results in a candidate list of differentially expressed genes, reducing the wastage of resources in follow up assays and assisting in the interpretation of the results.
Declaration

This is to certify that

(i) the thesis comprises only my original work towards the PhD except where indicated in the Preface,

(ii) due acknowledgement has been made in the text to all other material used,

(iii) the thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Signed,
Preface

The data sets used in this thesis were provided by the investigators listed below.

The K562 versus Pool experiment which features in chapter 3 was generated in David Erle’s lab at University of California, San Francisco. Yu Chuan Tai from the University of California, Berkeley analysed the Affymetrix data and mapped the probes to Unigene clusters on both platforms. All further analysis of this data set is my own work.

The Pax5 experiment used in chapters 3 and 5 was provided by Melissa Holmes from the Division of Immunology at the Walter and Eliza Hall Institute of Medical Research. All analyses of these data represents my own work.

The QC data set featured in chapters 4 and 5 was provided by Ryan van Laar, Dileepa Diyagama and Andrew Holloway from the Peter MacCallum Cancer Centre. All analyses of these data are my own work.

The METH experiment used in chapter 4 was provided by Jody Neilson and Alexander Dobrovic from the Peter MacCallum Cancer Centre. The analysis of this data set is my own work.

The DC experiment used in chapters 4 and 5 and the QDC experiment which features in chapter 5 were provided by Mireille Lahoud from the Division of Immunology at the Walter and Eliza Hall Institute of Medical Research. All analyses of these data sets are my own work.
Acknowledgments

Many people need to be thanked for supporting me throughout my PhD. First thanks go to my principal PhD supervisors, Gordon and Terry, for their guidance and direction on this project. I’m especially grateful to ‘Crazy’ Gordon who encouraged me to take the plunge and move from sunny Brisbane to the cooler climes of Melbourne, a decision which I haven’t regretted. Thank you Gordon for your patience, generosity with your time, and for being such a social supervisor on our many Friday visits to the pub and dinners around North Melbourne and Carlton. The next beer is definitely my shout! Thanks to Terry for allowing me to join his vibrant Bioinformatics research group, which has been a great learning environment. Your input on this thesis, and on my Wednesday seminar slides have made big differences - thanks for making the time. Cheers also for the references. To Sean, the silent third supervisor - I hope the end product is ok by you!

To my colleagues, both past and present, at the UQ Department of Mathematics and the WEHI; thanks for your friendship, enthusiasm, and constructive criticism of this work over the past 3 years and 10 months. Coming into the office is always a pleasure when you’re surrounded by such a lovely group of people. In particular, I must mention my old office buddy Nat, who taught me so much about microarray data analysis when I was starting out, and current officemate Alicia, for being ever-cheerful and helping proofread this thesis. Thanks also to the other members of the Bioinformatics group; James (for his non-stop enthusiasm and R help), Ken (purveyor of fine footy tips and advice on Affy analysis), Thommo (for laughs and GeneTraffic jam), Tim (fellow gym junkie and last minute printing expert!), Fred (comrade in free food, PhD deadlines and Wednesday seminar stresses), Asa (super fish, dining companion and computing guru), Lavinia (another super fish and tea-time shepherd), Melanie (for introducing me to Tiamo 2 and tortellini
tartufo!), Tony (Belgian beer drinking buddy), Jim (roomie and all-round nice guy), Keith (owner of our favourite ski lodge and midnight taxi service), Toby (for IT support in a second) and Cyclone Tracey (the super server). I promise a special stubby club meeting very soon to thank you all. To the international microarrayers, Jean and Henrik, thanks for useful discussions on this work and on other matters. To the rest of the division of Genetics and Bioinformatics, thanks for being so welcoming and friendly, and to Stephen and Melanie O in the AGRF for always happily answering any microarray technology queries, and for proofreading chapter 1. The ITS at the WEHI also deserve recognition for their efficiency and professionalism (especially Nick, Tri and John N). Without our computers, this thesis would be a lot shorter - thanks for keeping them running smoothly.

I'm also extremely grateful to the many collaborators who have generously provided their data for analysis and inclusion in this thesis. These people include, David Erle, Melissa, Andy, Ryan, Dileepa, Jody, Alex and Mireille. Thank you also to the early collaborators (Kirst, Wendy C, Luke, Wendy D, Lynn, Joëlle, Catherine and Toni to name but a few) from whom I learnt so much. I've really enjoyed hearing about the biology behind these varied experiments.

During my PhD candidature I have been very fortunate to have received financial support from the following sources: the Australian Postgraduate Award (APA), the Edith Moffat Scholarship fund, the WEHI and the GeneCRC.

To my friends and family in Melbourne and back home in Queensland; the long asked question “so when will you be finished” now has an answer! Thanks for all the support, both personal and financial, in particular from Mum, Dad and Shane.

To Rachel, my best friend in the world; thanks for being so patient over the past 3 years, especially during the stressful last few months. I promise we’ll have a real holiday very soon! With love always.

This thesis is dedicated to three very special people: my grandparents Helen Ritchie, Alma Keily and the late Charles Keily. These generous individuals have been great role models and advisors throughout my life, always full of enthusiasm and encouragement in whatever I was up to, particularly my academic endeavours. For their love and support I am eternally grateful.
# Table of Contents

Abstract  
Declaration  
Preface  
Acknowledgments  
Table of Contents  
List of Figures  
List of Tables  
List of Abbreviations  
List of Symbols  

1 Introduction to Microarray Technology and Data Analysis  
1.1 Measuring gene expression with microarrays  
1.2 The biological question  
1.3 Experimental design  
1.4 The two-colour microarray experiment  
  1.4.1 Array fabrication
1.4.2 Labelling and hybridisation .................................................. 7
1.4.3 Scanning ............................................................................. 8
1.5 Image analysis ......................................................................... 8
1.6 Pre-processing and normalisation ............................................. 9
  1.6.1 Background correction ....................................................... 9
  1.6.2 $M$ and $A$ ............................................................... 9
  1.6.3 Normalisation ................................................................... 10
1.7 Analysis .................................................................................. 10
  1.7.1 Summarising and testing for differential expression using
         linear models .................................................................. 10
  1.7.2 Software for the analysis of microarray data ...................... 13
1.8 Biological verification of results ............................................. 13
1.9 Thesis outline ......................................................................... 14

2 Data Sets and Controls .......................................................... 15
  2.1 K562 vs Pool experiment .................................................... 15
  2.2 Pax5 experiment ................................................................ 16
  2.3 QC data set ......................................................................... 18
  2.4 METH experiment ............................................................ 19
  2.5 DC experiment ................................................................... 21
  2.6 QDC experiment ............................................................... 21
  2.7 External control systems ................................................. 22
  2.8 Generalising the results .................................................... 25

3 Background Correction .......................................................... 26
  3.1 Introduction ......................................................................... 26
  3.2 Literature review ............................................................... 27
    3.2.1 Different background estimation and correction procedures .. 27
3.2.2 Approaches which deal with ‘fanning’ and missing values .......... 32
3.2.3 Observations from the analysis of Affymetrix data ................. 33
3.3 Comparison of background correction methods .................... 33
   3.3.1 Normal-exponential convolution model ..................... 34
   3.3.2 Details of other background processing methods .............. 37
   3.3.3 Using independent truth to compare the methods ............. 38
3.4 Results .................................. 38
   3.4.1 Assessment of variability ................................ 38
   3.4.2 Detecting differential expression ........................ 43
   3.4.3 Assessment of bias .................................. 45
3.5 Discussion ................................ 46

4 Array Quality Weights 52
   4.1 Literature review of current array quality methods ............. 52
   4.2 Array quality weights .................................. 56
      4.2.1 Mean and dispersion models .......................... 56
      4.2.2 The underlying model ................................ 57
      4.2.3 Tailoring the model for use on microarray data .......... 59
      4.2.4 A gene-by-gene update algorithm ...................... 62
      4.2.5 An example of the update algorithm .................... 64
      4.2.6 Simulated data .................................. 66
   4.3 Results .................................. 67
      4.3.1 Array weights for the QC data set ..................... 67
      4.3.2 Array weights for the METH experiment ................ 71
      4.3.3 Array weights for the DC experiment ................... 76
   4.4 Discussion .................................. 77
5 Spot Quality Weights

5.1 Literature review .......................................................... 79
  5.1.1 Methods based on spot specific measures ...................... 80
  5.1.2 Methods based on log-ratio reproducibility ................. 82

5.2 Spot quality weights ..................................................... 84
  5.2.1 Spot specific measures ............................................ 85
  5.2.2 Measuring reproducibility ....................................... 88
  5.2.3 Quality trends ...................................................... 88
  5.2.4 Spot quality weighting schemes ................................ 93
  5.2.5 Spot quality weights in the normalisation and gene-wise
       linear models ....................................................... 99

5.3 Results: Spot quality weights in the linear models ............. 101
  5.3.1 Spot weights applied to the QC data set ..................... 101
  5.3.2 Spot weights applied to other data sets ...................... 109
  5.3.3 Are spot and array quality weights complementary? .......... 114
  5.3.4 Summary .......................................................... 115

5.4 Discussion ............................................................... 117

Concluding Comments ....................................................... 120

Bibliography ................................................................. 122

Appendix A ................................................................. 133

Appendix B ................................................................. 136
List of Figures

1.1 The central dogma of molecular biology ..................................... 3
1.2 The microarray lifecycle .................................................. 4
1.3 Robotic microarray printer at the AGRF Microarray Facility ........... 6
1.4 Steps in a two-colour microarray experiment ............................. 7
1.5 MA plot for a typical microarray before and after print-tip loess normalisation 11

2.1 Design of the K562 vs Pool experiment .................................. 16
2.2 Venn diagram of probes represented on the Affymetrix GeneChip® arrays and the two-colour microarrays used in the K562 vs Pool experiment . . . . . 17
2.3 Design of the Pax5 experiment ........................................... 17
2.4 Representation of a print batch from the QC data set ...................... 19
2.5 Design for the QC data set ............................................... 19
2.6 Design of the METH experiment ......................................... 20
2.7 Design of the DC experiment ............................................. 22
2.8 Design of the QDC experiment ......................................... 22
2.9 MA plot with LMS controls before and after normalisation ............ 24

3.1 Foreground and background measures for Spot and GenePix on the same array ................................................................. 29
3.2 MA plot illustrating ‘fanning’ of log-ratios at low intensities ........... 31
3.3 Smoothed histograms of the red and green foreground intensities from three arrays from the QC data set ............................................. 35
3.4 MA plots for a given microarray obtained using different background correction methods .............................................. 40
3.5 Averaged MA plots obtained using different background correction methods and Affymetrix for the 7,547 common probes .......................................................... 42
3.6 Residual $\hat{\sigma}$ for different background correction methods ................................................... 43
3.7 ROC curves for different background correction methods on the false positive range 0 to 0.2 .................................................. 48
3.8 ROC curves for different background correction methods on the false positive range 0.2 to 1 .................................................. 49
3.9 $M$ values for LUS controls from slide 1 of the Pax5 experiment for different background correction methods .................................................. 50
3.10 Bias of $M$ values for LUS controls from the Pax5 experiment for different background correction methods ............................................. 51

4.1 Estimated versus actual array variance parameters from the simulated data set ............................................. 68
4.2 Estimated array variance parameter versus gene number for each array from the simulated data set ............................................. 69
4.3 Relative array weights for the QC data set ............................................. 71
4.4 An example of poorly printed grids from array HD45 from the QC data set ............................................. 72
4.5 $t$ statistics and SEs with and without array weights for the QC data set ............................................. 73
4.6 $t$ statistics for LMS controls with and without array weights from the QC data set ............................................. 73
4.7 Relative array weights for the METH experiment ............................................. 74
4.8 Background levels and % flagged spots for the arrays from the METH experiment ............................................. 75
4.9 Relative array weights for the DC experiment ............................................. 76
5.1 Close-up view of a spot ............................. 85
5.2 Example of reproducibility measure $e_{ij}$ ............................. 89
5.3 Reproducibility trends for spot area ............................. 90
5.4 Reproducibility trends for spot intensity ............................. 91
5.5 Reproducibility trends for average SNR ............................. 91
5.6 Reproducibility trends for average spot foreground variability ........ 92
5.7 Reproducibility trends for average spot foreground variability divided by area 93
5.8 Reproducibility stratified by GenePix flag ............................. 94
5.9 Weight functions for area, intensity and average foreground variability ... 95
5.10 Regression trees of spot specific measures for Spot and GenePix data . ... 97
5.11 $t$ statistics for LMS controls from the QC data set (Spot data) with and without spot weights ............................. 102
5.12 $Z$-scores for LMS controls from the QC data set (Spot data) with and without spot weights ............................. 102
5.13 $t$ statistics for genes from the QC data set (Spot data) with and without spot weights ............................. 103
5.14 Ratio of SEs of coefficients with:without spot weights for the QC data set (Spot data) ............................. 104
5.15 $M$ versus relative spot weights for D03 from the QC data set (Spot data) . 105
5.16 $M$ versus relative spot weights for D10 from the QC data set (Spot data) . 105
5.17 $M$ versus relative spot weights for U03 from the QC data set (Spot data) . 106
5.18 $M$ versus relative spot weights for U10 from the QC data set (Spot data) . 106
5.19 $M$ versus relative spot weights for DR from the QC data set (Spot data) . . 107
5.20 $t$ statistics for LMS controls from the QC data set (GenePix data) with and without spot weights ............................. 107
5.21 $Z$-scores for LMS controls from the QC data set (GenePix data) with and without spot weights ............................. 108
5.22 Ratio of SEs of coefficients with:without spot weights for the QC data set
(GenePix data) ........................................ 109

5.23 Z-scores for LUS controls from the Pax5 experiment (Spot data) with and
without spot weights .................................. 111

5.24 Z-scores for LUS controls from the Pax5 experiment (GenePix data) with
and without spot weights ................................ 112

5.25 Z-scores for LUS controls from the DC experiment with and without spot
weights .................................................... 113

5.26 Z-scores for LUS controls from the QDC experiment with and without spot
weights .................................................... 114

5.27 Plot of t statistics for LMS controls from the QC data set (Spot data) before
and after combining array and spot weights .................. 116

B.1 Reproducibility trends for perimeter and circularity .................. 136

B.2 Average SNR versus spot intensity ............................. 137

B.3 Ratio of SEs of coefficients with:without spot weights for the Pax5 experi-
ment (Spot data) ........................................ 137

B.4 Ratio of SEs of coefficients with:without spot weights for the Pax5 experi-
ment (GenePix data) ..................................... 138

B.5 Ratio of SEs of coefficients with:without spot weights for the DC experiment 138

B.6 Ratio of SEs of coefficients with:without spot weights for the QDC experiment 139
List of Tables

2.1 Summary of LMS controls ............................ 23
2.2 Summary of LUS controls ............................ 25

3.1 Summary of the background correction methods compared ................................. 34
3.2 Missing values by background correction method ............................................. 39
3.3 Correlations between gene expression measurements from Affymetrix and
    spotted arrays for the different background correction methods ..................... 41
3.4 Area under ROC curves ordered by size of background estimate used .............. 45

4.1 Sources of experimental variability ............................................................ 53
4.2 Number of DE genes for the METH experiment obtained using different
    weighting methods ............................................................................... 75

5.1 Summary of the spot specific measures available from Spot and GenePix
    image analysis software .......................................................................... 86
5.2 Branches of the regression tree for Spot data .............................................. 98
5.3 Branches of the regression tree for GenePix data ......................................... 99
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, C, T, G</td>
<td>Four nitrogen bases: adenine, cytosine, thymine, guanine</td>
</tr>
<tr>
<td>AGRF</td>
<td>Australian Genome Research Facility</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of variance</td>
</tr>
<tr>
<td>CC</td>
<td>Calibration controls (LUS)</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementary deoxyribonucleic acid</td>
</tr>
<tr>
<td>Cy3</td>
<td>Cyanine 3; green fluorescent dye</td>
</tr>
<tr>
<td>Cy5</td>
<td>Cyanine 5; red fluorescent dye</td>
</tr>
<tr>
<td>D03</td>
<td>Down-regulated 3-fold controls (LMS)</td>
</tr>
<tr>
<td>D03Low</td>
<td>Down-regulated 3-fold low intensity controls (LUS)</td>
</tr>
<tr>
<td>D03High</td>
<td>Down-regulated 3-fold high intensity controls (LUS)</td>
</tr>
<tr>
<td>D10</td>
<td>Down-regulated 10-fold controls (LMS)</td>
</tr>
<tr>
<td>D10Low</td>
<td>Down-regulated 10-fold low intensity controls (LUS)</td>
</tr>
<tr>
<td>D10High</td>
<td>Down-regulated 10-fold high intensity controls (LUS)</td>
</tr>
<tr>
<td>DC</td>
<td>Dendritic cells</td>
</tr>
<tr>
<td>DE</td>
<td>Differentially expressed</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>DR</td>
<td>Dynamic range controls (LMS)</td>
</tr>
<tr>
<td>GLM</td>
<td>Generalised linear model</td>
</tr>
<tr>
<td>IQR</td>
<td>Inter quartile range</td>
</tr>
<tr>
<td>limma</td>
<td>Linear models for microarray data (R package)</td>
</tr>
<tr>
<td>LMS</td>
<td>Lucidea Microarray ScoreCard</td>
</tr>
<tr>
<td>loess</td>
<td>Locally weighted regression</td>
</tr>
<tr>
<td>LUS</td>
<td>Lucidea Universal ScoreCard</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>MDS</td>
<td>Myelodysplastic syndromes</td>
</tr>
<tr>
<td>morph</td>
<td>Morphological opening background estimate</td>
</tr>
<tr>
<td>morph.close.open</td>
<td>Morphological closing followed by opening background estimate</td>
</tr>
<tr>
<td>MA</td>
<td>$M$ versus $A$</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum likelihood</td>
</tr>
<tr>
<td>MM</td>
<td>Mismatch probe (Affymetrix)</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean square error</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger ribonucleic acid</td>
</tr>
<tr>
<td>MSP</td>
<td>Microarray sample pool</td>
</tr>
<tr>
<td>PM</td>
<td>Perfect Match probe (Affymetrix)</td>
</tr>
<tr>
<td>PMCC</td>
<td>Peter MacCallum Cancer Centre</td>
</tr>
<tr>
<td>PMT</td>
<td>Photomultiplier tube</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal-to-noise ratio</td>
</tr>
<tr>
<td>QC</td>
<td>Quality control</td>
</tr>
<tr>
<td>REML</td>
<td>Residual/restricted maximum likelihood</td>
</tr>
<tr>
<td>RMA</td>
<td>Robust multichip average</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver operator characteristic</td>
</tr>
<tr>
<td>SD</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>SE</td>
<td>Standard error</td>
</tr>
<tr>
<td>SRG</td>
<td>Seeded region growing</td>
</tr>
<tr>
<td>tiff</td>
<td>Tag image format file</td>
</tr>
<tr>
<td>U</td>
<td>Uracil, a nitrogen base found in RNA (in place of thymine)</td>
</tr>
<tr>
<td>U03</td>
<td>Up-regulated 3-fold controls (LMS)</td>
</tr>
<tr>
<td>U03Low</td>
<td>Up-regulated 3-fold low intensity controls (LUS)</td>
</tr>
<tr>
<td>U03High</td>
<td>Up-regulated 3-fold high intensity controls (LUS)</td>
</tr>
<tr>
<td>U10</td>
<td>Up-regulated 10-fold controls (LMS)</td>
</tr>
<tr>
<td>U10Low</td>
<td>Up-regulated 10-fold low intensity controls (LUS)</td>
</tr>
<tr>
<td>U10High</td>
<td>Up-regulated 10-fold high intensity controls (LUS)</td>
</tr>
<tr>
<td>WEHI</td>
<td>Walter and Eliza Hall Institute of Medical Research</td>
</tr>
</tbody>
</table>
List of Symbols

$A$ Average log-intensity, $\frac{1}{2} \log_2(R \cdot G) = \frac{1}{2}(\log_2 R + \log_2 G)$

e$_{ij}$ Drop-one-out residual for spot $i$ on array $j$

$G$ Background corrected spot intensity; green channel

$G_b$ Raw background spot intensity measurement; green channel

$G_{b\text{IQR}}$ Inter quartile range of logged background pixels; green channel

$G_{b\text{SD}}$ Standard deviation of background pixels; green channel

$G_f$ Raw foreground spot intensity measurement; green channel

$G_{f\text{IQR}}$ Inter quartile range of logged foreground pixels; green channel

$G_{f\text{SD}}$ Standard deviation of foreground pixels; green channel

$M$ Log-ratio of expression, $\log_2(R/G) = \log_2 R - \log_2 G$

$R$ Background corrected spot intensity; red channel

$R_b$ Raw background spot intensity measurement; red channel

$R_{b\text{IQR}}$ Inter quartile range of logged background pixels; red channel

$R_{b\text{SD}}$ Standard deviation of background pixels; red channel

$R_f$ Raw foreground spot intensity measurement; red channel

$R_{f\text{IQR}}$ Inter quartile range of logged foreground pixels; red channel

$R_{f\text{SD}}$ Standard deviation of foreground pixels; red channel

$w_{ij}$ prior weight for spot $i$ on array $j$
Chapter 1

Introduction to Microarray Technology and Data Analysis

This chapter gives a brief introduction to the two-colour microarray technology used for profiling gene expression. The data that arises from these experiments along with the general pre-processing steps and basic statistical analysis which is applied in search of differentially expressed genes is also described.

1.1 Measuring gene expression with microarrays

The study of functional genomics has emerged as an important field following the availability of completed genome sequences for many organisms (225 as at 20 October 2004 according to http://www.genomesonline.org/). Understanding the function of each gene and the pathways and networks it influences is an enormous challenge, and high-throughput technologies such as microarrays have emerged as a popular tool for providing a first clue as to which genes are important in different conditions.

In eukaryotes, the vast majority of genes are encoded in the double stranded DNA found in the nucleus of most cells. According to the central dogma of molecular biology, genes are transcribed into single stranded mRNA before exiting the nucleus where they are used as a template for protein synthesis (refer to figure 1.1). Whilst there are many determinants of the amount of protein in a cell, such as post-translational modifications,
the transcriptional regulation of genes is a key process worthy of careful study. Gene expression experiments measure the amount of mRNA produced during transcription by exploiting the complementary base-pairing of DNA (A pairs with T and G pairs with C, or in the case of RNA, A pairs with U). Two complementary single-stranded sequences can combine, or ‘hybridise’ to form a double helix. On a microarray, hybridisation occurs between fluorescently labelled single-stranded sequences in solution, referred to as the ‘target’ and immobilised single-stranded sequences on the array surface known as ‘probes’. The high density arrangement of probes at precise locations on the microarray enables gene expression measurements to be collected for many thousands of genes simultaneously.

Microarrays are not restricted to measuring gene expression. For instance, cytogenetic studies make use of CGH (comparative genome hybridisation) arrays to detect chromosomal deletions or rearrangements, and for genotyping applications, SNP (Single Nucleotide Polymorphisms) chips have become available to give dense marker coverage across a genome. All of these technologies utilise genomic DNA.

Gene expression microarrays are available in different formats. For example, Affymetrix GeneChip® arrays are hybridised with one labelled sample, and measure absolute expression on each array with multiple in situ synthesised oligonucleotide (25-mer) probes per gene. Two-colour microarrays measure relative gene expression between two different target samples, with one population of mRNA labelled with Cy3 (green) dye and a second sample labelled with Cy5 (red) dye before being competitively hybridised to the probes on the slide. Each gene is typically represented by one probe, which may be either single stranded oligonucleotide sequences (usually of a specific length, e.g., 60-mer) or cDNA sequences (of variable length), which are ‘spotted’ on the array using a robotic printer. An image for both red and green channels is acquired by a scanner and the intensity of fluorescence (i.e. hybridisation) in each channel is quantified for each spot. In principle, the amount of hybridisation is proportional to the abundance of the particular mRNA species in the target sample, and the ratio between the two intensities for a given spot is a measure of the ‘differential expression’ for the particular gene between the two samples. A ‘differentially expressed’ gene is more highly transcribed in the red labelled sample compared to the green labelled sample, or vice versa. With Affymetrix technology, two chips are required to measure the difference in expression between two samples.
Figure 1.1: The central dogma of molecular biology; DNA is transcribed into RNA which is translated into protein outside the nucleus. Image courtesy of Natalie Thorne (reproduced from Access Excellence Graphics gallery, http://www.accessexcellence.org/RC/VL/GG/central.html).
The stages involved in a microarray experiment are shown in figure 1.2. In the following sections, the steps in this process will be described in more detail for two-colour spotted microarrays.

### 1.2 The biological question

Two-colour microarrays are commonly used to investigate which genes are differentially expressed (DE) between two samples. Some examples of specific questions of interest to researchers are given in chapter 2, and include the identification of novel genes expressed on the surface of different subtypes of dendritic cells (DC), or determining the target genes regulated by Pax5, a gene which is essential for B-cell-lineage commitment.

In other studies, the broad pattern of gene expression may be of interest for clustering ‘like’ samples or discriminating between disease states using the gene expression measures. Many large studies such as Alon et al. [1999] and Alizadeh et al. [2000] have profiled the gene expression in normal and cancerous tissue samples in patients and used these profiles...
to group tumour samples into clinical subtypes, and as predictors of patient survival or response to therapy. The ultimate aim of such investigations is to develop diagnostic tests which would allow more targeted treatment and a better prognosis for cancer patients.

1.3 Experimental design

Design considerations can include which probes and controls to include and the layout of spots on the array, but more commonly it refers to the allocation of mRNA samples to arrays. Whether to pool mRNA samples, compare samples directly or indirectly using a common reference and how many times to replicate the experiment are issues which must be considered to ensure that the primary question of interest can be answered. Smyth et al. [2003] includes a good summary of the design issues, and some examples of different designs used in real experiments are given in chapter 2. In this thesis, the term ‘designed experiment’ will be used to refer to experiments with some level of replication.

1.4 The two-colour microarray experiment

A good overview of the subjects covered in sections 1.4.1 to 1.4.3 is given in Eisen and Brown [1999] or chapter 1 of Schena [1999].

1.4.1 Array fabrication

A collection of probes (or clones), known as a ‘library’ can be representative of all genes in the genome of the model organism, or be a specialised custom set of probes (e.g. the B-cell specific library enriched for genes involved in B-cell development or the DC specific clones in chapter 2). The probes are stored in 96- or 384-well plates ready for printing.

In addition to the experimental probes, control probes are also included on arrays to add ‘truth’ to each microarray. Housekeeping genes which are ubiquitously expressed can be used as positive controls for expression, while buffer solution, water or genes from other organisms act as negative controls which should show no hybridisation. MSP (Microarray Sample Pool) controls which consist of a pool of all, or a large proportion of the genes printed on an array can be diluted out in a titration series to provide useful normalisation.
controls (see Yang et al. [2002c]). Commercial control systems such as Lucidea Microarray ScoreCard (LMS) or Lucidea Universal ScoreCard (LUS, see section 2.7 for a description) have a mix of negative, positive and spiked-in DE and non-DE controls.

For spotted arrays, probe is deposited on each array using a robotic printer. Figure 1.3 shows the printing robot used at the AGRF Microarray Facility in Melbourne, which produced many of the arrays analysed in this thesis. It consists of a print-head with a number of pins, which dip into the plates of probes, and then deposit small amounts of probe at regularly spaced intervals on contact with the poly-L-lysine-coated glass microscope slides. Typically a large number of arrays (100) are printed in parallel, giving rise to the term ‘print batch’. After printing is complete, several post-processing steps such as UV cross-linking of printed DNA to the slide surface and blocking to limit the non-specific binding of DNA to the array are performed after which the array is ready for hybridisation.

Figure 1.3: The robotic microarray printer at the AGRF Microarray Facility (Melbourne, Australia). The printer-head with 8 pins is shown ready to dip into a 384-well plate before spotting the probe onto glass slides. Photo courtesy of Communication Services at the WEHI.
Figure 1.4 gives a schematic representation of the steps involved in hybridising a microarray. Firstly, the total cellular RNA is extracted from cells or a tissue sample, from which the mRNA is purified. The mRNA is used as a template for the production of cDNA which is fluorescently labelled with either Cy3 or Cy5 dye. The two samples are mixed and allowed to hybridise to the probes on the array for an appropriate incubation period. The slide is then washed to remove excess sample which has not hybridised to the spots.

1.4.2 Labelling and hybridisation

Figure 1.4 gives a schematic representation of the steps involved in hybridising a microarray. Firstly, the total cellular RNA is extracted from cells or a tissue sample, from which the mRNA is purified. The mRNA is used as a template for the production of cDNA which is fluorescently labelled with either Cy3 or Cy5 dye. The two samples are mixed and allowed to hybridise to the probes on the array for an appropriate incubation period. The slide is then washed to remove excess sample which has not hybridised to the spots.
1.4.3 Scanning

The scanner is an optical device used to measure the fluorescence emitted by the Cy3 and Cy5 dyes present on the array after excitation using a light source at appropriate wavelengths. An image of the fluorescence emitted for each channel is stored as a 16-bit tag image format file (tiff). For most arrays analysed in this thesis, a GenePix 4000B scanner (Axon Instruments, Union City, CA) which uses lasers at wavelengths 532 nm and 635 nm to excite the Cy3 and Cy5 dyes, respectively, acquired the images. The settings of the scanner, such as the laser power (10%, 33% and 100%) and the gain of the photomultiplier tubes (PMT, which are the detectors of emission) can be adjusted. The PMT voltage should be adjusted to ensure minimal saturation of the pixels, and equivalent intensities between the channels.

1.5 Image analysis

The raw tiffs are quantitated in a process known as image analysis, which can be divided into three main tasks (Yang et al. [2002b]). After the green tiff is overlaid with the red tiff, regular grids are aligned over the spots on each array to mark the coordinates of each spot; this is known as ‘addressing’ or ‘gridding’. Secondly, ‘segmentation’ of the images into foreground and background pixels is performed. In this step, masks are placed over each spot and pixels which fall within the mask area are classified as foreground, and the remaining pixels as background. Different image analysis software uses slightly different segmentation algorithms. Spot (Buckley [2000]) implements seeded region growing (SRG, Adams and Bischof [1994]) where the mask is grown from an initial seed and may have an irregular shape, whereas GenePix Pro version 4 (GenePix Pro [2001]) uses adaptive circle segmentation, which assumes spots are circular but may vary in diameter. Finally, the intensities for the foreground and background are extracted and summarised for each spot. These values, along with other quality measures are stored in a file. A separate file which contains identification information for each spot, sometimes known as a GAL (GenePix Array List) file may also be necessary for further analysis.
1.6 Pre-processing and normalisation

Before assessing which genes are DE in an experiment or applying clustering or discrimination techniques, various pre-processing and normalisation steps are recommended (see Smyth et al. [2003]). These are explained briefly in the following sections.

1.6.1 Background correction

The process of background correction involves adjusting the measured foreground intensity for background fluorescence. The local background measure for each channel \((R_b, G_b)\) is usually subtracted from the foreground measure \((R_f, G_f)\) to give an estimate of the true signal arising from hybridisation \((R = R_f - R_b, G = G_f - G_b)\). In chapter 3, some of the problems arising from background subtraction are discussed, and it is shown that more sophisticated background correction methods can be advantageous.

1.6.2 \(M\) and \(A\)

Dudoit et al. [2002] found it useful to convert the \(\log_2 R\) and \(\log_2 G\) values into \(M\) (for ‘minus’) and \(A\) (for ‘add’) values. Here

\[
M = \log_2(R/G) = \log_2 R - \log_2 G
\]

measures the relative difference in the amount of hybridisation for each spot, and

\[
A = \frac{1}{2} \log_2(R.G) = \frac{1}{2}(\log_2 R + \log_2 G)
\]

measures the average log-intensity of a spot. A gene which is expressed equally in the two samples will have \(M = 0\), and a gene which has a 2-fold difference in the red sample compared with the green will have \(M = 1\). For the \(A\) values, an increase of 1 unit represents a 2-fold increase in the average brightness of a spot.

An MA plot, where the \(M\) values are plotted against the \(A\) values for a given slide is a very useful diagnostic plot for microarrays, allowing detection of spatial artifacts and the need for normalisation as well as the general success of the hybridisation (arrays with
spot intensities spanning the full dynamic A range are generally better than arrays with a small A range). Figure 1.5 shows an example MA plot.

1.6.3 Normalisation

The process of normalisation refers to the removal of systematic biases from the log-ratios which are due to technical rather than biological sources. Normalisation can be applied to data from one array, or between arrays. Popular within-array methods are discussed in Yang et al. [2001c], and include global normalisation, which adjusts all M values by a constant amount (usually the mean or median M value from a given slide), and global intensity dependent normalisation, which removes A-dependent bias from the M values. Normalisation of microarray data using intensity (A) based print-tip loess by subtracting the corresponding value for the print-tip loess curve (\( M_{\text{norm}} = M - \text{loess}_p(A) \)), where the loess curve is a function of A for the pth print-tip group) is recommended in Smyth and Speed [2003]. This method is intended to remove systematic red-green dye-biases and some of the spatial effects from the log-ratios (see figure 1.5). Between slide normalisation to remove scale differences between the arrays in an experiment may also be useful (see Smyth and Speed [2003]). This is done with a scaling factor which ensures each array has the same median absolute deviation of M.

1.7 Analysis

1.7.1 Summarising and testing for differential expression using linear models

For designed microarray experiments with some level of replication, linear models can be used to summarise the log-ratios for each gene before testing for differential expression. Smyth [2004] gives an introduction to this topic and the general model with slightly different notation will be repeated here. For data from a two-colour microarray experiment, assume there is a response vector of log-ratios \( y_i^T = (M_{i1}, \ldots, M_{iJ}) \) for the ith gene measured on J arrays, and assume the log-ratios have been normalised appropriately and that missing values may be present. Assume that
Figure 1.5: MA plot for a typical microarray (a) before and (b) after print-tip loess normalisation. The different coloured curves represent the intensity based loess fits through the $M$ values from different print-tip groups. After normalisation, the $M$ values are centered around 0 at all intensities.
\[ E(y_i) = X_{\text{gene}}\beta_i \]  

where \( X_{\text{gene}} \) is the design matrix of full column rank, \( \beta_i \) is a coefficient vector and 

\[ \text{var}(y_i) = W_i \sigma_i^2 \]

where \( W_i \) is a known non-negative definite weight matrix. In chapter 4, explicit modelling of the variance of \( y_i \) in an array dependent manner is explored, while in chapter 5, the assignment of prior spot weights using different quality measures obtained during image analysis is investigated.

Having fitted the linear model, estimated coefficient vectors \( \hat{\beta}_i \), and variance estimators \( s_i^2 \) are obtained along with variance-covariance matrices 

\[ \text{var}(\hat{\beta}_i) = V_i s_i^2 \]

where \( V_i \) is a positive definite matrix. In this thesis, the coefficients are estimated using weighted least squares or generalised weighted least squares. To test the \( H_0 : \beta_{ik} = 0 \), (where there are \( k = 1, \ldots, K \) coefficients) under the assumption of normality, the test statistics

\[ t_{ik} = \frac{\hat{\beta}_{ik}}{\text{SE}(\hat{\beta}_{ik})} \]

are calculated. Here \( \text{SE}(\hat{\beta}_{ik}) = s_i \sqrt{v_{ik}} \), where \( v_{ik} \) are the diagonal elements of \( V_i \). In cases where a moderated \( t \) statistic is to be used, \( s_i \) is replaced with \( \tilde{s}_i \) (such as the posterior variances of Smyth [2004]). Ranking the genes using the \( t_{ik} \) statistics or the resulting \( p \)-values, which are often adjusted for multiple testing of thousands of genes simultaneously, can be used as evidence for differential expression. Genes with large \( |t_{ik}| \) (or equivalently small adjusted \( p \)-values) are likely to be truly DE. In cases where the contrasts of the coefficients are of interest, these are defined as \( \phi_i = C^T \beta_i \), where \( C \) is the contrast matrix, and 

\[ \text{var}(\hat{\phi}_i) = C^T V_i C s_i^2. \]
For the calculation of the $t$ statistics for the contrasts, $\hat{\beta}_{ik}$ is replaced with $\hat{\phi}_{ik}$ and $v_{ik}$ are now the diagonal elements of $C^T V_i C$.

1.7.2 Software for the analysis of microarray data

There are many open source (see Dudoit et al. [2003] for a selected review) and commercial software packages available for microarray data analysis. For two-colour microarray data, the limma (Smyth [2004]) package was used extensively in this thesis. limma facilitates reading in of the data, the generation of diagnostic plots, the fitting of linear models and testing for differential expression. This software is implemented in the statistical programming language R (Ihaka and Gentleman [1996]) and is available as part of the Bioconductor project (Gentleman et al. [2004]). A standard laptop with a Pentium 2.0GHz processor and 1024Mb of RAM was used to perform all analyses in this thesis.

1.8 Biological verification of results

To confirm that the changes observed in a microarray study are ‘real’, other techniques of quantitating mRNA levels such as Northern blot or quantitative reverse transcription-polymerase chain reaction (qRT-PCR) are used. These methods are generally low-throughput and this independent confirmation is only usually followed up for a few genes. Further cellular investigations to hone in on the function and processes that these genes are involved in are also important.

Using the available gene annotations for the DE genes, common functional themes for well characterised genes can be examined using their gene ontology (GO) terms (http://www.geneontology.org/). Beißbarth and Speed [2004] provide software (GOstat, http://gostat.wehi.edu.au/) which tests for over-represented GO terms in a list of DE genes compared to a reference set (typically all genes on the array). Such an analysis may provide a clue as to which biological processes are important in the system being studied.

For libraries of unknown clones, sequencing of the clones followed by alignment against known sequences can be used to identify these genes. To save time and expense, sequencing
is often only carried out on the candidate DE clones from an experiment which may be subject to further study.

1.9 Thesis outline

Methods that deal with the underlying variability of microarray data, arising from the pre-processing step of background correction, or the varying quality of data measured on different arrays in different spots, are developed in this thesis. In chapter 2, the various experimental data sets analysed and used to test the methods proposed will be described. The issue of background correction of two-colour microarray data will be addressed in chapter 3, with various approaches including model-based correction procedures compared to find a ‘best’ alternative for use in the remainder of the thesis.

The issue of data quality will be addressed in chapters 4 and 5, with the former describing the use of array weights which can be empirically derived using the data from designed experiments, and the latter outlining the use of prior spot weights derived from spot specific measures. These weights are used in the linear model analysis of the log-ratios to improve the detection of DE genes. Some final remarks on the topics discussed in this thesis are given in the Concluding Comments section.
Chapter 2

Data Sets and Controls

Several microarray data sets are used to demonstrate and assess the methods proposed in this thesis. All of the data sets considered have some form of independent truth available, either in the form of gene expression measures from another microarray platform, or via non-DE and DE spike-in controls. Such truth provides a basis for comparing different analysis alternatives, where methods which recover the known results more closely are preferred. These data sets and the external control systems used on many of the microarrays are briefly described in this chapter.

To summarise the experimental design used to allocate mRNA samples to arrays, the directed arrow notation of Kerr and Churchill [2001a], which points from the sample labelled green towards the sample labelled red will be used throughout this chapter.

2.1 K562 vs Pool experiment

This data set is published in Barczak et al. [2003], and compares the gene expression measurements across two microarray platforms. Two-colour arrays printed with the Operon Human Genome Oligo Set Version 1, which consists of 13,971 oligo probes (mostly 70-mers) were produced at the Department of Medicine, University of California, San Francisco. A total of 6 arrays were hybridised with mRNA isolated from K562 erythroleukemia cells in one channel and pooled mRNA from 10 different cell lines (Universal Human Reference RNA, Stratagene) in the other channel, as indicated in figure 2.1. The arrays were scanned
using a GenePix 4000B scanner, and image analysed using Spot 2.0 and GenePix Pro 3.0 software. The log-ratios from these arrays were normalised using intensity dependent print-tip loess.

Figure 2.1: Design of the K562 versus Pool experiment.

In addition, Affymetrix U95Av2 GeneChip® arrays with 25mer probes (11-20 perfect match (PM) probes per gene) synthesised in situ were also hybridised: 2 chips with K562 mRNA and 3 with pool mRNA. The PM data was background corrected, quantile normalised between chips and summarised using the Robust Multichip Average method of Bolstad et al. [2003] using the affy package (Gautier et al. [2004]) to yield expression measures for 12,558 genes. This analysis was performed by Yu Chuan Tai from the University of California, Berkeley.

The expression measures for each of the 12,558 genes were summarised between chips using a two-coefficient linear model ($\beta_i^T = (\beta_{K562}^T, \beta_{Pool}^T)$), and contrasts of the coefficients were calculated to give log-ratios equivalent to those obtained from the two-colour arrays. The linear model analysis was performed using limma (Smyth [2004]).

To compare results between platforms, the probes were matched using Unigene clusters. The Genbank accession numbers for each probe were mapped to Unigene clusters using the SOURCE database (http://source.stanford.edu). A total of 7,547 unique genes were in common to the Operon and Affymetrix arrays and these were used for comparison (see figure 2.2). This mapping was done by Yu Chuan Tai. This data set will be used in chapter 3 of this thesis to compare different background correction methods.

2.2 Pax5 experiment

The Pax5 gene encodes the B-cell-specific activator protein (BSAP), a transcription factor which is essential for lineage commitment of pro-B cells. To discover novel target genes regulated by Pax5, a microarray experiment comparing pro-B cells from Pax5 knock-out
Figure 2.2: Venn diagram of probes represented on the Affymetrix GeneChip® arrays and those on the two-colour microarrays printed with the Operon Human Genome Oligo Set Version 1. 7,547 probes mapped to unique Unigene clusters, and these were regarded as genes common to both platforms.

(-/-) and wild-type (+/+) mice was designed. Pro-B cells isolated from mouse spleen were extracted and cell lines were generated. mRNA from the wild-type cell line (Wt) and 3 knock-out cell lines (Mt1, Mt2 and Mt3) were compared on 9 microarrays according to the experimental design in figure 2.3.

The arrays were printed with a B cell specific cDNA library made up of 12,863 clones and the LUS control series described in section 2.7. Spots were printed in duplicate side-by-side on glass slides at the AGRF in Melbourne.

Figure 2.3: Design of the Pax5 experiment.

The hybridised arrays were scanned using a GenePix 4000B scanner, and image analysed using Spot 2.0 and GenePix Pro 4.0. The log-ratios from these arrays were normalised separately for the genes (using intensity dependent print-tip loess) and controls
(using global loess) and summarised using linear models, with appropriate design matrices \(X_{\text{gene}}\) to estimate the coefficients \(\beta_{Wt-Mt1}, \beta_{Wt-Mt2}, \beta_{Wt-Mt3}\) for each gene and \(\beta_{\text{LUS}}\) for each control. Since each probe was printed in duplicate, the linear models were fitted using generalised least squares with an estimated value for the correlation between duplicate spots, which has the effect of increasing the degrees of freedom available to estimate the gene-wise variances compared with averaging the duplicate log-ratios (Smyth et al. [2005]). This data set will be referred to as the Pax5 experiment and is used in chapters 3 and 5 of this thesis.

2.3 QC data set

This data set was provided by Andrew Holloway, Ryan Van Laar and Dileepa Diyagama from the Microarray core facility at the PMCC. The microarrays were printed with a Human 10.5K cDNA clone set, made up of 8K sequence verified clones corresponding to known genes from the I.M.A.G.E. consortium (distributed by Research Genetics/Invitrogen, Carlsbad, CA) combined with a further 2.5K human EST’s from the PMCC and the LMS control series described in section 2.7. The arrays, printed in 2001-2002, were taken from the beginning, middle and end of 22 print batches (refer to figure 2.4). Each microarray was assigned a label using the batch name (HA, HB, HC, etc.) followed by the array number \((j = 1, \ldots, 100)\) within a print batch; so the 15th array from batch HA is labelled HA15.

The 4-6 arrays from each batch were hybridised with Cy3 (green) labelled mRNA from the MCF-7 breast cancer cell line and Cy5 (red) labelled mRNA from the Jurkat T cell leukemia cell line (refer to figure 2.5) as a quality control step in the printing process. A total of 111 arrays were available. The earlier slides made use of indirect labelling protocol and reagents, while the later arrays moved to a direct labelling method. Arrays from a given batch may be regarded as technical replicates, insofar as the mRNA was derived from a single extraction, which should reduce the variability within batches.

Once hybridised, the slides were scanned using the scanner available at the PMCC at the time. The scanners included a ScanArray 5000, ScanArray 4000 (Packard BioChip Technologies, Billerica, MA) and an Agilent DNA Microarray Scanner (Agilent, Palo Alto,
Figure 2.4: Representation of a print batch from the QC data set. Each rectangle in this diagram represents a microarray. After printing was completed, between 4 and 6 arrays were taken from the beginning, middle and end (black rectangles) of the batch and hybridised as a quality control step, to ensure the spots had been printed properly and the 384-well plates had been inserted the correct way in the printer. Arrays from 22 slide print batches manufactured during 2001-2002 at the PMCC made up the QC data set.

CA). The tiff images from these arrays were image analysed using Spot 2.0 and GenePix Pro 4.0.

The log-ratios for the genes and controls were normalised separately and summarised between arrays by gene-wise linear models which estimated the coefficient $\beta_{\text{Jurkat}-\text{MCF7}}$ for each spot. This collection of arrays will be referred to as the QC data set, and will be used in chapters 4 and 5 to test the array and spot weighting methods proposed.

Figure 2.5: Experimental design for the QC data set. 111 replicate microarrays comparing mRNA from Jurkat and MCF-7 cell lines were hybridised.

2.4 METH experiment

This experiment, courtesy of Jody Neilson and Alexander Dobrovic from the PMCC also used Human 10.5K cDNA microarrays. The study aimed at identifying novel methylated markers in myelodysplastic syndromes (MDS) using the cell line KG1a. A known inhibitor of DNA methylation, 5-azacytidine (Azacitidine) was added to KG1a cells in varying doses
(0mM, or untreated, 1mM and 3mM) over time. Both direct and indirect comparisons between the 1mM and 3mM treatments and the 0mM treatment were made on a total of 10 arrays (refer to figure 2.6 for the experimental design). The panel reference mRNA consisted of a pool of mRNAs from 11 cancer cell lines as described in Pollack [2002].

The arrays were scanned using a GenePix 4000B scanner, and GenePix Pro 4.0 was used for the image analysis of the tiff images from these microarrays.

The data were normalised using intensity based print-tip loess, and the log-ratios were summarised by linear models, using an appropriate design matrix ($X_{gene}$) to estimate the coefficients ($\beta_{0 - Panel}, \beta_{1 - Panel}, \beta_{3 - Panel}$) for each gene. Contrasts of the coefficients gave the gene expression effects of main interest, ($\beta_{1 - 0}, \beta_{3 - 0}$). This data set will be referred to as the METH experiment and is used in chapter 4.

Figure 2.6: Design of the METH experiment, which compared 3 mRNA sources of interest (0mM, 1mM and 3mM) directly on the first 4 arrays and indirectly via a Panel reference on a further 6 arrays. The arrays were hybridised in pairs on different days (i.e. array 1 was hybridised on the same day as array 2, array 3 was hybridised on the same day as array 4, and so on).
2.5 DC experiment

The identification of novel surface markers on dendritic cells (DC) was the aim of an experiment conducted by Mireille Lahoud from the Division of Immunology at the WEHI. The experiment compared mRNA from 3 activated splenic dendritic cell subsets: CD4$^{+}$8$^{-}$, CD4$^{-}$8$^{+}$ and CD4$^{-}$8$^{-}$ which were isolated from mouse. A total of 6 microarrays were hybridised in 3 dye-swap pairs according to the design in figure 2.7. Genes which were DE in one subtype and not in others were of interest, especially those which encode for cell surface proteins.

Each array was printed with a custom cDNA library containing approximately 10,000 clones. This library was generated from a pool of mouse DC of all subtypes which had been normalised and had genes in common with fibroblasts subtracted. The clones, along with the LUS control series described in section 2.7 were printed on arrays in duplicate side by side at the AGRF in Melbourne. After hybridisation, the 6 arrays were scanned using a GenePix 4000B scanner, and the tiff images were quantitated using GenePix Pro 4.0.

The data were normalised by print-tip loess for the genes and global loess for the controls, and the log-ratios were summarised by linear models, using appropriate design matrices ($X_{\text{gene}}$) to estimate the coefficients ($\beta_{(CD4^{-}8^{+})-(CD4^{+}8^{-})}$, $\beta_{(CD4^{-}8^{+})-(CD4^{-}8^{-})}$) for each gene and $\beta_{LUS}$ for each control. Since each probe was printed in duplicate, a common correlation between duplicate spots was used in the regressions (see Smyth et al. [2005]) instead of averaging the log-ratios from duplicate spots within an array. This data set will be referred to as the DC experiment, and is used in chapter 5.

2.6 QDC experiment

This experiment is also courtesy of Mireille Lahoud from the Division of Immunology at the WEHI, and was carried out prior to the DC experiment as a pilot study to test reagents and the hybridisation protocol. In this experiment, mRNA from quiescent (resting) splenic DC from mouse of the CD4$^{+}$8$^{-}$ and CD4$^{-}$8$^{+}$ subpopulations were compared directly on 2 arrays, with one dye-swap (refer to figure 2.8). The arrays were printed with the DC
specific cDNA library and LUS controls, as described in section 2.5. After hybridisation, the 2 arrays were scanned using a GenePix 4000B scanner, and the tiff images were analysed using GenePix Pro 4.0.

The data were normalised appropriately, and the log-ratios for each gene were summarised by linear models to estimate the coefficient $\beta_{(\text{CD}4^-8^+)-(\text{CD}4^+8^-)}$ for each gene and $\beta_{\text{LUS}}$ for each control. The duplicate correlation approach of Smyth et al. [2005] was used in the regressions instead of averaging the log-ratios from duplicate spots within the same array for the reasons noted previously (see section 2.2). This data set will be referred to as the QDC experiment in chapter 5 of this thesis.

### 2.7 External control systems

Lucidea Microarray ScoreCard (LMS) controls described in Samartzidou et al. [2001] were printed in the last row of each print-tip group on arrays from the QC and METH experiments, and spiked into the mRNA mixes prior to labelling for most of the arrays. The DE (D03, D10, U03, U10) and non-DE (DR) control genes have known log-ratios, as shown in

---

**Figure 2.7: Design of the DC experiment.**

**Figure 2.8: Design of the QDC experiment.**
table 2.1, and provide independent ‘truth’ of gene expression which can be used to assess a method’s ability to detect known differential expression. The number of each control spotted on these arrays is also indicated in table 2.1.

Table 2.1: Summary of LMS controls.

<table>
<thead>
<tr>
<th>Control</th>
<th>Abbrev.</th>
<th>M</th>
<th>A</th>
<th>Spots per slide</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up-regulated 3-fold</td>
<td>U03</td>
<td>(\log_2(3))</td>
<td>med/high</td>
<td>12</td>
</tr>
<tr>
<td>Up-regulated 10-fold</td>
<td>U10</td>
<td>(\log_2(10))</td>
<td>med/high</td>
<td>12</td>
</tr>
<tr>
<td>Down-regulated 3-fold</td>
<td>D03</td>
<td>(-\log_2(3))</td>
<td>med/high</td>
<td>12</td>
</tr>
<tr>
<td>Down-regulated 10-fold</td>
<td>D10</td>
<td>(-\log_2(10))</td>
<td>med/high</td>
<td>12</td>
</tr>
<tr>
<td>Dynamic Range controls</td>
<td>DR</td>
<td>0</td>
<td>low/med/high</td>
<td>72</td>
</tr>
</tbody>
</table>

The Lucidea Universal ScoreCard (LUS) controls, described in the Lucidea Universal ScoreCard User’s Guide [2002], are the second generation control series printed on arrays from the AGRF microarray facility, including those from the Pax5, DC and QDC experiments. Table 2.2 shows the controls which make up the LUS series. Ratio controls at both high (U03High, U10High, D03High, D10High) and low-medium (U03Low, U10Low, D03Low, D10Low) intensities, as well as calibration controls (CC) are spiked into the mRNA mixes of each channel to give known fold-changes which again provide some truth in gene expression for each data set.

The spiked-in LMS and LUS controls often exhibit slightly different intensity dependent biases to the regular clones, and require separate normalisation (as noted in chapter 3 of Thorne [2004]). This is a result of their artificial spike-in nature; if the amount spiked into each sample (Cy3 and Cy5) before labelling differs by even a small amount, then the level of log-ratios for all spike-in controls will be shifted up or down (see figure 2.9 (a)). These controls were printed in different print-tip groups on the arrays, so a global intensity based loess normalisation \(M_{\text{norm}} = M – \text{loess}(A)\) is recommended to adjust the log-ratios for the spike-in controls on each array (see figure 2.9 (b)). All spike-in controls used in this thesis have been normalised in this way.
Figure 2.9: MA plot with LMS controls before (a) and after (b) normalisation. Different systematic dye-biases are evident for the genes (fine black points) and the controls (colour-coded points), which motivates separate normalisation. For the 120 LMS control spots on each array, a single intensity dependent loess fit with a span of 0.75 is used to adjust the log-ratios. The genes are normalised in the usual way using intensity dependent print-tip loess. These values are from array HD60 (slide 22 out of 111) from the QC data set.
Table 2.2: Summary of LUS controls.

<table>
<thead>
<tr>
<th>Control</th>
<th>Abbrev.</th>
<th>$M$</th>
<th>$A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up-regulated 3-fold Low Intensity</td>
<td>U03Low</td>
<td>$\log_2(3)$</td>
<td>low-med</td>
</tr>
<tr>
<td>Up-regulated 10-fold Low Intensity</td>
<td>U10Low</td>
<td>$\log_2(10)$</td>
<td>low-med</td>
</tr>
<tr>
<td>Down-regulated 3-fold Low Intensity</td>
<td>D03Low</td>
<td>$-\log_2(3)$</td>
<td>low-med</td>
</tr>
<tr>
<td>Down-regulated 10-fold Low Intensity</td>
<td>D10Low</td>
<td>$-\log_2(10)$</td>
<td>low-med</td>
</tr>
<tr>
<td>Up-regulated 3-fold High Intensity</td>
<td>U03High</td>
<td>$\log_2(3)$</td>
<td>high</td>
</tr>
<tr>
<td>Up-regulated 10-fold High Intensity</td>
<td>U10High</td>
<td>$\log_2(10)$</td>
<td>high</td>
</tr>
<tr>
<td>Down-regulated 3-fold High Intensity</td>
<td>D03High</td>
<td>$-\log_2(3)$</td>
<td>high</td>
</tr>
<tr>
<td>Down-regulated 10-fold High Intensity</td>
<td>D10High</td>
<td>$-\log_2(10)$</td>
<td>high</td>
</tr>
<tr>
<td>Calibration controls</td>
<td>CC</td>
<td>0</td>
<td>low/med/high</td>
</tr>
</tbody>
</table>

2.8 Generalising the results

By looking at data from a wide range of sources, it is expected that the results from this thesis will be generalisable to two-colour microarray data from other laboratories. The microarrays analysed in this thesis were printed with a number of different cDNA and oligo libraries at various core facilities and have been hybridised by many different experimenters. They were also scanned using different scanners and quantified using 2 image analysis software packages. These factors guarantee that there will be considerable variability in the gene expression measurements, which will have arisen from many different sources. These characteristics are ideal for a study of data quality.

One major limitation that must be kept in mind, comes from the reliance on the LMS and LUS spike-in controls for truth in differential expression. While these controls should give predictable log-ratios and have known differential expression status, they are artificial genes, which often have the largest log-ratios on an array, and mostly occur at high intensities (refer to figure 2.9). This means that results obtained by looking at these controls should be interpreted carefully, and may not generalise perfectly to the regular probes, particularly those at lower intensities. However, until some other form of true gene expression is available in experiments, comparing analysis methods using these controls is the best that can be done.
Chapter 3

Background Correction

In this chapter, the effects of different background estimates and correction methods for two-colour microarray data will be examined. Section 3.1 gives an introduction to this topic and section 3.2 presents a literature review on the different approaches of estimating the background and adjusting the foreground signal for background fluorescence. In section 3.3, a number of these methods are selected for comparison on data sets where some independent truth in gene expression is known. Results of this comparison are given in section 3.4 and discussed in section 3.5.

3.1 Introduction

In many optical technologies where signal is measured by an instrument, the observed signal, \( O \) can be considered to consist of a true signal, \( S \), along with some added background signal, \( B \), such that \( O = S + B \). The measurement of the intensity in each channel on a two-colour microarray using a fluorescent scanner is a good example of this. Excitation of the array with light sources at two different wave-lengths (lasers at 532 nm and 635 nm in the case of the GenePix 4000B scanner) give intensity readings on the scale \([1, 65536]\) for each pixel in each channel.

The pixels which make up the foreground region of a spot are likely to be brighter than the pixels surrounding this region, since probe (cDNA or oligonucleotide sequence) has been deposited there and hybridisation may have occurred. However, the surrounding
pixels don’t tend to have the lowest intensity value, due to non-specific binding of labelled sample on the glass around the spot, other processing effects such as deposits left after the wash stage, as well as general background noise to name a few examples. The need to remove this part of the signal which is not due to hybridisation of mRNA to the spot from the foreground intensity to obtain an accurate measure of the true intensity is referred to in the literature as ‘background correction’.

In most image analysis packages a summarised measure for the foreground ($o$) and background ($b$) is obtained for each spot after segmentation of the image. Since the background is assumed to be additive to the true signal on the original scale, the true signal is estimated by subtracting these two quantities ($s = o - b$). For each spot on a two-colour microarray, a background corrected intensity in each channel ($R = R_f - R_b$, and $G = G_f - G_b$) is typically used to form the log-ratio ($M$).

In this chapter, different estimators of background ($R_b$ and $G_b$) as well as some alternate processing methods (variants on subtracting the background) will be compared. Their effect on the phenomenon of ‘fanning’ of log-ratios at low intensities, the number of missing values, the precision and bias of the gene expression measurements and ability to detect known differential expression will be assessed for each method in search of a ‘best’ alternative.

### 3.2 Literature review

#### 3.2.1 Different background estimation and correction procedures

Different approaches to estimating background on an array are implemented in different image analysis software. Yang et al. [2002b] groups these estimates into three categories.

- **Constant background**, where a global value is used in each channel, which may be taken as the foreground intensity from negative control spots, or a percentile of the foreground intensities (e.g. 3%) of regular genes.

- **Local background**, estimated from pixels surrounding the spot mask, usually taken as the mean or median of the pixel intensities.
• Morphological background, estimated by applying a non-linear filter to each image using a large local window. For instance a morphological opening (known as morph background) consists of applying an erosion operator (local minimum) followed by a dilation (local maximum) using a window of fixed size for each image. The morph.close.open background is obtained after applying a dilation, followed by an erosion and another dilation. These operations remove the bright foreground pixels and replace these with nearby background values (minimum for an erosion and maximum for dilation), thus generating an image which is an estimate of the background over the entire slide. These methods are implemented in the Spot software (Buckley [2000]). The morph background tends to give a less variable background estimate which is not biased upwards by the inclusion of bright pixels belonging to the foreground or edge of a spot.

Bengtsson [2003] investigates methods of background estimation, looking at so called local ‘region’ methods as employed in GenePix and the morphological methods used in Spot and proposes two variants on these existing approaches. A local background method which uses a fixed number of pixels to estimate the background in a local region near the spot differs from the default GenePix local background which allows the number of background pixels to vary depending on the radius of the foreground mask. A variant on the morphological method, known as rank filtering is also introduced. Rather than taking the minimum (erosion), followed by the maximum (dilation), a percentage filter of say 10% followed by 70% using the same local window is applied to each image. The finding on this study are that both methods can give less biased, more precise estimates of the background. However, optimising the percentage cut-off used in the rank filtering was found to be non-trivial, which detracts from the usability of such an approach.

Yang et al. [2002b] conclude that the segmentation algorithm has less of an effect on the log-ratios than the background estimate used. Figure 3.1 plots the foreground and background values in each channel for Spot and GenePix data from the same slide (array 1 from the K562 versus Pool experiment) to highlight the differences. While the foreground intensities are broadly similar despite the use of different segmentation algorithms (SRG for Spot and adaptive circle method for GenePix), for the background, the Spot morph estimates are systematically lower than the local median background mea-
Figure 3.1: Foreground and background measures for Spot and GenePix on the same array. The red dashed lines represent equality (slope 1, intercept 0). While the foreground measures are similar in each channel, Spot’s `morph` background is systematically lower than the GenePix local median background.

Yang et al. [2002b] also notes that subtracting these local background values from the foreground before constructing log-ratios adds considerable variability to the log-ratios within and between slides, compared with not background correcting. Using an intermediate background value such as the `morph` estimate in Spot was found to perform best in terms of bias-variance trade-off, producing more extreme t statistics for known DE genes compared with using no background, constant background or local background.

In Yang et al. [2001b], MA plots for a given slide using data from several image analysis programs (GenePix, ScanAlyze and Spot) are presented. Background correction can be seen to have a more marked effect on the variability of low intensity spots where $R_f \approx R_b$ or $G_f \approx G_b$. This behaviour can be expected, since the ratio of two quantities ($R/G$) tend to be very variable when at least one of those quantities is small. The authors recommend not background correcting the data when morphological background estimates are not available.
This ‘fanning’ of log-ratios at very low intensities, depicted in figure 3.2, has been noted by many authors including Beißbarth et al. [2000], Baggerly et al. [2001], Yang et al. [2002a], Finkelstein et al. [2002], Kooperberg et al. [2002] and Bilban et al. [2002] to name a few. Another undesirable side effect of subtracting background are negative intensities, which occur when \( R_f < R_b \) or \( G_f < G_b \) and lead to missing log-ratios (NAs in R and S-Plus). Genes expressed at low levels \( (R_f \approx R_b, G_f \approx G_b) \) are more likely to be censored by background correction, which has the potential to remove interesting genes from the analysis, particularly in experiments with multiple conditions, such as a time-course, where genes may become expressed over time. Missing values are also problematic when normalising between arrays, or when using multivariate techniques such as principal components analysis (PCA) and singular value decomposition (SVD).

How to best use the background values at hand to minimise these negative side effects has been discussed in many papers. Tran et al. [2002] choose not to background correct GenePix data, or add an offset to avoid negative signals.

Kooperberg et al. [2002] suggest an empirical Bayes model made up of a convolution of normal distributions to background adjust the signals. Observed foreground and background means \( (X_f \text{ and } X_b) \), standard deviations \( (SD_f \text{ and } SD_b) \) and the number of foreground and background pixels \( (n_f \text{ and } n_b) \) for each spot in a given channel are used in the model,

\[
p(\mu|\sigma_b, \sigma_f, X_f) = \frac{\phi \left( \frac{X_f - \mu - X_b}{\sigma_d} \right) \Phi \left( \frac{(X_f - \mu) \sigma_f^2 + X_b \sigma_b^2}{\sigma_f \sigma_b \sigma_d} \right)}{\sigma_d \int_0^\infty \Phi \left( \frac{X_f - v}{\sigma_f} \right) \phi \left( \frac{X_b - v}{\sigma_b} \right) dv}
\]  

(3.1)

where \( \phi(.) \) is the density of the standard normal distribution, \( \Phi(x) = \int_{-\infty}^x \phi(y)dy \) is the cumulative standard normal distribution, \( \sigma_f = a SD_f / \sqrt{n_f} \), \( \sigma_b = a SD_b / \sqrt{n_b} \), \( \sigma_d = \sqrt{\sigma_b^2 + \sigma_f^2} \) and \( a \) is a scaling factor. Numerical integration is applied to obtain the expected value of the true signal \( E(\mu|X_b, X_f, \sigma_b, \sigma_f) \) in each channel for each spot. This model-based method avoids missing values and was demonstrated to reduce the high variability of low intensity log-ratios when used on data from a self-self hybridisation where there is no differential expression.
Figure 3.2: MA plot illustrating ‘fanning’ of log-ratios at low intensities ($A < 7$). These data are from a self-self hybridisation where there should be no differential expression. GenePix data were used, with the median background subtracted from the foreground measures before forming log-ratios. The control spots are colour-coded. Figure courtesy of Gordon Smyth.

A simpler method with the same aim is suggested in Edwards [2003], who adjusts the foreground intensities as follows:

\[
R = \begin{cases} 
R_f - R_b & \text{if } R_f - R_b > \delta \\
\delta \exp[1 - (R_b + \delta)/R_f] & \text{otherwise} 
\end{cases}
\]

\[
G = \begin{cases} 
G_f - G_b & \text{if } G_f - G_b > \delta \\
\delta \exp[1 - (G_b + \delta)/G_f] & \text{otherwise} 
\end{cases}
\]
In this model, subtraction of the background is done as usual when the difference between the foreground and background is larger than a threshold value $\delta$, however when the difference is small or negative ($\leq \delta$), subtraction is replaced by a smooth monotonic function. In this paper, $\delta = 1$ was used in the analysis.

### 3.2.2 Approaches which deal with ‘fanning’ and missing values

The case for using an alternative transformation to the logarithm is put forward in Durbin et al. [2002], Huber et al. [2002], Rocke and Durbin [2003], Kafadar and Phang [2003], Cui et al. [2003] and Durbin and Rocke [2004]. Transformations such as the $\text{arcsinh}$, log-linear, shift-log and started-log have been suggested, sometimes for the purpose of removing curvature of the data at low intensities as a form of normalisation, and other times ($\text{arcsinh}$ and log-linear) to stabilise the variance at low intensities.

In Durbin and Rocke [2004], a generalised log transform which stabilises the variance of the differences of the transformed intensities for all spots using the delta method is described. A started-log and generalised log transform which approximately stabilise the delta method variance are also described, and all of these approaches were found to be superior to the ordinary log transform in stabilising the variance at all intensities.

Cui et al. [2003] recommends that transformations should be used as appropriate. Examining a plot of the smoothed IQR of the untransformed, or transformed measure of differential expression versus intensity is suggested as a simple diagnostic plot.

Modelling log-ratio variability based on intensity is another approach that deals with this problem. This is done in various ways by Newton et al. [2001], Baggerly et al. [2001] and Yang et al. [2002a] for the purpose of choosing DE genes.

Methods to impute missing values when analysis techniques are used which need the complete data have been proposed by Troyanskaya et al. [2001] and Zhou et al. [2003]. In Ouyang et al. [2004] a number of methods for imputing missing values for large microarray data sets are compared.

All of these approaches deal with the symptoms rather than the cause of the problem which is rooted in the practice of background correction. When no background correction is applied, and the data are appropriately normalised, variability of the log-ratios tends
not to change dramatically with intensity, and missing values are avoided. This is arguably a more convenient solution than the ones proposed in this section, however whether this method, or some of the other model-based correction methods still gives the appropriate genes as DE is worthy of further investigation.

3.2.3 Observations from the analysis of Affymetrix data

In chapter 2 of Bolstad [2004], the estimation and processing of background for Affymetrix GeneChip® data are examined in detail. For this single-channel technology, background is not explicitly measured using local pixels as is done for two-colour spotted arrays. An early approach of adjusting the signal was to subtract the intensity of each mismatch probe from its corresponding perfect match probe (PM-MM), however this tended to result in many negative values, which become missing when a log transform is taken. Alternative methods, including a convolution model which consists of an exponential (signal) and normal (background) component, not background correcting, and a standard curve adjustment method were investigated. These methods were compared using the Affymetrix HGU95A spike-in data set, and found to be favourable in different ways. In particular, not background correcting the data gave the best results in terms of selecting known DE genes, although the true fold-changes were systematically underestimated (biased). Performing some form of background adjustment gave more accurate log-ratios at the expense of more variability.

3.3 Comparison of background correction methods

Given the large effect of background on log-ratio variability, a more detailed study on the effects of the various estimates and background processing methods will be undertaken, with the aim of choosing the most appropriate combination of the two for use in the study of quality in chapters 4 and 5.

Table 3.1 lists the methods for comparison. Log-ratios were formed using Spot (Buckley [2000]) data with \texttt{morph} and \texttt{morph.close.open} background subtracted from the mean foreground in each channel. Log-ratios from GenePix (GenePix Pro [2001]) data using no background or median local background subtracted from the mean foreground were
also considered, along with log-ratios formed using the alternative processing methods introduced in the literature review (section 3.2.1). A competing model-based alternative which will be referred to as the normal-exponential convolution model, and the specifics of the moving minimum method and other model-based approaches will be discussed in sections 3.3.1 and 3.3.2. The $M$ values for each method were normalised using intensity based print-tip loess.

### Table 3.1: Summary of the background correction methods compared.

<table>
<thead>
<tr>
<th>Method</th>
<th>Image Analysis Software</th>
<th>Details</th>
<th>Processing method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spot morph</td>
<td>Spot 2.0</td>
<td>morph</td>
<td>subtraction</td>
</tr>
<tr>
<td>Spot morph.c.o</td>
<td>Spot 2.0</td>
<td>morph.close.open</td>
<td>subtraction</td>
</tr>
<tr>
<td>GP no bg</td>
<td>GenePix Pro 3.0/4.0</td>
<td>none</td>
<td>-</td>
</tr>
<tr>
<td>GP median</td>
<td>GenePix Pro 3.0/4.0</td>
<td>local median</td>
<td>subtraction</td>
</tr>
<tr>
<td>GP moving min</td>
<td>GenePix Pro 3.0/4.0</td>
<td>local median</td>
<td>moving minimum</td>
</tr>
<tr>
<td>GP Kooperberg</td>
<td>GenePix Pro 3.0/4.0</td>
<td>local median</td>
<td>model (see equation 3.1)</td>
</tr>
<tr>
<td>GP Norm+Exp</td>
<td>GenePix Pro 3.0/4.0</td>
<td>local median</td>
<td>model (see equation 3.3)</td>
</tr>
<tr>
<td>GP Edwards</td>
<td>GenePix Pro 3.0/4.0</td>
<td>local median</td>
<td>model (see equation 3.2)</td>
</tr>
</tbody>
</table>

As mentioned in section 3.2.1, `morph` and `morph.close.open` backgrounds differ in the order in which the morphological operators are applied. The local median background measured in GenePix and the `morph.close.open` background from Spot are generally similar and larger than the `morph` background (see figure 3.1).

### 3.3.1 Normal-exponential convolution model

A convolution model for two-colour microarray data adapted from Bolstad [2004] will be described in this section. The motivation for this model comes from looking at the distribution of the observed foreground signals ($o$) in each channel for a given array (see figure 3.3). Assume that the foreground ($O$), true signal ($S$) and background signal ($B$) are additive ($O = S + B$) and independent, and that $S$ is exponentially distributed with mean $\alpha$, and $B$ is normally distributed with mean $\mu$ and standard deviation $\sigma$. The joint density of $S$ and $B$ is

$$p(s, b) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{1}{2\sigma^2} (b - \mu)^2 \right) \frac{1}{\alpha} \exp \left( -s/\alpha \right)$$
for $s > 0$. The joint density of $O$ and $S$ is therefore

$$p(o, s) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(o - s - \mu)^2\right) \frac{1}{\alpha} \exp\left(-\frac{s}{\alpha}\right).$$

Now

$$-\frac{1}{2\sigma^2}(s - o + \mu)^2 - s/\alpha = -\frac{1}{2\sigma^2} (s^2 - 2(s - \mu - \sigma^2/\alpha)s + (s - \sigma^2/\alpha)^2)$$

$$= -\frac{1}{2\sigma^2} (s - (o - \mu - \sigma^2/\alpha))^2 - \frac{1}{2\sigma^2} \left((o - \mu)^2 - (o - \mu - \sigma^2/\alpha)^2\right)$$

$$= -\frac{1}{2\sigma^2} (s - (o - \mu - \sigma^2/\alpha))^2 - \frac{1}{2\sigma^2} (2(o - \mu)\sigma^2/\alpha) - (\sigma^2/\alpha)^2$$

$$= -\frac{1}{2\sigma^2} (s - (o - \mu - \sigma^2/\alpha))^2 - (o - \mu)/\alpha + \frac{1}{2\sigma^2}/\alpha^2$$
so we can re-write the joint density as

\[ p(o, s) = \frac{1}{\alpha} \exp \left( -\frac{(o - \mu)}{\alpha} + \frac{1}{2} \frac{\sigma^2}{\alpha^2} \right) \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{1}{2\sigma^2} (s - \mu_{s,o})^2 \right) \]

with \( \mu_{s,o} = o - \mu - \sigma^2/\alpha \).

The marginal distribution of \( O \) arises from integrating with respect to \( s \) as follows

\[ p(o) = \frac{1}{\alpha} \exp \left( -\frac{(o - \mu)}{\alpha} + \frac{1}{2} \frac{\sigma^2}{\alpha^2} \right) \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{1}{2\sigma^2} (s - \mu_{s,o})^2 \right) ds \]

\[ = \frac{1}{\alpha} \exp \left( -\frac{(o - \mu)}{\alpha} + \frac{1}{2} \frac{\sigma^2}{\alpha^2} \right) \left( 1 - \Phi(0; \mu_{s,o}, \sigma^2) \right) \]

where \( \Phi \) is the cumulative normal distribution. The parameters \( \alpha, \mu \) and \( \sigma^2 \) can be estimated via maximum likelihood (ML) using the log likelihood

\[ \log p(o) = -\log \alpha - \frac{(o - \mu)}{\alpha} + \frac{1}{2} \frac{\sigma^2}{\alpha^2} + \log \left( 1 - \Phi(0; \mu_{s,o}, \sigma^2) \right) . \]

The conditional distribution of \( S \) given \( O \) is the truncated normal distribution

\[ p(s|o) = \frac{p(o, s)}{p(o)} = \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{1}{2\sigma^2} (s - \mu_{s,o})^2 \right)}{1 - \Phi(0; \mu_{s,o}, \sigma^2)} \]

for \( s > 0 \). Now

\[ \frac{\partial \log p(s|o)}{\partial \mu_{s,o}} = \frac{1}{\sigma^2} (s - \mu_{s,o}) - \frac{\phi(0; \mu_{s,o}, \sigma^2)}{1 - \Phi(0; \mu_{s,o}, \sigma^2)} \]

which gives

\[ E(S|O = o) = \mu_{s,o} + \sigma^2 \frac{\phi(0; \mu_{s,o}, \sigma^2)}{1 - \Phi(0; \mu_{s,o}, \sigma^2)} \]

where \( \phi \) is the normal density function.

The major distinction between the convolution model proposed in Bolstad [2004] and this model, is the incorporation of the observed background in the calculations. For the \( i \)th spot we observe the foreground \( (o_i) \) and background \( (b_i) \) intensities. Assuming that \( B_i|b_i \sim N(\beta + b_i, \sigma^2) \) and \( S_i \sim \exp(\alpha) \) gives the common parameters \( \alpha, \beta \) and \( \sigma^2 \) which are assumed constant across an array. The modified log likelihood function

\[ \log p(o) = -\log \alpha - \frac{(o - \mu_i)}{\alpha} + \frac{1}{2} \frac{\sigma^2}{\alpha^2} + \log \left( 1 - \Phi(0; \mu_{s,o,i}, \sigma^2) \right) \]
where $\mu_{s,o,i} = a_i - \mu_i - \sigma^2/\alpha$ and $\mu_i = \beta + b_i$ was used to obtain ML estimates $\hat{\alpha}, \hat{\beta}$ and $\hat{\sigma}^2$ in each channel using the Nelder-Mead simplex algorithm. These estimates are used to obtain adjusted signals according to

$$E(S_i|b_i, O_i = o_i) = \mu_{s,o,i} + \sigma^2 \frac{\phi(0; \mu_{s,o,i}, \sigma^2)}{1 - \Phi(0; \mu_{s,o,i}, \sigma^2)},$$

(3.3)

Equation 3.3 gives the background corrected intensity for the $i$th spot in a given channel; the corrected signals from both channels are used to form the $M$ and $A$ values for each spot.

### 3.3.2 Details of other background processing methods

The moving minimum method uses a given spot’s background and the background from its 8 nearest neighbours (where available) and takes the minimum of these values as the new background estimate. The moving minimum is then subtracted from the foreground to obtain an estimate of the true signal. This method is applied to the GenePix data with median local background in the data sets examined.

For the model proposed in Edwards [2003], the parameter $\delta$ was adaptively chosen from the data. The quantile of the difference between foreground and background in each channel which was 10% above the number of negative background corrected values was chosen when negatives were present. If there were no negative values, the minimum value was used. Adjustments as per equation 3.2 were then made.

The function `kooperberg` was implemented in the `limma` package to adjust the foreground signals according to the normal convolution model specified in Kooperberg et al. [2002]. The $a$’s are first estimated in each channel by regressing the observed background variability ($SD_b/\sqrt{m_b}$) on the empirical standard deviation of the background from the 3 (for spots on an outer row/column) or 4 nearest neighbour spots within a print-tip group using `lm`. The R code to do this was modified from Charles Kooperberg’s S-Plus code (supplied in personal communication). The `integrate` function with 10,000 subdivisions was used to perform the numerical integration to obtain the expected signal, $E(\mu_t|X_b, X_f, \sigma_b, \sigma_f)$ in each channel for each spot. On a standard desktop computer (Pentium 2.0GHz processor with 1024Mb of RAM) a typical array takes around 3 minutes
to process. The corrected signals were used to obtain $M$ and $A$ values for each spot. This method was previously unavailable in public domain software, and was implemented for the purpose of this evaluation.

The other methods compared in this chapter are implemented in the backgroundCorrect function in limma by changing the method argument ("none", "subtract", "movingmin", "normexp" or "edwards" are the relevant options).

3.3.3 Using independent truth to compare the methods

The K562 vs Pool experiment (described in section 2.1), will be used to compare the variability of the log-ratios derived using the different background methods and to see how well each one performs at choosing DE genes. In this experiment, the Affymetrix data independently measures gene expression, and can be used to choose the ‘true’ DE genes. For our purposes, it is not necessary for the Affymetrix data to give perfect truth, provided that deviations from truth do not selectively advantage any of the methods being compared, which is probably a reasonable assumption. The ability of the background methods applied to the two-colour data at recovering these DE genes can be used as a performance indicator to rank the alternatives.

To assess the bias of the different methods, the LUS controls from the Pax5 experiment (refer to section 2.2) were used. These controls, expressed at known fold-changes and varying intensities (see table 2.2, section 2.7), allow bias to be assessed at different intensities. The Pax5 experiment was chosen as it had the largest number of these controls along with GenePix and Spot data. The log-ratios for the controls were normalised separately for each method using a global intensity based loess normalisation (with span=0.6).

3.4 Results

3.4.1 Assessment of variability

The normalised MA plots for all genes on the two-colour arrays for slide 1 of 6 are shown in figure 3.4. For Spot data with morph background subtracted and GenePix data with no background subtracted, the log-ratios are least variable at all intensities. The ranges of $A$
values are also reduced for these methods. Visually, the most variation in \( M \) values within a slide is evident for Spot data with \textit{morph.close.open} background subtracted, GenePix with median background subtracted or Edwards background adjustment, especially at low intensity. Both the Kooperberg and normal-exponential model-based adjustment of GenePix data reduce the amount of ‘fanning’ at low intensities.

Table 3.2 shows the number and percentage of missing log-ratios for each method for the 7,547 probes in common between the platforms across the 6 two-colour arrays. For this data set, median background subtraction resulted in the most missing values (4.73%), followed by moving minimum subtraction (2.57%) and \textit{morph.close.open} subtraction (2.08%). The remaining methods had very few (0.02%) or no missing values. It is worth noting that the proportion of missing values for this data set is quite low; more typically rates of between 5% and 20% are obtained with GenePix data when median background is subtracted.

Table 3.2: The number and percentage of missing \( M \) values for each background correction method (maximum shown in bold).

<table>
<thead>
<tr>
<th>Method</th>
<th>Count</th>
<th>% missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spot morph</td>
<td>9</td>
<td>0.02</td>
</tr>
<tr>
<td>Spot morph.c.o</td>
<td>940</td>
<td>2.08</td>
</tr>
<tr>
<td>GP no bg</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GP median</td>
<td>2,141</td>
<td>4.73</td>
</tr>
<tr>
<td>GP moving min</td>
<td>1,165</td>
<td>2.57</td>
</tr>
<tr>
<td>GP Kooperberg</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GP Norm+Exp</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GP Edwards</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The averaged \( M \) and \( A \) values across these 6 slides for the 7,547 common probes are shown in figure 3.5. The ‘fanning’ phenomena is no longer evident in these plots, indicating that the excessive low intensity variability observed in figure 3.4 for some background correction methods is probably artifactual. Alternatively, it could indicate that the genes in common between the platforms are all highly expressed, which would unfortunately mean that this data set is not going to be very sensitive to the background method chosen. It also appears that more genes are up regulated in the Pool mRNA compared to the K562 mRNA, since there are more large negative \( M \) values than positive values. This
Figure 3.4: MA plots obtained using different background correction methods for array 1 out of 6. The y-axes for some plots have been truncated slightly. The full range of $M$ values for each method is given in the top right-hand corner of each plot.
would be expected given that the Pool mRNA is constructed from 10 different cell lines.

The correlations of the averaged $M$ values for the ‘same’ genes across platforms is quite high (around 0.78) for all methods, and a little lower (around 0.58) for the averaged $A$ values (see table 3.3). For the $M$ values, Spot using morph background, GenePix with no background subtracted, and the Kooperberg and normal-exponential model-based adjustment methods had the highest correlation ($r = 0.79$) with the Affymetrix data, and for the $A$ values, the same methods tended to give the highest correlations ($r = 0.59$).

Table 3.3: Correlations between gene expression measurements from Affymetrix and spotted arrays for the different background correction methods. The Pearson correlation coefficient ($r$) between the average log-ratios ($M$) from the spotted microarrays versus the Affymetrix GeneChip® arrays were calculated for the common genes, and rounded to 2 decimal places. The same was done for the average intensities ($A$). The maximum correlations are shown in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>Correlation ($r$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M$</td>
</tr>
<tr>
<td>Spot morph</td>
<td><strong>0.79</strong></td>
</tr>
<tr>
<td>Spot morph.c.o</td>
<td>0.76</td>
</tr>
<tr>
<td>GP no bg</td>
<td><strong>0.79</strong></td>
</tr>
<tr>
<td>GP median</td>
<td>0.76</td>
</tr>
<tr>
<td>GP moving min</td>
<td>0.77</td>
</tr>
<tr>
<td>GP Kooperberg</td>
<td><strong>0.79</strong></td>
</tr>
<tr>
<td>GP Norm+Exp</td>
<td><strong>0.79</strong></td>
</tr>
<tr>
<td>GP Edwards</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Figure 3.6 plots the residual standard deviations ($\hat{\sigma}$) for each method using the common clones which have no missing log-ratios under any method (5,912 out of 7,547). The log-ratios formed from the Affymetrix data are the least variable on average, followed by GenePix with no background subtracted and Spot with morph background subtracted. $M$ values formed using Spot with morph.close.open background correction are most variable on average, followed closely by GenePix data with median background subtracted. The moving minimum, normal-exponential model, Kooperberg and Edwards background adjustment methods all reduce the variation compared with GenePix median background subtraction by small amounts. The normal-exponential model reduces the residual variation the most, followed by the moving minimum method, Kooperberg and finally Edwards adjustment.
Figure 3.5: Averaged MA plots obtained using different background correction methods and Affymetrix for the 7,547 common probes.
Figure 3.6: Residual $\hat{\sigma}$ for different background correction methods. $\log_2(\hat{\sigma})$ were plotted for the common probes which had no missing values for any method on any array (5,912 out of 7,547 common probes).

3.4.2 Detecting differential expression

For the K562 versus Pool experiment, the true positive DE genes were selected using the Affymetrix data, by ranking the genes using the empirical Bayes log odds statistics of Smyth [2004]. Genes falling below a certain cut-off for the log odds were regarded as non-DE and those above were classified as DE. Cut-offs of 5 (High), 2.5 (Medium) or 0 (Low) were chosen and resulted in 93 (1.2%), 311 (4.1%) and 783 (10.4%) DE genes respectively.

The choice of background and processing method has been shown in section 3.4.1 to dramatically affect the $M$ values for two-colour microarray data, which will have a bearing on the performance of a test based on $|M|$ to detect differential expression. To measure this for a particular method given a set of DE genes, percentiles of the $|M|$ values were taken for each slide, and the number of true positives and false positives at
each percentile were accumulated. The results were then averaged between slides to get a receiver operator characteristic (ROC) curve. This process was repeated for each method at different stringencies for differential expression to give the ROC curves in figures 3.7 and 3.8. A ROC curve shows the rate of true positives (y-axis) against the rate of false positives (x-axis). An ideal test should get 100% true positives with 0% false positives; in practice, a higher curve represents a better test criteria.

Although the results are close, using Spot data with *morph* background and GenePix data with no background subtracted seems to give slightly more true positives for a given number of false positives, particularly at low false positive rates (see figure 3.7) and at higher false positive rates for medium and low stringencies of differential expression (see figure 3.7). The Kooperberg method and normal-exponential models perform best when the rate of false positives is higher (see figure 3.8) and when the stringency for differential expression is high, and GenePix median background subtraction is consistently the worst, due to a greater number of missing values and high variability of the $M$ values at low intensities.

The area under the ROC curves (see table 3.4) can be used to rank the methods globally, which is useful when the curves intersect. For all 3 stringencies, GenePix data with median background subtraction is worst, followed by Spot with *morph.close.open* background and moving minimum adjustment. For medium and low stringency, the no or low background alternatives (GenePix no background and Spot *morph*) perform better than the model-based approaches, however for high stringency, Kooperberg adjustment is marginally better than the normal-exponential model, followed closely by the Edwards method, then Spot *morph* and finally GenePix with no background subtraction. For the comparison made in this experiment (i.e. K562 mRNA versus Pool mRNA from 10 different cell lines), the results from the low stringency scenario (i.e. quite a lot of differential expression) are probably the most realistic. In all cases, the differences between the no or low background subtraction methods and the 3 model-based alternatives are very small. It is also worth noting that the performance of the methods which subtract background from foreground, almost uniformly worsens as the median level of the background estimate subtracted increases.
Table 3.4: Area under ROC curves ordered by the median size of the background estimates (smallest to largest) used by the method. The maximum values are shown in bold.

| Method            | Median Background 
<table>
<thead>
<tr>
<th></th>
<th>$(R_b, G_b)$</th>
<th>Differential expression cut-off</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>High</td>
</tr>
<tr>
<td>GP no bg</td>
<td>(0, 0)</td>
<td>0.872</td>
</tr>
<tr>
<td>Spot morph</td>
<td>(48.5, 56.3)</td>
<td>0.875</td>
</tr>
<tr>
<td>Spot morph.c.o</td>
<td>(116.6, 139.2)</td>
<td>0.849</td>
</tr>
<tr>
<td>GP moving min</td>
<td>(125.4, 146.8)</td>
<td>0.852</td>
</tr>
<tr>
<td>GP median</td>
<td>(125.4, 146.8)</td>
<td>0.822</td>
</tr>
<tr>
<td>GP Kooperberg</td>
<td>(125.4, 146.8)</td>
<td><strong>0.887</strong></td>
</tr>
<tr>
<td>GP Norm+Exp</td>
<td>(125.4, 146.8)</td>
<td>0.878</td>
</tr>
<tr>
<td>GP Edwards</td>
<td>(125.4, 146.8)</td>
<td>0.877</td>
</tr>
</tbody>
</table>

3.4.3 Assessment of bias

The $M$ values for the ratio controls from the Pax5 experiment tended to be systematically lower than the expected spike-in fold-changes for all background correction methods. This is likely to be caused by adding slightly less of the spike-in mix to each mRNA sample than is recommended which may be due to pipetting errors (Stephen Wilcox, personal communication).

Despite this, some methods gave log-ratios which were closer to the expected log fold-changes than others. Figure 3.9, shows the log-ratios for the LUS controls from slide 1 for each of the background correction methods. For the low-medium intensity ratio controls (D03Low, D10Low, U03Low, U10Low) the no or low background correction methods (GenePix no background and Spot morph) under-estimate the log-ratios more than the other alternatives. This is less evident for the high intensity ratio controls (D03High, D10High, U03High, U10High), and there appears to be no differences for the calibration controls, which tend to have appropriate $M$ values (0) for all alternatives.

The log-ratios from the LUS controls were summarised across all 9 slides by the coefficients $\hat{\beta}_{LUS_i}$, and the bias $(\Sigma(\hat{\beta}_{LUS_i} - \mu)/n)$ was calculated for each class of control, where $n$ and $\mu$ varied depending on the ScoreCard control. The biases by class of control are plotted in figure 3.10. These show that the no or low background subtraction methods tend to be more biased than the other methods for all controls, excepting the calibration and D03High spots. In addition, bias generally decreases as the level of the
background estimate used for background correction increases. The alternative processing methods (moving minimum, Kooperberg, normal-exponential and Edwards) tended to be roughly as biased as GenePix with median background subtracted. Across all classes of controls, the Kooperberg method gave the least biased results, followed by Edwards adjustment, GenePix with median background subtracted, Spot with \textit{morph.close.open} background correction, GenePix with moving minimum background subtracted, the normal-exponential model-based adjustment, Spot with \textit{morph} background subtracted and finally GenePix with no background subtracted.

### 3.5 Discussion

The main objective of this chapter was to determine whether the choice of background estimate and processing method has any dramatic effect on the results from a microarray experiment. The answer to this question is clearly ‘yes’. Subtraction of high background estimates such as local median or \textit{morph.close.open} introduces more missing values than other approaches and increases the overall variability of the log-ratios, particularly in the low intensity range. These findings are consistent with the results in Yang \textit{et al.} [2002b], and these characteristics make it harder for such methods to detect DE genes compared to the no or low background methods (such as GenePix no background and Spot \textit{morph} background) when $|M|$ is used to select differential expression. Using a moving minimum of the background on the GenePix data improved the situation slightly, with fewer missing values and more correctly called DE genes. However, the best results obtained using a local median background were achieved with the more sophisticated model-based correction procedures of Edwards [2003], Kooperberg \textit{et al.} [2002] or the normal-exponential convolution model described in this chapter. These methods eliminate missing values, reduce within slide log-ratio variability at low intensities, and slightly decrease the between slide variability. They also perform very well at choosing DE genes, doing marginally better than the no or low background methods when a high stringency was used to determine the DE genes.

Ranking the methods based on bias of the log-ratios puts any method which uses local background estimates ahead of the low or no background alternatives, which systematically
under-estimate log fold-changes of the ratio controls the most. Using Kooperberg model-based correction gives the least biased log-ratios across the different spike-in controls examined. Studies of background correction of Affymetrix data have arrived at similar conclusions. Background correction tends to increase the variation of the data which can result in more false positive calls when looking for DE genes, however the log-ratios are estimated with less bias than if the data were not background corrected (refer to chapter 2 of Bolstad [2004]).

Although the Kooperberg model was marginally better than the normal-exponential model and Edwards method, the latter approaches are more widely applicable since they only rely on the foreground and background estimates to correct the signal. The Kooperberg model uses the standard deviations of the foreground and background pixel intensities as well as the number of pixels, which are not measured in all image analysis software. In addition the computational burden of having to perform numerical integration twice for each spot on the array detracts from its appeal. For these reasons, the normal-exponential model would be the recommended alternative.

In practice, the choice of background method will depend on the purpose of the analysis. If accurate estimates of the log fold-changes are of primary interest, then using a local background estimate combined with an adjustment method which avoids missing values, such as the Kooperberg (for GenePix data) or normal-exponential model (more generally) will be most useful. For selecting DE genes, the low or no background methods have been shown to perform better.

Bearing these results in mind, it was decided not to background correct GenePix data when used in the remainder of this thesis. The properties of no missing values and low variability are both desirable, particularly when looking at the issue of spot and array quality, where variability is important. It is hoped that this will reduce the widely observed phenomena of intensity dependent variability of log-ratios. Where Spot data are used in the remainder of this thesis, the morph background will be subtracted from the foreground before forming log-ratios. It is worth noting that the latest version of GenePix Pro (version 6.0) implements the morph background as an option.
Figure 3.7: ROC curves at varying stringencies (High, Medium, Low) of differential expression for different background correction methods on the false positive range [0, 0.2].
Figure 3.8: ROC curves at varying stringencies (High, Medium, Low) of differential expression for different background correction methods on the false positive range [0.2, 1].
Figure 3.9: $M$ values for LUS controls from slide 1 of the Pax5 experiment for different background correction methods. The dashed black lines represent the theoretical spike-in log-ratios for each control type.
Figure 3.10: Bias of $M$ values for LUS controls from the Pax5 experiment for different background correction methods.
Chapter 4

Array Quality Weights

In this chapter, a novel approach to quality which makes use of array weights in the analysis of two-colour microarray data from designed experiments will be outlined. Section 4.1 presents a literature review covering the current methods available for dealing with array quality. In section 4.2, a joint mean and dispersion model applicable to microarray data is described along with a gene-by-gene update algorithm to estimate the array-specific variances in the dispersion model. The convergence of the algorithm and recovery of the true array variance parameters for a simulated data set is also given in this section. Results obtained using the array variances as inverse weights in the linear model analysis of the log-ratios for experimental data are presented in section 4.3. This added quality control step in the analysis aids in the detection of DE genes.

4.1 Literature review of current array quality methods

The quality of results from a microarray experiment can be affected by many factors at the printing and hybridisation stages. Schuchhardt et al. [2000] lists a number of the experimental steps involved in producing a microarray, which are summarised in table 4.1. Various authors have investigated the variability introduced at different stages in the process.

Spruill et al. [2002] focused on variability introduced by print-tips, 96-well plates used in array fabrication as well as gene and sample by fitting an ANOVA model to data from
Table 4.1: Sources of experimental variability in the production, hybridisation and scanning of microarrays (from Schuchhardt et al. [2000]). Many of these variables can affect all spots on an array.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>mRNA preparation</td>
<td>Depends on tissue, degradation</td>
</tr>
<tr>
<td>Transcription</td>
<td>Reverse transcription results in cDNAs of varying length</td>
</tr>
<tr>
<td>Labelling</td>
<td>Will depend on the method, direct or indirect</td>
</tr>
<tr>
<td>Amplification</td>
<td>Library is subject to amplification which may fail for some clones, and not for others</td>
</tr>
<tr>
<td>Print-tip variation</td>
<td>Different shapes in the tip will result in varying amounts of target being deposited</td>
</tr>
<tr>
<td>Fluctuations in target volume</td>
<td>Randomly fluctuates for a given pin</td>
</tr>
<tr>
<td>Target fixation</td>
<td>The amount of target which is chemically linked to the surface is unknown</td>
</tr>
<tr>
<td>Hybridization parameters</td>
<td>Efficiency of hybridisation is influenced by temperature, time, amount of probe used etc.</td>
</tr>
<tr>
<td>Slide inhomogeneities</td>
<td>Slide thickness may vary or spatial effects may arise due to handling</td>
</tr>
<tr>
<td>Non-specific/cross hybridisation</td>
<td>Typical error which is hard to correct for</td>
</tr>
<tr>
<td>Image analysis</td>
<td>Saturation effects</td>
</tr>
</tbody>
</table>

custom arrays. Multiple prints of each gene were present on these arrays, with replication performed using 2 plates containing the same clones and 2 pins, to enable estimation of the fixed effects. Of the technological variation, plate added considerably more variability than print-tip, and the biological variation arising from the samples (mRNA) was intermediate of the two.

In a similar investigation, Wildsmith et al. [2001] varied 6 factors including type and age of dye, type of enzyme, incubation time and type of RNA (total or messenger) to optimise their protocol. A fractional factorial experiment was set up and an ANOVA of the log-ratios indicated that type of RNA and dye contributed the most to variation in the signal.

Novak et al. [2002] studied the technical and biological variation due to RNA preparation. RNA derived from cell-lines, and whole tissue from the same, and different animals were compared using Affymetrix GeneChip® arrays. Expression values obtained from cell-lines were found to be less variable than those obtained from tissue samples.

In practice, the amount of variability introduced at each step in the protocol will vary between labs, and may change over time. While replicating the important steps of the experimental process enables us to get a handle on these errors, in routine experiments it
may not be practical due to limited availability of sample or cost considerations. Some replication is still essential though as stressed by Lee et al. [2000], Kerr and Churchill [2001b], Yang et al. [2002a], Tran et al. [2002] and Smyth et al. [2003] to name a few. In Lee et al. [2000] and Yang et al. [2002a] at least three repeat hybridisations are recommended, while Pavlidis et al. [2003], after examining a number of publicly available data sets, concludes that a minimum of 5 and a maximum of 15 arrays (depending on the study) should produce ‘stable’ results. Other authors have made attempts to recommend the number of replicate arrays to include in a designed experiment based on simple power analyses (see Pan et al. [2002], Hwang et al. [2002] and Zien et al. [2003]).

Kerr and Churchill [2001b] modelled array and dye effects explicitly in their ANOVA of log-intensities. Using the same model, Draghici et al. [2001] computes array variances as measures of quality, and although no guidance as to how these would be used is given in the paper, it is presumed that arrays with high variation are removed from the analysis. A recent paper by Chen et al. [2004], takes an ANOVA model similar to the one proposed in Kerr and Churchill [2001b] and treats some of the components in the model (such as the array, and experimental batch terms) as random effects. Batch variation was larger than animal to animal variability, and the authors suggest minimising the number of runs in an experiment to reduce this source of error. Unfortunately the estimation of an array component of variance doesn’t help identify poor quality arrays individually.

A number of articles (Sawitzki [2002], Gollub et al. [2003], Smyth et al. [2003] and Petri et al. [2004]) stress the importance of visualising microarray data using diagnostic plots to assess data quality. Spatial plots of the foreground or background intensities, or M or A values for an array can reveal intensity gradients, background problems and systematic red-green imbalances on a slide. MA plots are also useful, allowing detection of spatial artifacts and the need for normalisation. The overall success of a hybridisation can also be assessed; arrays with spot intensities spanning the full dynamic A range are generally better than arrays with a small A range, and the behaviour of the control spots (which can be highlighted on the MA plot) gives a clue as to how well the experimental protocol was followed. Histograms of the R and G measures can be used to assess signal saturation and labelling efficiency.

Packages such as arrayQuality (by Yang et al. [2004]) can generate standardised
diagnostic plots for two-colour microarrays with a single command. By comparing a plot from a new array to plots from previous ‘good’ quality hybridisations, slides which are irregular in some way can be determined and removed from further analysis. One problem with identifying outlier arrays in this way is its subjectivity; different practitioners may flag different arrays to remove from the analysis, which is undesirable.

Finkelstein et al. [2002] observed that arrays with spatial biases were quite common (1 in 5 for their data), and recommends that affected arrays be repeated. Colantuoni et al. [2002] advocates the use of normalisation as a slide quality control procedure to remove systematic variation across the surface of an array due to hybridisation artifacts.

Brown et al. [2001] notes that during image analysis, a measure of the average correlation between the red and green pixels of all spots can be used as an indicator of scan quality. Such a measure can be used to determine when the red and green images are misaligned, which is a common problem with images from certain scanners, and was observed to reduce the average correlation considerably. They also advocate visual inspection, and used this as a basis for removing an array in their weighted analysis of expression across 10 replicate arrays. Where duplicate spots are available on an array Steinfath et al. [2001] uses duplicate correlation as an array quality measure.

For large scale experiments, a statistical process control approach is taken in Model et al. [2002] to detect outlier chips. Experiments made up of 207, 647 and 433 arrays were used to demonstrate this approach. For a data set where repeat hybridisations were available, a robust profile of expression $\bar{m}_j$, was calculated for comparison $j$. Chips which deviated from this profile could be easily seen in a plot of the individual array profile against the robust summary; poorly correlated plots indicate hybridisation problems or other systematic variations. Where replicate data were not available, robust principal components analysis was used to identify outlier arrays. A multivariate distance between chips was calculated using the first $d$ principal components. Outlier arrays which were regarded as ‘unacceptable’ by visual inspection were generally (97%) detected using this method. Statistical process control added the time of hybridisation variable to the analysis, and the $T^2$ control chart made obvious the different batch effects resulting from changes in process variables at different times.
4.2 Array quality weights

For designed experiments with some level of replication, an approach which is more systematic than removing outlier arrays by visual inspection and fits within a linear modelling framework was required. As highlighted in the literature review (4.1), a multitude of factors may affect the variability of the results from a microarray experiment, and many of these can affect all spots on an array equally, such as handling/operator effects, day effects etc. In this chapter, an empirical approach which assigns precisions to arrays by looking at how a given array agrees with other arrays in an experiment will be taken. To do this, a dispersion model with array specific terms is fitted along with the usual mean model (equation 1.1 in section 1.7.1, chapter 1). The final estimates of the array specific variances were converted to weights and used in re-fitting the gene-wise mean models to improve the precision of the estimates of the gene expression coefficients.

In the sections that follow, a brief literature review of joint mean and dispersion modelling will be given. A general model will be specified, and adapted to suit microarray data. To allow efficient estimation of the common array parameters, which are shared by many thousands of observations, a novel gene-by-gene update approach which uses modified Residual Maximum Likelihood (REML) scoring will be described.

4.2.1 Mean and dispersion models

An early example of a mean and dispersion (or variance) model is given in Park [1966]. Harvey [1976] presents a generalisation of this model along with the ML scoring iterations for the mean and variance parameters which are updated together. The simplifications of the iterations under the assumption of normally distributed responses, and the benefits of a multiplicative over an additive heteroscedastic model (negative or zero estimates of the variance parameters are avoided, and the likelihood ratio test has a simpler form) are also discussed.

Cook and Weisberg [1983] present a score test for heteroscedasticity for the same general model, and Aitkin [1987] implements a ML estimation procedure in GLIM (Baker and Nelder [1978]) for the mean and variance parameters when they are functionally independent. This simplifies to using weighted least squares regression to fit the mean
parameters and gamma regression with a log-link function and scale parameter of 2 for the deviances to obtain the variance parameters.

Smyth [1989] extends the model to other distributions (inverse gaussian, gamma) and allows a more general form for the variance model and discusses the possibility of using quasi-likelihoods. Iterations which cycle between estimating the mean parameters while fixing the variance parameters, and the variance parameters while fixing the mean parameters, are found to be effective, since the two sets of coefficients are orthogonal. This issue is explored more fully in Smyth [1996].

Verbyla [1993] gives the REML scoring iterations for the model of Harvey [1976], and compares the estimates with those obtained using ML. REML estimation is preferred as it uses an unbiased score for the estimation of the variance parameters, and is also found to be more robust to outliers than the ML estimators. More generally, REML estimators of the variance parameters are known to be less biased, as the likelihood of a set of zero mean contrasts is maximised, which adjusts the degrees of freedom for the mean parameters already estimated.

In Smyth and Verbyla [1999], approximate REML is used to fit double GLMs for the mean and dispersion model simultaneously in S-Plus, using a $\chi^2(1)$ approximation of the deviances and a saddle-point approximation to the exponential dispersion model.

Smyth et al. [2001] compares exact and approximate REML for heteroscedastic linear models, to find that using double GLMs may not give correct standard errors for the variance coefficients. Smyth [2002] suggests an efficient algorithm for exact REML, using Levenberg-Marquardt damping of the information matrix to ensure convergence and efficient computation.

The theory and estimation methods discussed in these papers will be used and adapted in the sections that follow.

4.2.2 The underlying model

The model described in Smyth et al. [2001] will be outlined in this section. Assume we have $l = 1, \ldots, r$ identically and independently distributed random variables, with
\( y_t \sim N(x_T^t \beta, \sigma^2_t / w_t) \). A model for the mean is

\[
E(y_t) = \mu_l = x_T^t \beta
\]

(4.1)

and for the variance

\[
g(\sigma^2_t) = \log(\sigma^2_t) = z_T^t \gamma
\]

(4.2)

where the prior weights are \( w_l \), and the link function \( g \) is \( \log \). The vectors \( x_T^t \) and \( z_T^t \) are the \( l \)th rows of the full rank matrices \( X \) and \( Z \) (which are \( r \times s \) and \( r \times t \), respectively), containing covariates for predicting the mean and variance, and \( \beta \) and \( \gamma \) are \( s \times 1 \) and \( t \times 1 \) vectors of the coefficients of interest. The ML score vector and Fisher expected information are

\[
U_{\text{ML}}(\beta, \gamma ; y) = \begin{pmatrix} X^T \Sigma^{-1} (y - X\beta) \\ \frac{1}{2} Z^T (\Sigma^{-1} d - 1_r) \end{pmatrix}
\]

and

\[
I_{\text{ML}}(\beta, \gamma) = \begin{pmatrix} X^T \Sigma^{-1} X & 0 \\ 0 & \frac{1}{2} Z^T Z \end{pmatrix}
\]

(4.3)

where \( \Sigma \) is a \( r \times r \) matrix with \( l \)th diagonal element \( \sigma^2_t / w_t \), \( y \) is the vector of response variables, \( d \) is the \( r \)-vector of elements \( d_l = w_l(y_l - \mu_l)^2 \) and \( 1_r \) is an \( r \times 1 \) vector of unit elements. The \( p \)th iteration of the method of scoring is

\[
\hat{\beta}_{(p)} = (X^T \Sigma_{(p-1)} X)^{-1} X^T \Sigma_{(p-1)}^{-1} y
\]

\[
\hat{\gamma}_{(p)} = \gamma_{(p-1)} + (Z^T Z)^{-1} Z^T (\Sigma_{(p-1)}^{-1} d - 1_r)
\]

= \((Z^T Z)^{-1} Z^T (\Sigma_{(p-1)}^{-1} d - 1_r + Z \gamma_{(p-1)})\).

These iterations are equivalent to performing normal weighted least squares regression for \( \beta \), with weights \( w_l / \sigma^2_t \) while the iterations for \( \gamma \) are the same as that for a gamma GLM with dispersion parameter 2 (since \( d_l \sim \sigma^2_t \chi^2(1) \)) and link function \( g \) is \( \log \).

Verbyla [1993] derived the REML marginal likelihood for \( \gamma \) for the same model as

\[
\ell_{\text{REML}}(\gamma ; y) = -\frac{1}{2} (\log |\Sigma| + y^T P y + \log |X^T \Sigma^{-1} X|) + \frac{1}{2 \sigma^2_t} \chi^2_{(1)}
\]
where \( P = \Sigma^{-1} - \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1} \) and the score vector and Fisher expected information for \( \gamma \) are

\[
U_{\text{REML}}(\gamma ; y) = \frac{1}{2}Z^T(\Sigma^{-1}d - 1_r + h)
\]

and

\[
I_{\text{REML}}(\gamma) = \frac{1}{2}Z^TVZ
\]

respectively. Here \( h \) is an \( r \times 1 \) vector made up of the diagonal elements from the hat matrix,

\[
H = \Sigma^{-1/2}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1/2}
\]

and \( V \) is an \( r \times r \) matrix with diagonal elements \( (1 - h_{mm})^2 \) and off diagonals of \( h_{mn}^2 \).

The corresponding REML scoring iterations for \( \gamma \) are

\[
\hat{\gamma}(p) = \gamma(p-1) + I_{\text{REML}}(\gamma(p-1))^{-1}U_{\text{REML}}(\gamma(p-1) ; y) \]

\[
= \gamma(p-1) + (Z^TVZ)^{-1}Z^T(\Sigma(p-1)d - 1_r + h).
\]

Approximate REML replaces \( V \) with a diagonal approximation \( V^* \), which has diagonal elements \( (1 - h_{mm})^q \) and off diagonals of 0. Smyth et al. [2001] shows that \( q = 1 \) often gives the best results amongst diagonal approximations to \( V \). Sufficient iterations are performed until convergence is reached.

### 4.2.3 Tailoring the model for use on microarray data

Adapting the mean and variance models (equations 4.1 and 4.2) to suit microarray data is straightforward. Recall the gene-wise linear models (equation 1.1 in section 1.7.1) of the form

\[
E(y_i) = \mu_i = X_{\text{gene}}\beta_i
\]

where \( y_i^T = (M_{i1}, \ldots, M_{iJ}) \) is a response vector of log-ratios (which may contain missing values), \( X_{\text{gene}} \) is the design matrix of full column rank, \( \beta_i \) is a \( K \times 1 \) coefficient vector and

\[
\text{var}(y_i) = W_i\sigma_i^2
\]

59
where $W_i$ is a known non-negative definite weight matrix. For the complete data the linear model for the mean would be specified as

$$E(y) = X\beta$$

(4.4)

where $y$ is a vector of the log-ratios for all genes ($i = 1, \ldots, I$) across $J$ arrays, i.e., $y^T = (M_{11}, \ldots, M_{1J}, \ldots, M_{I1}, \ldots, M_{IJ})$. The design matrix, $X$ for the complete data would have dimension $(J \times I) \times (K \times I)$ and a block diagonal structure,

$$X = \left( \begin{array}{cccc}
X_{\text{gene}} & 0 & \ldots & 0 \\
0 & X_{\text{gene}} & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & X_{\text{gene}}
\end{array} \right)$$

since the coefficients for each gene, $\beta_i$ are not shared. The vector of mean coefficients would be $\beta^T = (\beta_1^T, \ldots, \beta_I^T)$. A log-linear model for the variances would include both gene ($\delta_i$) and array ($\gamma_j$) terms such that

$$g(\sigma_{ij}^2) = \log(\sigma_{ij}^2) = \delta_i + \gamma_j.$$ 

Applying sum to zero constraints on the $\gamma_j$’s gives the gene-wise variance $\exp(\delta_i)$. This model allows each array to have a multiplicative effect $\exp(\gamma_j)$ in the variances, which is intended to summarize the overall quality or reliability of that array.

The design matrix, $Z$ for the variance model (equation 4.2) has a regular structure
too:

\[
Z = \begin{pmatrix}
Z_1 \\
\vdots \\
Z_J
\end{pmatrix}
= \begin{pmatrix}
\begin{pmatrix}
1 & 0 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
1 & 0 & \ldots & 0
\end{pmatrix}
& Z_{\text{array}} \\
0 & 1 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 1 & \ldots & 0 \\
0 & 0 & \ldots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & 1
\end{pmatrix}
\]

where \( Z_i \) is a \( J \times I \) matrix of 0’s except for the \( i \)th column which consists of 1’s, and

\[
Z_{\text{array}} = \begin{pmatrix}
1 & 0 & \ldots & 0 \\
0 & 1 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & 1 \\
-1 & -1 & \ldots & -1
\end{pmatrix}
\]  

is a \( J \times (J - 1) \) matrix with a standard form when sum to zero constraints are placed on the array parameters (i.e. \( \gamma_J = -\sum_{j=1}^{J-1} \gamma_j \)). The vector of \( (I + J - 1) \) coefficients for the variance model would be \( \gamma^T = (\delta_1, \ldots, \delta_I, \gamma_1, \ldots, \gamma_{J-1}) \).

In principle this heteroscedastic model can be fitted to the complete data using either ML or REML scoring iterations as described in section 4.2.2. However, with between 10,000 and 30,000 spots on each array and multiple arrays making up an experiment, the design matrices \( X \) and \( Z \) and variance-covariance matrix \( \Sigma \) are too large to be stored in the memory of a standard computer. This presents a problem, since matrix operations are required to estimate the parameters \( \beta \) and \( \gamma \).

The mean parameters (\( \beta \)) as well as the gene variance terms (\( \delta_i \)) are nuisance parameters, from the point of view of estimating the common array variances (\( \gamma_j \)) and are parameterised to be orthogonal or approximately orthogonal to the array variance terms.
In this context, orthogonality refers to having off-diagonals elements of zero in the information matrix, which has been shown to be the case for the mean and variance parameters (equation 4.3). When using approximate REML to update the variance parameter estimates, the off-diagonals $h_{mn}$ of $V^*$ are set to zero, hence the variance parameters ($\delta_i$’s and $\gamma_j$’s) are only approximately orthogonal to each other.

To sensibly estimate these shared parameters, all of the genes can be used in a gene-by-gene update algorithm which exploits the orthogonality of the parameters. This approach will be described in the following section.

### 4.2.4 A gene-by-gene update algorithm

By convenient parameterisation, the REML scoring iterations described in section 4.2.2 can be made gene specific, and rather than iterating until convergence in the usual way, we can iteratively update the estimates of the shared variance parameters gene-by-gene, finishing after $I$ iterations when we have run through all the genes.

An appropriate gene-specific variance model would be

$$g(\sigma_i^2) = Z_{\text{gene}} \gamma_i,$$

where $\gamma_i^T = (\delta_i, \gamma_1, \ldots, \gamma_{J-1})$ is the vector of coefficients, $Z_{\text{gene}}$ is the design matrix, and $g$ is a log-link function. The variance model was parameterised using sum to zero contrasts, with array variance parameters $\gamma_{\text{array}}^T = (\gamma_1, \ldots, \gamma_{J-1})$, where $\gamma_J = -\sum_{i=1}^{J-1} \gamma_i$, relative to the gene specific parameter, $\delta_i$, which gives a $J \times J$ design matrix,

$$Z_{\text{gene}} = \begin{pmatrix}
1 \\
\vdots \\
Z_{\text{array}} \\
1
\end{pmatrix},$$

for the dispersion model, where $Z_{\text{array}}$ has been defined previously (matrix 4.5). Since the gene variance term, $\delta_i$ is a nuisance parameter, which is approximately orthogonal to the array variance terms, we can use the reduced design matrix $Z_{\text{array}}$ in our iterations.

For a given $\gamma_{\text{array}}$, weighted least squares can be used to fit the mean model (equa-
tion 1.1) using weights $w_{ij}/\sigma_{ij}^2$, where $\sigma_{ij}^2 = \exp(z_iT \gamma_{array})$ to give an estimate of the mean parameters ($\hat{\beta}_i$) and variance $\hat{\sigma}_i^2$ for gene $i$. The gene specific term in the variance model ($\delta_i$) is available in closed form as $\delta_i = \log\hat{\sigma}_i^2$, which also depends on $\gamma_{array}$. The array specific parameters ($\gamma_{array}$) can be updated using the $i$th gene by fitting the variance model (equation 4.6) to the squared residuals ($d_{ij} = w_{ij}(y_{ij} - \mu_{ij})^2$) from the mean fit. Using a REML scoring approach, the score vector

$$U_{REML(i)} = \frac{1}{2} Z_{array}^T (\Sigma_i^{-1} d_i - 1_J + h_i)$$

is calculated for each gene. The vector, $h_i$ is the $J$-vector containing the diagonal elements from the hat matrix

$$H_i = \Sigma_i^{-1/2} X_{gene}(X_{gene}^T \Sigma_i^{-1} X_{gene})^{-1} X_{gene}^T \Sigma_i^{-1/2},$$

from the regression of $\beta_i$, where $\Sigma_i$ has $\sigma_{ij}^2/w_{ij}$ as its $j$th diagonal element and $1_J$ is a unit vector of length $J$. The expected (Fisher) information matrix has the form

$$I_{REML(i)} = \frac{1}{2} Z_{array}^T V_i Z_{array},$$

in general, where $V_i$ has diagonal elements $(1 - h_{mm})^2$, and off diagonal elements $h_{mn}^2$ from the hat matrix, $H_i$. Here we will use approximate REML and replace $V_i$ with $V_i^*$, which has diagonal elements $(1 - h_{mm})$ and off diagonal elements zero (as suggested in Smyth et al. [2001]). A modified version of the information, $I_{REML(i)}^*$ which was allowed to accumulate on each iteration, defined as

$$I_{REML(i)}^* = \frac{1}{2} \sum_{u=1}^{i} Z_{array}^T V_i^* Z_{array}^T$$

was used in the modified scoring iterations.

The update step for $\hat{\gamma}_{array}$ is

$$\hat{\gamma}_{array(i+1)} = \gamma_{array(i)} + (I_{REML(i)}^*)^{-1} U_{REML(i)}$$

$$= \gamma_{array(i)} + (\sum_{u=1}^{i} Z_{array}^T V_i^* Z_{array})^{-1} Z_{array}^T (\Sigma_i^{-1} d_i - 1_J + h_i).$$

63
The update algorithm was repeated $i = 1, \ldots, I$ times. An initial value of $\gamma_{\text{array}(0)} = 0$, i.e., equal array variances was assumed, and the initial information $I_{\text{REML}(1)}^* = 10Z_{\text{array}}^T V_1^# z_{\text{array}}$, where $V_1^#$ is a matrix similar to $V_1^*$, except with the leverages calculated using a variance-covariance matrix with equal weights, assuming equal variances and no missing values (i.e. $H_1 = X_{\text{gene}}(X_{\text{gene}}^T X_{\text{gene}})^{-1} X_{\text{gene}}^T$). This ensured the stability of the initial iteration. Summing over the past information matrices has the effect of decreasing the step size of the iterations as the number of genes used increases.

The final array variance estimates ($\hat{\sigma}_{Ij}^2 = \exp(z_{\text{array}}^T \gamma_{\text{array}(I)})$) were used in the gene-wise linear models with the prior weights ($w_{ij}$) to obtain estimates of the gene expression coefficients. $t$ statistics were calculated using the coefficients and their SEs. The R code for calculating the $\gamma_{\text{array}}$ terms is included as Appendix A.

### 4.2.5 An example of the update algorithm

Suppose we have three microarrays ($j = 1, 2, 3$) with 5 genes on each array ($i = 1, 2, 3, 4, 5$), with the same mRNA samples labelled red and green on arrays 1 and 3, and reversed (dye-swapped) on array 2. Assume equal prior weights ($w_{ij} = 1$). Since there are only two sources of mRNA, the design matrix for the mean model will be

$$X_{\text{gene}} = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix},$$

where the coefficient, $\beta_i$ is a scalar quantity representing the (weighted) mean of the log-ratios. For the linear model for the variances, the design matrix is

$$Z_{\text{gene}} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} \text{\textbf{1}} \\ z_{\text{array}} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \end{pmatrix},$$

and the variance parameters are $\gamma_I^T = (\delta_1, \gamma_1, \gamma_2)$, where $\gamma_3 = -\gamma_1 - \gamma_2$ because of the sum to zero constraints. The second and third columns of $Z_{\text{gene}}$ are relevant for our iterations,
so we have $\gamma_{array}^T = (\gamma_1, \gamma_2)$ and the reduced design matrix

$$Z_{array} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix}.$$ 

Starting with an initial estimate of $\gamma_{array(0)}^T = (0, 0)$, gives the variance-covariance matrix

$$\Sigma_1 = \begin{pmatrix} \sigma_{11}^2/w_{11} & 0 & 0 \\ 0 & \sigma_{12}^2/w_{12} & 0 \\ 0 & 0 & \sigma_{13}^2/w_{13} \end{pmatrix} = \begin{pmatrix} e^0 & 0 & 0 \\ 0 & e^0 & 0 \\ 0 & 0 & e^0 \end{pmatrix} = I,$$

which gives weights $w_{ij}/\sigma_{ij}^2 = 1$ to use in the weighted least squares regression of $y_1^T = (y_{11}, y_{12}, y_{13})$ on $X_{\text{gene}}$ to obtain $\hat{\beta}_1$ and $\hat{\sigma}_1^2$. The squared residuals

$$d_1 = \begin{pmatrix} (y_{11} - \hat{\beta}_1)^2 \\ (y_{12} + \hat{\beta}_1)^2 \\ (y_{13} - \hat{\beta}_1)^2 \end{pmatrix},$$

along with the matrix

$$V_1^# = \begin{pmatrix} 1 - h_{11} & 0 & 0 \\ 0 & 1 - h_{22} & 0 \\ 0 & 0 & 1 - h_{33} \end{pmatrix},$$

and vector $h_{1}^T = (h_{11}, h_{22}, h_{33})$ are used in the scoring iterations for the array variance parameters. The elements $h_{mm}$ are the diagonal entries from the hat matrix $H_1 = X_{\text{gene}}(X_{\text{gene}}^TX_{\text{gene}})^{-1}X_{\text{gene}}^T$.

The first iteration is thus

$$\hat{\gamma}_{array(1)} = \hat{\gamma}_{array(0)} + (10Z_{array}^T V_1^# Z_{array})^{-1}Z_{array}^T (\Sigma_1^{-1}d_1 - I_3 + h_1),$$
For $i = 2$, we update the variance-covariance matrix,

$$
\Sigma_2 = \begin{pmatrix}
\sigma_{21}^2/w_{21} & 0 & 0 \\
0 & \sigma_{22}^2/w_{22} & 0 \\
0 & 0 & \sigma_{23}^2/w_{23}
\end{pmatrix} = \begin{pmatrix}
e^{\hat{\gamma}_1(1)} & 0 & 0 \\
0 & e^{\hat{\gamma}_2(1)} & 0 \\
0 & 0 & e^{-\hat{\gamma}_1(1) - \hat{\gamma}_2(1)}
\end{pmatrix},
$$

and estimate the coefficient $\hat{\beta}_2$ and variance $\hat{\sigma}_2^2$ via weighted least squares regression of $y_2^T = (y_{21}, y_{22}, y_{23})$ on $X_{\text{gene}}$ using the inverse diagonal elements of $\Sigma_2$ as the weights. After $d_2$, and $h_2$ and $V_2^*$ from

$$
H_2 = \Sigma_2^{-1/2} X_{\text{gene}} (X_{\text{gene}}^T \Sigma_1^{-1} X_{\text{gene}})^{-1} X_{\text{gene}}^T \Sigma_1^{-1/2}
$$

are formed, the next iteration is

$$
\hat{\gamma}_{\text{array}(2)} = \hat{\gamma}_{\text{array}(1)} + (10 Z_{\text{array}}^T V_1^* Z_{\text{array}} + Z_{\text{array}}^T V_2^* Z_{\text{array}})^{-1} Z_{\text{array}}^T (\Sigma_2^{-1} d_2 - 1_3 + h_2).
$$

The iterations are repeated for genes $i = 3, 4, 5$, with the information matrix $I_{\text{REML}}^*$ accumulating with each gene. From the final parameter estimates, $\hat{\gamma}_{\text{array}(5)}$, the array variances, $\sigma_{2i}^2 = (e^{\hat{\gamma}_1(5)}, e^{\hat{\gamma}_2(5)}, e^{-\hat{\gamma}_1(5) - \hat{\gamma}_2(5)})$ are used to re-fit each of the gene regressions (equation 1.1) to get final estimates of the gene expression coefficients, which are then used to calculate $t$ statistics and assess differential expression.

### 4.2.6 Simulated data

A simulated data set was used to test the ability of the algorithm described in section 4.2.4 to recover known array variances ($\gamma_j$’s). Log-ratios from 10 arrays ($j = 1, \ldots, 10$), with $I = 10,000$ genes were simulated in R using the `rnorm` function. Individual gene means ($\mu_i$), and gene variance parameters ($\delta_i$) were drawn from $N(0, 1)$ distributions, while array variances ($\gamma_j$) were drawn from a U[-1,1] distribution and subject to sum to zero constraints. The log-ratios ($M_{ij}$) were drawn from $N(\mu_i, \exp(\delta_i + \gamma_j))$ distributions.

Figure 4.1 shows the estimated versus actual array variance parameters obtained from the gene-by-gene update algorithm using 100, 1,000 or all 10,000 genes in the iterations. For $I = 100$ genes, the relative array variances aren’t very well estimated (correlation coef-
ficient between estimated and actual values of 0.885), with only 5/10 of the $\gamma_j$’s correctly ordered. As more genes are used ($I = 1,000$ and $I = 10,000$), the estimates improve considerably (correlation coefficients of 0.968 and 0.999 respectively), and the ordering of the estimated parameters is correct for 5/10 arrays for 1,000 genes and 10/10 when all 10,000 are included in the iterations. Figure 4.2 shows the estimate for each array variance parameter ($\hat{\gamma}_j$) at each gene update ($i$) iteration. After iterating through $I = 100$ and 1,000 genes, the estimates tend not to have converged, however after using all 10,000 genes, the estimates recovered are close to the true values.

Microarrays generally contain more than 10,000 genes (in the case of the QC and METH data sets $I = 10,368$), which ensures this approach of estimating the array variances is a practical one.

### 4.3 Results

The calculation of the array variance parameters takes around 3 minutes on a standard desktop computer for the data sets analysed. In the following sections, the appropriateness of the resulting weights will be addressed.

To answer the question ‘do microarrays with obvious defects receive less weight than other microarrays?’ qualitative information on the arrays, such as intensity and background levels over the slide, and how well the printing was carried out will be used.

When array quality weights are used in the gene-wise linear models, their ability to reduce the variability of the estimated gene expression coefficients will also be assessed, by focusing on the ordinary $t$ statistics from the coefficients and the SEs with and without array quality weights. The ‘truth’ available through the spike-in controls will be exploited to determine if improvements in our ability to detect known differential expression are made.

#### 4.3.1 Array weights for the QC data set

The QC data set, which is known to vary in quality, is a useful collection of arrays for testing the array weighting methodology. The QC arrays are from 22 separate print
Figure 4.1: Estimated ($\hat{\gamma}_j$) versus actual ($\gamma_j$) array variance parameters from the simulated data set using $I = 100, 1,000$ and $10,000$ genes and $J = 10$ microarrays.
Figure 4.2: Estimated array variance parameter ($\hat{\gamma}_j$, after iterating through $i$ genes) versus gene number ($i$) for each array from the simulated data set. The number of genes used ($I = 100, 1,000$ and $10,000$) are colour-coded (blue, green and red respectively). The estimates appear to have converged to the true value, indicated by the black dashed line, after $I = 10,000$ iterations.
batches and were hybridised over an extended period by several experimenters. These factors ensure the measurements from these arrays are subject to many different sources of variability.

The final estimates for the array weights \( \exp(-\mathbf{z}_{\text{array}}^T \hat{\gamma}_{\text{array}(i)}) \) for the QC data set obtained using Spot image analysis output are shown in Figure 4.3. For this data set, arrays with poorly printed spots (HD45, HJ45, HJ75) received the lowest weights (0.64, 0.53 and 0.54 respectively). On closer inspection, it can be seen that these arrays have spots printed over the top of each other in each grid (see figure 4.4 for an example), which creates problems for the segmentation. Each grid is short by one row, which results in a lot of mis-addressed spots.

Arrays with low signal, such as HE60, HR75 and HW75 were also assigned lower weights of 0.83, 0.91 and 0.83 respectively. A visual inspection of the remaining arrays reveals no obvious spatial artifacts or background problems, which is consistent with these arrays mostly receiving relatively higher weights than the arrays discussed. In particular, arrays from the HN and HQ batches were visually some of the best arrays in the series, with clearly printed spots, good signal over the entire array and low background, and were assigned amongst the highest weights. Although some variation in the signal on the remaining arrays is evident, this obviously doesn’t affect the variability of the log-ratios.

Observations from low weighted arrays have less influence in the re-fits of the linear models, which should improve the precision of the gene expression estimates \( \hat{\beta}_{ik} \), and enhance our ability to detect differential expression. To examine the extent to which this occurs, the ordinary t statistics \( \hat{\beta}_{ik}/\text{SE}(\hat{\beta}_{ik}) \) which are used to assess differential expression, were plotted before and after applying array weights. Figure 4.5 shows that the \(|t|\) for the most highly DE genes are slightly more extreme after applying array quality weights compared to using equal weights. As the right panel shows, this is caused by improved precision in estimating the coefficients, with the ratio of SEs with:without weights decreasing by 3% on average. This represents a small gain in power, and is an indication that the weights are appropriate (i.e. array weights are inversely proportional to the variability).

Figure 4.6 shows the effect of array weights on the t statistics of the LMS controls. The array variances were estimated using the controls from 100 out of the 111 arrays, for which the spike-in controls had been added, and the t statistics were calculated with equal
Figure 4.3: Relative array weights for the QC data set obtained using Spot image analysis output. The relative weights were calculated using 10,368 genes from 111 arrays which were manufactured and hybridised in 22 batches. The arrays are ordered by batch, with bars of the same colour belonging to the same batch. Microarrays HD45, HJ45 and HJ75 received the lowest weights (0.64, 0.53 and 0.54 respectively).

weights, and relative array weights. The median $t$ statistics for the DE controls (D03, D10, U03 and U10) all increase in absolute terms when log-ratios from suspect arrays are down-weighted in the analysis, which shows improvement in power to detect differential expression. The $t$ statistics for the DR, which are not DE, do not change.

4.3.2 Array weights for the METH experiment

The experimenters who conducted the METH experiment were suspicious about the reliability of the first 4 arrays in the experiment, which they believed weren’t giving consistent results with the last 6 arrays. Figure 4.7 shows the array weights assigned to this experiment. Arrays 1 and 4, which were assigned the lowest weights (0.78 and 0.38 respectively),
Figure 4.4: An example of poorly printed grids from array HD45 from the QC data set. Grids 19 and 20 are shown in this picture as a RGB overlay of the Cy3 and Cy5 images (top) and after segmentation, with the SRG masks shown for each spot (bottom). Spots in rows 7 and 8 (indicated with arrowheads) are printed over the top of each other in both grids, and the spots in rows 1 to 7 are smudged in grid 20. This causes problems during image analysis with about 1/3 of the spots mis-addressed.
Figure 4.5: Difference in ordinary \( t \) statistics calculated with array weights and with equal weights versus \( t \) statistics with equal weights (left) and the ratio of the SEs with:without array weights (right).

Figure 4.6: Ordinary \( t \) statistics calculated with and without array weights for the LMS controls from the QC data set.
both suffered from high background in the green and red channels (see figure 4.8, left panel). This is known to introduce noise into the log-ratios. Array 4 also received the most ‘not found’ flags (36% of spots) from GenePix (see figure 4.8, right panel), which is a further indication that this slide is problematic. It is worth noting that the arrays were hybridised in pairs (1 with 2, 3 with 4 and so on), so different handling effects within each batch are the cause of the high background problem.

When these weights are used in the analysis, slightly more genes are suggested as DE for a given threshold for the $p$ values (obtained from the $t$ statistics and adjusted for multiple testing using the FDR method of Benjamini and Hochberg [1995]). Table 4.2 shows the number of genes ranked as DE for the 1 vs 0 and 3 vs 0 contrasts, which are of most interest. For the 1 vs 0 comparison, which has two poor quality arrays directly comparing these mRNA sources, removal of the poor quality arrays throws away most of the information on this comparison, and results in only 1 DE gene. Treating the observations from different arrays equally (i.e. equal weights) suggests 62 genes are DE for the 1 vs 0 contrast. Using relative array weights gives the most candidate DE genes with 97 for this comparison, all of which are considered DE in the 3 vs 0 comparison,
Figure 4.8: The log of the Cy3 (green) and Cy5 (red) background intensities for each slide from the METH experiment (left, top and bottom). The percentage of spots on each array flagged by GenePix as ‘not found’ is shown on the right.

Table 4.2: Number of DE genes for the METH experiment obtained using different weighting methods. Alternative 1 represents a typical analysis where the data from each array is treated equally. The filtering alternative (2) removes the poor arrays (1 and 4) and approach 3 uses the array weights from Figure 4.7 in the analysis. Using array weights gives a few more candidate DE genes for a given p-value cut-off (p < 0.05) compared to the other alternatives.

<table>
<thead>
<tr>
<th>Weighting Method</th>
<th>Number of DE genes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 vs 0</td>
</tr>
<tr>
<td>1. Equal array weights (w_j = 1)</td>
<td>62</td>
</tr>
<tr>
<td>2. Filtering arrays (remove arrays 1 and 4)</td>
<td>1</td>
</tr>
<tr>
<td>3. Array weights</td>
<td>97</td>
</tr>
</tbody>
</table>
Figure 4.9: The relative array weights for the high quality DC experiment. All arrays receive equivalent weights ($\approx 1$).

which would be expected since the same drug is administered. Also the majority (59% or 57) of these genes respond in a predictable way, being DE at a higher level in the 3 vs 0 comparison compared to the 1 vs 0 comparison. This suggests that the ‘rescued’ genes are worthy candidates for further validation, and demonstrates how using array weights can help in the search for DE genes.

4.3.3 Array weights for the DC experiment

As a kind of ‘positive’ control, the performance of this method on a good quality data set was also examined. The DC experiment is made up of 6 arrays, which are all of a very high standard, with low background, well behaved LUS controls, and a good dynamic range of spot intensities. The estimated variances are roughly equal for these arrays, giving rise to equivalent array weights (refer to figure 4.9) which is consistent with our visual assessment of the slides. Thus in situations when the arrays are of high quality, the log-ratios are treated equally, which is equivalent to the default linear model analysis where no weights are specified.
4.4 Discussion

In this chapter, examples of array quality weights which give less influence to the log-ratios from microarrays which suffer from hybridisation or printing problems and relatively more influence to the log-ratios from arrays with good signal and low background in the gene-wise linear models have been presented. The use of such weights allows the gene-wise regression coefficients $\beta_i$ to be estimated with less error, which improves our ability to detect highly DE genes using $|t|$ statistics. For a given cut-off, $C$ for which genes with $|t| > C$ are deemed to be probably DE, such a change represents an increase in the statistical power for detecting differential expression.

The computation of the array weights using the gene-by-gene update method is quite efficient, taking a few minutes per data set on a standard desktop computer. It has also been shown that for simulated data, the variance parameters obtained by the algorithm are well estimated when sufficient genes (at least several thousand) are available on each array. Since a typical microarray contains tens of thousands of genes, the gene-by-gene update method for estimating the array variance parameters is a practical one.

Another strength of the method, is its use of log-ratios only to derive array quality weights, which makes it applicable to two-colour microarray data from any image analysis package. This method could equally be applied to the log-intensity data from single-channel gene expression technologies such as Affymetrix. Missing values also present no problems, although arrays with many missing values may receive lower weight than those with fewer missing values.

One further topic that deserves some discussion is the use of robust linear models to estimate the gene expression coefficients. Lloyd et al. [2003] has shown that for small factorial experiments ($J = 8$), fitting the gene-wise linear models using robust MM-estimation compared to least squares can give misleading results when the tuning constants aren’t appropriate, with systematically downward biased SEs. While robust methods are well suited to large sample problems, most microarray data sets such as the METH and DC experiments consist of a small number of arrays, and in these situations, robust methods are probably not suitable. It is also worth noting that for the large QC data set ($J = 111$), the results in terms of the $|t|$ of the DE control genes obtained by robustly fitting the linear
models were equivalent to those obtained using array weights.

The algorithm described in this chapter could be easily incorporated in existing microarray analysis software such as limma for routine use. The method is however restricted for use on data from designed experiments which include some level of replication. A minimum of 3 arrays, due to the sum to zero constraints imposed on the array variance parameters is also necessary. This method will also be of little benefit if all arrays in an experiment are poor quality.

A more fine grade approach, which uses the same methodology to fit print-tip group variances to the model instead of a single array wide variance term is a further possibility. Such parameters may be useful for characterising local spatial variations on slides, or variability introduced by the occasional misalignment of a grid. It is worth noting that the estimates obtained from the gene-by-gene update algorithm may not be stable in this situation, since the number of iterations will now be reduced to the number of spots in a print-tip group, which typically numbers in the hundreds of spots (18 x 24 = 432 in the case of the arrays used in the QC and METH experiments). With a smaller number of genes, it may also be possible to fit the heteroscedastic linear model in usual way (see Smyth et al. [2001]).

Of particular interest in this chapter, has been the novel gene-by-gene update procedure, which uses modified REML scoring with accumulating information to estimate the array variance parameters in the log-linear dispersion model. This represents an adaptation of an established approach for fitting GLMs which may be useful in other settings where shared parameters need to be estimated in large data sets.
Chapter 5

Spot Quality Weights

In this chapter, the use of prior spot weights derived from various spot specific measures in the analysis of log-ratios from designed microarray experiments will be demonstrated. The chapter begins with a literature review of current approaches that deal with spot quality (section 5.1) before looking at the methodology used to derive several different weighting schemes (section 5.2). In section 5.3, the methods are compared on various two-colour microarray data sets by examining their effect in the linear model fits. The effectiveness of the weighting schemes is demonstrated on known DE control genes, which are generally easier to detect when spot quality weights are used in the analysis.

5.1 Literature review

Most microarray analysis papers address the importance of performing some kind of data quality control at the spot and/or array level. This issue is of particular importance when there is limited or no replication of an experiment, and decisions about which genes are DE are to be made.

The spot quality methods in the literature can be broadly grouped into two categories. One class of methods use spot specific measures such as size, intensity and signal-to-noise ratio (SNR) to rule out unreliable spots; a graduated version along these lines will be described in this chapter. An alternate approach, which was used at the array level to assign the weights in chapter 4, involves looking at the reproducibility of the log-ratios
directly to determine outliers. In nearly all papers, spots judged as poor quality by such methods are removed from further analysis. This is commonly referred to as ‘filtering’ in the microarray literature.

5.1.1 Methods based on spot specific measures

In image analysis packages such as GenePix, spots can be flagged either manually after visual inspection, or based on automatic criteria. Removal of low intensity spots from further analysis is one commonly used criterion (see for instance Beißbarth et al. [2000], Yang et al. [2001a] or Fan et al. [2004]) motivated by the high variability of log-ratios at low intensities. For example, Fan et al. [2004] discards spots with $A < 6$ on a slide by slide basis. As discussed in chapter 3, this ‘fanning’ of log-ratios is closely related to the background correction method used, and where possible our analysis will be carried out in a way which avoids this.

Yang et al. [2001a] eliminated spots whose foreground intensity was less than $c$ times the background variation in either channel. A $c$ value of 8 was used and the background variation was defined as the standard deviation of the differences between each spot’s measured background and the backgrounds of its 8 nearest neighbours.

Tran et al. [2002] used the ‘correlation’ between the mean and median foreground intensity to filter unreliable spots. They argue that for a good spot, the mean and median intensities of the foreground pixels should be similar within each channel. Deviations, indicating a skewing of the foreground pixel intensity distribution, can be used to flag bad spots which are irregular in some way (such as ‘donut’ or other poorly printed spots). In this paper, ‘correlation’ is the ratio of the raw mean and median foreground (smallest over largest) for each channel. If a spot’s ratio is less than 85% in either channel, it is excluded from further analysis. The authors found the mean-median ‘correlation’ to be superior to the default GenePix image analysis flags, as spots of low intensity, and some variable high intensity spots were removed.

Finkelstein et al. [2002] found little consensus between replicate experiments, which led them to conclude that single experiments are inadequate for measuring reproducibility and assessing quality. They discuss an automatic flagging procedure based on correlation
of the red and green pixel intensities for each spot (also suggested in Brown et al. [2001]),
but conclude that this method may give high correlations to poorly printed spots with
uneven shapes, or low correlations when images are mis-aligned. An alternative method
suggested was to apply a threshold to the standard deviations of the foreground pixels, to
remove spots with highly variable foreground pixel intensities.

Jenssen et al. [2002] found that within-slide data filtering based on fixed cut-offs (such
as rejecting spots whose foreground intensity < 1.4 times their background intensity)
performed poorly when used blindly across all slides in several data sets. After filtering
spots flagged by the GenePix software and those whose foreground was less than or equal
to the background intensity, a variant of SNR called the SB ratio, which consists of the spot
intensity divided by background intensity for each channel, was used as a further filtering
criterion. Using a fixed threshold of 0.43 for the ‘repeatability coefficient’ (2.83 x \sqrt{\text{MSE}}),
a cut-off for the SB ratio is adaptively chosen for each data set.

The first paper which attempted to combine a number of spot specific quality scores
into a single indicator of quality is Wang et al. [2001]. In this paper, information about a
spot’s size, its signal-to-noise ratio (SNR), the level and variability of its local background
and the amount of foreground pixel saturation are combined to give a composite quality
score, \( q_{\text{com}} \). Low scores were indicative of low quality, and spots below a chosen threshold
(such as \( q_{\text{com}} < 0.3 \)) were eliminated from further analysis. Hessner et al. [2003a] and
Hessner et al. [2003b] used the \( q_{\text{com}} \) quality scores to remove unreliable spots. In a follow
up paper (Wang et al. [2003]), \( q_{\text{com}} \) based loess normalisation is performed to remove
biases from the data. Using microarray data produced at the WEHI, \( q_{\text{com}} \) was found to be
closely related to spot intensity (A, data not shown). The need for log-ratios to be
normalised according to their \( q_{\text{com}} \) (analogous to intensity based normalisation) is further
evidence of this relationship.

Buhler et al. [2000] categorised spots into ‘accept’, ‘reject’ or ‘show’ classes as part of
the image analysis process using a linear classifier. Implemented in the program Dapple,
background brightness (b-score) and position of a spot from the center of its bounding
rectangle (p-score) are considered in the classifier. A training data set where the spots
had been assigned to each group after visual inspection was used to build the classifier.
The classifier performed well on the 6 arrays tested, with a maximum mis-classification
rate of 2.05% observed for these data.

Hautaniemi et al. [2003] proposed a method which used 14 spot quality features (bleed-
ing, size, roundness, alignment error, as well as foreground and background intensity and
background noise in the Cy3 and Cy5 channels) to derive a Bayesian network model of
quality. The final spot quality states of the model are ‘good’ and ‘bad’ and the nodes
consist of different spot measures. A training set is required to derive the structure of the
network, and estimate the conditional probabilities between the nodes. The method was
applied to small data sets (160 spots in each) where the spot quality had been verified by
visual inspection, and to a self-self hybridisation on a single array. Features such as bleed-
ing, spot roundness, intensity, size and background intensity were found to be important
variables in the model, and further suggestions of signal-to-noise ratio and green-to-red
balance were proposed as important measures to include if the model were to be extended
to large microarray experiments.

Raffelsberger et al. [2002] proposed overall gene flags based on signal-to-noise ratio,
saturation and replicate reproducibility (measured by coefficient of variation). High flags
(greater than 1) indicated unreliable data, which were removed from downstream analysis.

5.1.2 Methods based on log-ratio reproducibility

Methods which use log-ratio variability/reproducibility as the basis for their quality judg-
ments are also popular. Repeat hybridisations or multiple prints of the same gene on each
array are necessary to use these approaches.

Brown et al. [2001] computes a spot ratio variability (SRV), by dividing the standard
deviation of the ratio, estimated using the delta method, by the scaled expression ratio.
Spots with higher values of SRV were observed to contain artifacts in spot morphology.
A weighted mean ratio was calculated to summarise the expression of each gene across 9
replicate hybridisations, with 1/SRV used as the weights. Weighting was found to improve
the overall quality of the data set, based on an aggregate quality score (Q) calculated
across all spots. Using image analysis control flags to eliminate spots which appeared to
be shifted from their ideal location or to be suffering from other flaws further improved
the quality score.
Comparing the log-ratios of duplicate spots printed within an array, and between arrays when replicate hybridisations are available was done in Yang et al. [2002a] and Jenssen et al. [2002]. Clones whose log-ratios differed by more than 2 standard deviations from the mean of the distribution of differences were removed within slides as well as between slides in Yang et al. [2002a], and this was shown to improve the correlation between repeated log-ratios in both instances. Triplicate hybridisations were recommended to allow this kind of filtering between arrays.

Konig et al. [2004] compared a number of filtering methods including intensity-based filtering (on normalised, un-transformed signals), foreground mean-median discrepancy filtering (using the absolute difference between the log-ratio formed using foreground medians and the log-ratio formed using foreground means calculated on the same spot) and differences between the log-ratios of duplicate spots on the same array, or the same spot between replicate arrays. The log-ratio differencing approach was found to be the best method, giving the highest percentage of consistently responding genes (up/down/not-changing) determined using log-ratios from direct and indirect hybridisations.

Jenssen et al. [2002] propose a correlation based method suitable for use on arrays which have clones spotted multiple times, and the experiment has been replicated. The Pearson correlation of the log-ratios (which was averaged if spots were printed more than twice on each array) was used as an indicator of repeatability for a clone. Further averaging of the correlations for all clones in a data set was used to indicate the overall repeatability of the data. As a within array quality indicator, the mean absolute pair-wise deviation was used, with lower values indicating a more consistent array. The authors compare these methods on several data sets, and demonstrate variations in quality of the different data sets using these measures.

Nadon et al. [2001] removes outlier spots judged from a normal distribution of the log-ratios from replicate arrays using the 3rd and 4th moments of the distribution. Genes with more than a minimum number of outliers (2 in this paper) are removed from further analysis.

Tseng et al. [2001] uses multiple spotting of clones (more than twice) on the same array to calculate the coefficient of variation \( CV = \frac{SD}{\mu} \) of its log-ratios. Unreliable genes are chosen using an intensity based windowing procedure, which removes genes with \( CV \)’s in
the top 10% for each subset. The authors found this procedure to not work as effectively between slides because of the large amount of variability between arrays compared to within arrays.

Fan et al. [2004] takes a similar approach, but calculated the CV using spot intensity (A) rather than log-ratio (M) between 6 replicate hybridisations. Genes with high CV’s, taken to be greater than 1.5 standard deviations of the median CV, were regarded as unreliable and deleted from further analysis.

5.2 Spot quality weights

Most approaches outlined in the literature review (section 5.1) are filtering methods, which accept spots or reject them from further analysis using a filtering criterion and cut-off value. They assume observations go from ‘good’ to ‘bad’ in a sharp way that depends on the cut-off, which is often chosen in an ad-hoc manner and results in throwing away good data.

The aim of this study was to make use of the spot specific measures gathered during image analysis to derive relative spot weights \( w_{ij} \) for gene \( i \) on array \( j \), which are used along with the log-ratios \( M_{ij} \) in the linear model analysis. Relative weighting deals with quality in a graduated way, although very bad spots can still be excluded from the analysis by assigning them 0 weight. Where possible though, log-ratios are down-weighted in the analysis rather than thrown away.

In theory, a weight should be proportional to the precision of a measurement; for instance if log-ratio \( M_{ij} \) is measured with variance \( \sigma^2_{ij} \), then \( w_{ij} \propto 1/\sigma^2_{ij} \) is a suitable weight. Using appropriate weights in the linear model fits will give minimum variance estimates of gene expression, which should improve the detection of DE genes using \( t \) statistics \( t_{ik} = \hat{\beta}_{ik}/\text{SE}(\hat{\beta}_{ik}) \).

The existence of empirical relationships between reproducibility (precision) and various spot specific measures were investigated using the 111 replicate arrays which make up the QC data set (refer to section 2.3). The general trends learned from this analysis were used to derive weight functions for the spot measures examined. Generalised weight functions were used to assign spot weights in other data sets, including the Pax5, DC, and QDC
experiments, to assess whether this approach is widely applicable.

5.2.1 Spot specific measures

During image analysis, measurements such as area and foreground and background intensities and variabilities for Cy3 and Cy5 channels are obtained for each spot. Figure 5.1 gives an indication of which pixels are used to calculate these measures. As discussed in section 1.5, different software uses different segmentation algorithms for determining which pixels make up the spot foreground. The program Spot (Buckley [2000]) implements seeded region growing (SRG, see Adams and Bischof [1994]) which allows for irregular spot shapes, while GenePix uses adaptive circle segmentation, which assumes each spot is circular, but may vary in diameter. The background estimates are also obtained in different ways (refer to section 3.2.1). Despite the different algorithms, figure 3.1 (refer to section 3.2) shows that foreground estimates are similar between the packages while the background estimates can be systematically different.

![Spot Area (in pixels)](image)

![Intensity and its variability for foreground and background pixels in each channel](image)

Figure 5.1: Image analysis programs quantify the size and signal (including foreground and background pixel intensities and variabilities) for each spot.

Table 5.1 gives a brief summary of the equivalent spot specific measures available from these two image analysis packages. Due to slight differences in the way some of these measures are calculated, the scales can differ, and it is anticipated that slightly different weighting schemes will be needed to account for these differences. Recall also from chapter 3 that GenePix data will not be background corrected.
Table 5.1: Summary of the measures from Spot and GenePix image analysis programs. The spot areas measured using GenePix can take on a restricted set of values which depend on the diameter of the circular mask used in segmentation. Also of note is that the foreground and background IQRs are calculated from the logged foreground and background pixels in Spot, whereas in GenePix the SDs are measured on the original scale.

<table>
<thead>
<tr>
<th>Spot Characteristic</th>
<th>Spot 2.0</th>
<th>Values</th>
<th>GenePix Pro 4.0</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreground Intensity</td>
<td>( R_f, G_f )</td>
<td>([1, 2^{16}])</td>
<td>( R_f, G_f )</td>
<td>([1, 2^{16}])</td>
</tr>
<tr>
<td>Background Intensity</td>
<td>( R_b, G_b )</td>
<td>([1, 2^{16}])</td>
<td>( R_b, G_b )</td>
<td>([1, 2^{16}])</td>
</tr>
<tr>
<td>Signal</td>
<td>( R = R_f - R_b, G = G_f - G_b ) ([1 - 2^{16}, 2^{16} - 1])</td>
<td>( R = R_f, G = G_f )</td>
<td>([1, 2^{16}])</td>
<td></td>
</tr>
<tr>
<td>Spot size (in pixels)</td>
<td>Based on SRG</td>
<td>[0,300]</td>
<td>Based on circular mask</td>
<td>[0,300]</td>
</tr>
<tr>
<td>Perimeter</td>
<td>Measured in pixels</td>
<td>[9,100]</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Circularity</td>
<td>( \frac{4 \pi \text{Area}}{\text{Perimeter}^2} )</td>
<td>[0.2]</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Foreground variability</td>
<td>( R_f \text{IQR}, G_f \text{IQR} )</td>
<td>[0.16]</td>
<td>( R_f \text{SD}, G_f \text{SD} )</td>
<td>[0.2(16])</td>
</tr>
<tr>
<td>Background variability</td>
<td>( R_b \text{IQR}, G_b \text{IQR} )</td>
<td>[0.16]</td>
<td>( R_b \text{SD}, G_b \text{SD} )</td>
<td>[0.2(16])</td>
</tr>
<tr>
<td>Intensity ((A))</td>
<td>( A = 0.5(\log_2 R + \log_2 G) )</td>
<td>[0.16]</td>
<td>( A = 0.5(\log_2 R + \log_2 G) )</td>
<td>[0.16]</td>
</tr>
<tr>
<td>Signal-to-noise ratio</td>
<td>( \frac{\log_2(R_f - R_b)}{R_b \text{IQR}}, \frac{\log_2(G_f - G_b)}{G_b \text{IQR}} )</td>
<td>[0.50]</td>
<td>( \frac{R_f}{\text{R_b SD}}, \frac{G_f}{\text{G_b SD}} )</td>
<td>[0.1000]</td>
</tr>
<tr>
<td>Quality flag</td>
<td>-</td>
<td>-</td>
<td>User assigned &amp; automatic</td>
<td>-100,-75,-50,0,100</td>
</tr>
</tbody>
</table>

Signal based quantities such as variability of intensity and SNR are measured in each channel and, given our desire to work with log-ratios, it is necessary to combine these measures.

For foreground variability, Spot calculates the inter quartile range (IQR) of the logged foreground pixel intensities as a measure of variability in each channel, while GenePix takes the SD of the foreground pixels on the original scale. The IQR should be roughly proportional to the SD, so the combined foreground variability (on the log scale) was taken to be an average, \((R_f^2 \text{IQR} + G_f^2 \text{IQR})/2\) for Spot data. To obtain the GenePix equivalent, the square of the SD was divided by the square of the foreground signal (motivated by the relationship \( \text{var}(\log Y) \approx \text{var}(Y)/E(Y)^2 \)) before averaging across the two channels to give

\[
\frac{1}{2} \left( \frac{R_f^2 \text{SD}}{R_f^2} + \frac{G_f^2 \text{SD}}{G_f^2} \right).
\]

Pickett et al. [2001] calculates the SNR as \((\text{Foreground} - \text{Background})/\text{(SD of Background)}\)
in each channel. An average SNR was used to get a combined measure for Spot data of

\[ \frac{1}{2} \left( \frac{\log_2(R_f - R_b)}{R_b \text{IQR}} + \frac{\log_2(G_f - G_b)}{G_b \text{IQR}} \right) \]

which is on the log scale since the pixels are logged before obtaining the background IQR. The GenePix equivalent, but on the original scale, is

\[ \frac{1}{2} \left( \frac{R_f}{R_b \text{SD}} + \frac{G_f}{G_b \text{SD}} \right) \]

since the data are not background corrected and the SDs are calculated using the raw pixel intensities.

Spot perimeter and circularity are quality measures specific to the Spot software. Circularity aims to assess the regularity of a spot’s shape. A theoretical circularity measure of 1 indicates a perfectly round spot, and values larger or smaller indicate the SRG mask has an irregular morphology. Since the spots are printed to be circular, this measure should give a good indication of mis-printed spots which may be of lesser quality than circular spots.

GenePix includes quality flags which are a mix of user defined and automatically assigned spot classes for each array. Flag values of ‘bad’ (−100) or ‘good’ (100) are assigned to spots after visually inspecting them. A flag of −75 indicates a spot is either not defined in the GAL file, or is labelled as ‘empty’ and −50 is an automatically generated flag which indicates the spot was ‘not found.’ According to the GenePix User’s Guide (GenePix Pro [2001]), the ‘not found’ flag is assigned to a spot when: it has fewer than 6 pixels; its diameter is too large, or it overlaps with a neighbouring spot. In practice, this flag value is associated with low intensity spots. All other spots receive a flag of 0. For the data sets analysed in this chapter, flags of −50 and 0 were the only flagged categories available. No manual calling of ‘good’ and ‘bad’ spots was performed, and no empty/not defined spots were present.

The extent to which each of the measures introduced in this section (spot area, perimeter, circularity, intensity, average SNR, average foreground variability and GenePix flags) are able to systematically explain the variability of the log-ratios across the 111 arrays from the QC experiment is of immediate interest here, and will be examined closely in the
5.2.2 Measuring reproducibility

Reproducibility or precision can be measured explicitly when repeat observations of gene expression are available. In the case of the QC data set, there are 111 repeat observations (log-ratios) for each gene from Spot and GenePix respectively.

An intuitive measure of reproducibility of expression for a given gene \((i)\) is measured by the drop-one-out residual, \(e_{ij}\). For gene \(i\), measured on \(J\) replicate arrays \((j = 1, \ldots, J)\), \(e_{ij}\) is calculated as the difference between the log-ratio on the \(j\)th array \((M_{ij})\) and the average expression value on the remaining arrays \((\bar{M}_{i(j)})\), i.e., \(e_{ij} = M_{ij} - \bar{M}_{i(j)}\). Hence a spot with a large absolute value of \(e_{ij}\) is not as reproducible as one with a small value (close to zero). Refer to figure 5.2 for an example.

The \(e_{ij}\) were calculated for each spot across \(J = 111\) arrays from the QC data set, and squared to give a variance like measure. A smaller \(e_{ij}^2\) indicates a more reproducible expression estimate for a spot compared to a larger value.

5.2.3 Quality trends

The ability of spot area, intensity, average SNR and foreground variability to predict spot reproducibility are shown in figures 5.3, 5.4, 5.5 and 5.6 respectively. The genes from the QC data set (10,368 genes x 111 arrays = 1,150,848 observations) were used for these figures.

The squared drop-one-out residuals \((e_{ij}^2)\) are generally very small, so \(\log_2(e_{ij}^2)\) was plotted to allow trends in reproducibility to be visualised. The spot specific measures were also binned, and box plot displays were used at regular intervals to show the distribution of the log variability. Although the individual variance like measures are very variable, the medians of these box plots show the general trends. In all figures in this section, the boxes are drawn with widths proportional to the square-root of the number of observations in that bin, and the red dashed lines in each plot on the y-axis indicate a 2-fold change in variation on the original scale.

In the case of area (figure 5.3), spots which are either smaller, or larger than the ideal
Figure 5.2: Example of reproducibility measure \((e_{ij})\) for a gene \(i\) from a set of \(J = 5\) replicate microarrays. For unknown reason, hybridisation of labelled mRNA hasn’t occurred for this gene on the 5th array. Assuming this is reflected in its log-ratio, which is quite different to the log-ratios measured on the first 4 arrays, this spot will have a large drop-one-out residual \(|e_{i5}| > 0\). The other 4, more reproducible spots will have smaller residual values \((e_{ij} \approx 0)\).

Figure 5.2: Example of reproducibility measure \((e_{ij})\) for a gene \(i\) from a set of \(J = 5\) replicate microarrays. For unknown reason, hybridisation of labelled mRNA hasn’t occurred for this gene on the 5th array. Assuming this is reflected in its log-ratio, which is quite different to the log-ratios measured on the first 4 arrays, this spot will have a large drop-one-out residual \(|e_{i5}| > 0\). The other 4, more reproducible spots will have smaller residual values \((e_{ij} \approx 0)\).

Perimeter and circularity give similar trends (refer to figure B.1 in Appendix B); spots which have smaller or larger measures than the ideal (which appears to be about 35 pixels for perimeter, and 0.8 for circularity, which interestingly is smaller than the theoretical value of 1 which should indicate a perfectly round spot) tend to be gradually less reproducible. The trend for circularity is less pronounced than those observed for area or perimeter, and perimeter isn’t measured in all image analysis packages, so area was chosen as the most suitable morphology measure for the purpose of spot quality weighting.

For average spot intensity, reproducibility generally improves with increasing intensity for both image analysis packages (see figure 5.4) at a similar rate. Variability is lowest for
spots with an intensity of between 12.5 and 13.5 in both packages, and higher intensity spots \( A > 14 \) show a slight decrease in reproducibility, caused by saturated pixels within spots, which results in a loss of information and less accurate log-ratios for affected spots.

The variability versus average SNR is given in figure 5.5, and shows a trend similar to that observed for intensity; spots with higher SNR are generally more reproducible than spots with low SNR. Further investigation revealed that average SNR is correlated with spot intensity (refer to figure B.2 in Appendix B), which is not surprising given that both measures use the intensities from each channel in their calculation. Given this redundancy, average SNR wasn’t pursued further for spot quality weighting purposes.

A sensible relationship for the average foreground pixel variability would be for more homogeneous spots, with low pixel variability to be more reproducible than spots with higher foreground pixel variability. This increasing trend is evident for Spot and GenePix data (see figure 5.6), with reproducibility decreasing gradually as foreground variability increases.

Figure 5.3: Reproducibility trends for spot area. The ideal spot area for these arrays, which is determined during printing, is between 70 and 80 pixels. Spots smaller and larger than this ideal size tend to have less precise log-ratios.
Figure 5.4: Reproducibility improves with increasing average spot intensity for both image analysis packages up to $A \approx 13$. A slight increase in variability is noticed after this, which is caused by pixel saturation.

Figure 5.5: Reproducibility trends for average SNR are similar for Spot and GenePix, with variability decreasing as the SNR increases.
Figure 5.6: Foreground variability trends are similar for Spot and GenePix. Spots with lower foreground variability are generally more reproducible than those with higher foreground variability, as measured by the average squared IQR for each channel in Spot, and the average standardised SD$^2$ for GenePix.

Combining foreground variability and spot area into a ratio, results in a similar relationship to that observed for foreground variability alone. Figure 5.7 shows the logged, and squared drop-one-out residual measure against (average foreground variability)/area. The motivation for forming a ratio is based on the idea that variability of the foreground pixels will be inversely proportional to the number of pixels used to estimate the foreground measures (i.e. smaller spots would tend to have higher foreground variability than larger ones, since they are estimated using a smaller sample). Since GenePix and Spot allow the number of foreground pixels to vary from spot to spot, dividing the variability by the number of pixels used to estimate it standardises this measure (analogous to converting a sample SD into a SE). The assumption that the pixels are independent is implicit in such a calculation.

As figure 5.8 shows, spots which receive an automatic GenePix flag of $-50$ (‘not found’) are slightly more variable on average than spots which receive a flag of 0. The general use of GenePix flags is to remove negatively flagged spots from further analysis (see Verdnik
Figure 5.7: Reproducibility trends for average foreground variability divided by area for Spot and GenePix. Reproducibility decreases as the foreground variability standardised by area increases, which is similar to the result observed using foreground variability alone (see figure 5.6).

[2004]), which would appear to be a reasonable practice based on this empirical evidence. However, it can result in throwing away a large proportion of the data, in this case 376,985 out of 1,150,848, or around 1/3 of all spots would be discarded.

5.2.4 Spot quality weighting schemes

For gene $i$ on array $j$, the log-ratio $M_{ij}$, will be assigned a weight $w_{ij}$, which could be used in the normalisation and linear models. Appropriately chosen weights should be proportional to the precision of the measurement. That is, if $M_{ij}$ is measured with variance $\sigma^2_{ij}$, then $w_{ij} \propto 1/\sigma^2_{ij}$ would be a suitable weight. In general $\sigma^2_{ij}$ can’t be measured directly, however it can be approximated by the squared drop-one-out residual $e^2_{ij}$ in the QC data set.

Most of the empirical trends in median reproducibility ($e^2_{ij}$) observed in section 5.2.3 can be generalised to give spot weight functions, which assign spot weights based on a spot’s area, intensity, average foreground variability or a combination of these.
Spot weights based on area have both intuitive and empirical appeal. Giving full weight to spots which have been quantified as being of the ideal size, which is a parameter largely controlled for during array printing, and relatively less weight to larger and smaller spots is a sensible choice with an empirical basis (refer to figure 5.3).

The function `wtarea` in `limma` has been implemented to generate relative area weights in the manner described above, using linear interpolation. The function takes an argument of the ideal size (between 65 and 85 pixels for the QC data set) and assigns full weight of 1 to these spots. The weight function decreases linearly for smaller and larger spots to a minimum of 0 weight for spots of size 0 and 150 pixels (150 = 65 + 85, the sum of the ideal pixel range). Figure 5.9 shows this weight function for the QC data set. This weight function can be generalised to any data set where the ideal spot sizes are known. Area is also measured in most image analysis programs, which makes it preferable to some of the more specialised morphology measures such as perimeter and circularity, which aren’t as widely available.

Intensity based filtering schemes have been proposed by many authors (see section 5.1),

Figure 5.8: Reproducibility stratified by GenePix flag. Spots receiving a flag of −50 (‘not found’) are on average slightly less reproducible than the remaining spots which are assigned a 0 flag. Around 1/3 of spots received a ‘not found’ flag in this data set.
so an intensity based weighting method was also considered. An empirically derived intensity weighting function was generated from the trend observed in figure 5.4. Cubic splines were used to interpolate the median of the inverse reproducibility measures ($1/e_{ij}^2$) from each intensity bin ($0, 1, \ldots, 16$) from the GenePix data. The resulting function is plotted in figure 5.9 (middle panel). The intensity weight function gives lowest weight ($\approx 0.4$) to spots with intensities below 6, increasing to maximum weight (1) for spots of intensity 13.5, before declining again for very bright spots which suffer from signal saturation. The same spline weight function was applied to Spot and GenePix data from various experiments to derive prior intensity based weights for each spot. No data set specific calibration was performed.

For the foreground variability measures, the reciprocal ($1/(\text{average foreground variability})$) was used to assign a weight to each spot (see figure 5.9, right panel). The motivation for an inverse weight function was pre-empted by the data (figure 5.6), which showed a decline in reproducibility as the average foreground pixel variability increased. Intuitively, if the
average foreground variability really is measuring the variance of \( M_{ij} \) (or something proportional to it), its inverse would be a natural weight. Using similar logic, the inverse of (average foreground variability)/area was used to derive prior spot quality weights which combine foreground variability and area. Spots which were assigned 0 weight by the area weight scheme (spots of size 0 and those greater than the sum of the ideal spot range) were also assigned zero weight for this method.

Spots can be of lesser quality for a variety of reasons, and the use of univariate spot quality weights such as area, intensity or foreground variability on their own, fails to address this. To get an idea of the relative importance of each variable, regression trees were fitted to the data using the `tree` function (from the `tree` package) in R. Regression trees can be used as a variable selection procedure and are able to highlight interactions between variables (Clark and Pregibon [1992]). A regression tree is a bifurcating tree, which recursively partitions the response variable, \( y \) (which has individual elements \( e_{ij}^2 \) in this case) using a set of predictor variables, which were chosen to be area, intensity, and foreground variability (on their original scale) for the Spot data, and the corresponding measures for GenePix, along with the flags. The partitioning algorithm uses successive splits in a predictor variable which maximise the change in deviance \( \Delta D = \sum_L D(\hat{\mu}_L, y_k) - \sum_R D(\hat{\mu}_R, y_k) \), between the left (L) and right (R) partitions for that variable, where \( D(\mu, y_k) = (y_k - \mu)^2 \).

Figure 5.10 shows the regression trees fitted to the data from each image analysis package. For Spot data, area, followed by average foreground variability and intensity were the most important variables included in the model. Table 5.2 shows the split rules and deviances for each branch of the tree.

The tree based on GenePix data uses average foreground variability as the major variable to partition the variation, followed by area and intensity in subsequent branches. The GenePix flags don’t feature in the tree, indicating that this variable is not as useful for characterising variability as some of the other variables. Table 5.3 shows the partitions, number of observations for each split and deviances for this tree.

Given the importance of area and average foreground variability, and to a lesser extent spot intensity in partitioning the variability in both trees, a combined spot weight along the lines of that proposed in Wang et al. [2001] was also considered. Since area,
Figure 5.10: Regression trees of spot specific measures for Spot (a) and GenePix (b). The response variables were the vector of $y_{ij} = e_{ij}^2$ in each tree and the predictors were area, intensity and average foreground variability for the Spot tree and the equivalent measures for the GenePix tree, along with the flags.
Table 5.2: Branches of the regression tree for Spot data. * denotes a terminal node.

<table>
<thead>
<tr>
<th>Node</th>
<th>Split</th>
<th>n</th>
<th>Deviance</th>
<th>( \hat{\mu} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Root</td>
<td>1,150,818</td>
<td>404,700</td>
<td>0.2123</td>
</tr>
<tr>
<td>2</td>
<td>Area &lt; 21.5</td>
<td>4,377</td>
<td>25,830</td>
<td>0.8952</td>
</tr>
<tr>
<td>4</td>
<td>AveFgIQR &lt; 3.505</td>
<td>3,981</td>
<td>11,080</td>
<td>0.7327</td>
</tr>
<tr>
<td>8</td>
<td>Intensity &lt; 9.685</td>
<td>2,426</td>
<td>2,983</td>
<td>0.4917</td>
</tr>
<tr>
<td>9</td>
<td>Intensity &gt; 9.685</td>
<td>1,555</td>
<td>7,736</td>
<td>1.1090</td>
</tr>
<tr>
<td>5</td>
<td>AveFgIQR &gt; 3.505</td>
<td>396</td>
<td>13,590</td>
<td>2.5280</td>
</tr>
<tr>
<td>10</td>
<td>Intensity &lt; 11.935</td>
<td>146</td>
<td>10,130</td>
<td>3.7750</td>
</tr>
<tr>
<td>11</td>
<td>Intensity &gt; 11.935</td>
<td>250</td>
<td>3,099</td>
<td>1.8000</td>
</tr>
<tr>
<td>3</td>
<td>Area &gt; 21.5</td>
<td>1,146,441</td>
<td>376,800</td>
<td>0.2097</td>
</tr>
<tr>
<td>6</td>
<td>AveFgIQR &lt; 0.875</td>
<td>670,440</td>
<td>193,800</td>
<td>0.1880</td>
</tr>
<tr>
<td>7</td>
<td>AveFgIQR &gt; 0.875</td>
<td>476,001</td>
<td>182,200</td>
<td>0.2403</td>
</tr>
<tr>
<td>14</td>
<td>Intensity &lt; 10.625</td>
<td>464,361</td>
<td>138,200</td>
<td>0.2342</td>
</tr>
<tr>
<td>28</td>
<td>Area &lt; 29.5</td>
<td>60,571</td>
<td>27,890</td>
<td>0.3047</td>
</tr>
<tr>
<td>29</td>
<td>Area &gt; 29.5</td>
<td>403,790</td>
<td>110,000</td>
<td>0.2236</td>
</tr>
<tr>
<td>15</td>
<td>Intensity &gt; 10.625</td>
<td>11,640</td>
<td>43,280</td>
<td>0.4855</td>
</tr>
<tr>
<td>30</td>
<td>AveFgIQR &lt; 4.915</td>
<td>11,210</td>
<td>33,980</td>
<td>0.4461</td>
</tr>
<tr>
<td>31</td>
<td>AveFgIQR &gt; 4.915</td>
<td>430</td>
<td>8,837</td>
<td>1.5140</td>
</tr>
<tr>
<td>62</td>
<td>Intensity &lt; 11.395</td>
<td>40</td>
<td>5,538</td>
<td>4.7680</td>
</tr>
<tr>
<td>124</td>
<td>AveFgIQR &lt; 9.035</td>
<td>33</td>
<td>1,181</td>
<td>1.9240</td>
</tr>
<tr>
<td>248</td>
<td>AveFgIQR &lt; 5.395</td>
<td>5</td>
<td>774</td>
<td>9.9730</td>
</tr>
<tr>
<td>249</td>
<td>AveFgIQR &gt; 5.395</td>
<td>28</td>
<td>25</td>
<td>0.4861</td>
</tr>
<tr>
<td>125</td>
<td>AveFgIQR &lt; 9.035</td>
<td>7</td>
<td>2,833</td>
<td>18.1700</td>
</tr>
<tr>
<td>63</td>
<td>Intensity &gt; 11.395</td>
<td>390</td>
<td>2,832</td>
<td>1.1800</td>
</tr>
</tbody>
</table>

Intensity and foreground variability weights were on different scales, they were combined by multiplication \((w_{\text{combined}}_{ij} = w_{\text{area}}_{ij} \times w_{\text{intensity}}_{ij} \times w_{\text{ave fg var}}_{ij})\) to give a combined weight. This penalises (or rewards) spots which receive low (or high) weights for multiple measures even more than is possible using an individual weighting method on its own.

Finally, a simple filtering method based on the GenePix flags which removes spots which receive a negative flag (as recommended in Verdnik [2004]) was also used. This is equivalent to assigning 0 weight to spots with negative flags, and full weight (1) to those receiving a positive flag. The function \texttt{wtflags} in \texttt{limma} implements this weighting scheme. The log-ratios from spots which receive 0 weight are ignored in the analysis.
Table 5.3: Branches of the regression tree for GenePix data. * denotes a terminal node.

<table>
<thead>
<tr>
<th>Node</th>
<th>Split</th>
<th>n</th>
<th>Deviance</th>
<th>$\mu$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Root</td>
<td>1,150,668</td>
<td>230,400</td>
<td>0.1856</td>
</tr>
<tr>
<td>2</td>
<td>AveFgVar &lt; 4.625</td>
<td>1,146,748</td>
<td>218,200</td>
<td>0.1834</td>
</tr>
<tr>
<td>4</td>
<td>Area &lt; 30</td>
<td>74,747</td>
<td>25,120</td>
<td>0.3038</td>
</tr>
<tr>
<td>8</td>
<td>AveFgVar &lt; 2.135</td>
<td>74,257</td>
<td>21,480</td>
<td>0.2973</td>
</tr>
<tr>
<td>16*</td>
<td>Intensity &lt; 13.005</td>
<td>74,176</td>
<td>19,330</td>
<td>0.2953</td>
</tr>
<tr>
<td>17*</td>
<td>Intensity &gt; 13.005</td>
<td>81</td>
<td>1,876</td>
<td>2.1360</td>
</tr>
<tr>
<td>9*</td>
<td>AveFgVar &gt; 2.135</td>
<td>490</td>
<td>3,166</td>
<td>1.2870</td>
</tr>
<tr>
<td>5</td>
<td>Area &gt; 30</td>
<td>1,072,001</td>
<td>191,900</td>
<td>0.1750</td>
</tr>
<tr>
<td>10*</td>
<td>AveFgVar &lt; 0.835</td>
<td>719,361</td>
<td>113,200</td>
<td>0.1613</td>
</tr>
<tr>
<td>11*</td>
<td>AveFgVar &gt; 0.835</td>
<td>352,640</td>
<td>78,270</td>
<td>0.2030</td>
</tr>
<tr>
<td>3</td>
<td>AveFgVar &gt; 4.625</td>
<td>3,920</td>
<td>10,530</td>
<td>0.8319</td>
</tr>
<tr>
<td>6</td>
<td>Area &lt; 22</td>
<td>227</td>
<td>1,564</td>
<td>1.7070</td>
</tr>
<tr>
<td>12*</td>
<td>Intensity &lt; 9.945</td>
<td>11</td>
<td>124</td>
<td>5.7920</td>
</tr>
<tr>
<td>13*</td>
<td>Intensity &gt; 9.945</td>
<td>216</td>
<td>1,247</td>
<td>1.4990</td>
</tr>
<tr>
<td>7*</td>
<td>Area &gt; 22</td>
<td>3,693</td>
<td>8,785</td>
<td>0.7781</td>
</tr>
</tbody>
</table>

5.2.5 Spot quality weights in the normalisation and gene-wise linear models

The spot quality weights described in section 5.2.4 can be used in the intensity-dependent loess normalisation (described in section 1.6.3). In the normalisation of the log-ratios from the $j$th array, each spot weight $w_{ij}$, is multiplied by a neighbourhood weight ($n_{ij}(A)$), i.e., $w_{ij}n_{ij}(A)$ in the initial iteration, and for the robust iterations, multiplied by a robustness weight ($r_{ij}$) and neighbourhood weight ($w_{ij}r_{ij}(A)n_{ij}(A)$) in the loess procedure (refer to Cleveland et al. [1992] for further detail). This has the effect of reducing the influence of log-ratios with small case weights compared to observations with relatively larger weights in the loess fit.

Although spot quality weights can be easily accommodated in the normalisation, they were not used for this purpose in the present study. Part of the reason was for convenience; there were 6 spot weighting methods to compare for Spot data and 7 for GenePix data which would mean 13 separate normalisations for each data set, which is both computationally and memory intensive. A more important consideration though is how to evaluate the effectiveness of spot weighting in the normalisation. The best way would be to look at MSP (Microarray Sample Pool) controls (refer to section 1.4.1) to see whether their bias is
reduced after using spot quality weights. Unfortunately none of the data sets used in this thesis have such controls. Generally speaking it would be anticipated that spot weights will have minimal impact in the normalisation given the existing robustness of the loess normalisation procedure. One obvious exception would be if the weights were themselves intensity dependent. If spots at a particular intensity were systematically down-weighted, then this may interfere with the intensity dependent normalisation.

In this chapter, our attention was directed at the usefulness of weights in the gene-wise linear models fitted to the log-ratios. Typically, when \texttt{lmFit} is called in \texttt{limma}, least squares regression is performed using the base function \texttt{lm.fit}. However, when weights are supplied, the base function \texttt{lm.wfit} is called to perform weighted least squares regression, which solves the normal equations

\[ X_{\text{gene}}^T W^{-1}_i X_{\text{gene}} \beta_i = X_{\text{gene}}^T W^{-1}_i y_i \]

where \( y_i^T = (M_{i1}, \ldots, M_{iJ}) \), \( X_{\text{gene}} \) is the design matrix, the weight matrix \( W_i \) has diagonal elements \( w_{ij} \), and \( \beta_i \) is a vector of coefficients measuring gene expression.

For a linear model with \( K \) coefficients \( (\beta_i^T = (\beta_{i1}, \ldots, \beta_{iK})) \), the estimate for the \( k \)th coefficient, \( \hat{\beta}_{ik} \) and its SE \( \text{SE}(\hat{\beta}_{ik}) \) will be used to form ordinary \( t \) statistics \( (t_{ik} = \hat{\beta}_{ik} / \text{SE}(\hat{\beta}_{ik})) \). For a given \( k \), genes can be ranked in order of evidence for differential expression using \( |t_{ik}| \) in descending order. The \( t \) statistics calculated using spot weights derived under the different weighting schemes described in section 5.2.4 will be compared with the standard \( t \) statistics obtained without using weights to assess any loss of signal or reduction in noise. The QC data set (which has \( K = 1 \) coefficient for each gene), as well as data from the Pax5 (\( K = 3 \)), DC (\( K = 3 \)) and QDC (\( K = 1 \)) experiments were used to evaluate each of the weighting methods. For the LMS and LUS control genes from each of these data sets, there is only one coefficient to estimate (\( K = 1 \)) for each spot (or spot pair if they are printed in duplicate) between arrays, since they are spiked-in to the green labelled and red labelled samples in the same way, on all arrays in these experiments.

Looking at the \( t \) statistics of the spike-in control genes (LMS or LUS) is informative since their differential expression status is known. The \( t \) statistics from the DR or CC give an indication of the null distribution, since these artificial genes are added in equal
amounts to both samples. On the other hand, the \( t \) statistics of the ratio controls can be used as a yardstick for assessing real differential expression, since they are up/down regulated at specific levels (refer to tables 2.1 and 2.2 in section 2.7).

5.3 Results: Spot quality weights in the linear models

5.3.1 Spot weights applied to the QC data set

Using Spot image analysis output

To look for improvements in our ability to detect differential expression, the ordinary \( t \) statistics of the LMS controls were plotted before and after applying spot weights. For the QC data set, the analysis of the LMS controls was restricted to 100 arrays for which the spike-ins had been added to the mRNA mixes prior to labelling.

In figure 5.11, the distributions of the \( t \) statistics by weighting scheme for different classes of controls are shown. The absolute medians of the \( t \) statistics for the U03, D03, U10 and D10 increase for all weighting methods. The improvements range from modest increases of 4%-14% achieved using area or intensity weights, to larger changes (15%-60%) for methods involving average foreground variability, combined foreground variability and area or the three spot weights (area, intensity and foreground variability) combined.

Figure 5.12 shows the \( Z \)-score equivalent of the \( t \) statistics plotted in figure 5.11 for each class of control. The \( Z \)-score represents the standard normal deviate with the same quantile as the \( t \) statistic in its null distribution. This puts \( t \) statistics on different degrees of freedom on the same scale. Recall that the degrees of freedom are allowed to differ depending on the weighting scheme due to the introduction of 0 weights in some methods. This figure recapitulates the improvements noted for the raw \( t \) statistics; all weighting methods give more extreme \( Z \)-scores for known DE ratio controls compared to not using weights. The ordering of improvements from largest to smallest starts with combined weighting as the best, followed by the other two weighting schemes involving foreground variability, intensity weights and finally area weighting. It is also clear from this figure that the null distributions of the DR are fairly constant across the methods.

Figure 5.13 presents the ordinary \( t \) statistics for the 10,368 genes by weighting method.
Figure 5.11: Plot of $t$ statistics for LMS controls for each spot weighting scheme, using 100 arrays from the QC data set, and Spot image analysis output.

Figure 5.12: Plot of $Z$-scores for LMS controls for various spot weighting schemes from the QC data set using Spot data.
The five genes with the largest negative and positive ordinary \textit{t} statistics computed without case weights are plotted (each with a different symbol) for each weighting scheme. As for the ratio controls, the most extreme genes become more extreme after applying weights, especially for the last three methods which involve foreground variability. In general, the distributions are fairly consistent between the methods, providing no evidence for loss of signal.

Figure 5.13: Plot of \textit{t} statistics for 10,368 genes for various spot weighting schemes from the QC data set using Spot data.

The main reason for these improvements is a reduction in the SEs. Using appropriate spot weights in the linear models should give more precise estimates of the coefficients than those obtained without weighting. Ratios of the weighted SEs to the regular SEs (based on equal weights) for the estimated coefficients are shown in figure 5.14. For the controls (right panel), the median decrease in SE ranges from 4% when area weights are used to 19% for combined weights. For the genes, the median reduction is much less, ranging from no change for intensity weights to a maximum of 6% when combined weights are used. Also of note in figure 5.14, is that the weighting doesn’t improve matters for every probe, with the SEs increasing to varying degrees for between 22% to 44% of genes and 2% to 10% of controls.
Figure 5.14: Ratio of SEs of coefficients with:without spot weights for the genes (left) and LMS controls (right) from the QC data set (Spot data). The different spot weighting schemes are colour-coded.

Since the log-ratios of the D03, D10, U03, U10 and DR should be at known levels, the effect of the weights on bias can be visually assessed. Figures 5.15 to 5.19 show the log-ratios for each class of LMS control versus the relative spot weights for each weighting scheme. Spots with log-ratios which differ from the desired level should receive lower weight than those at the ideal level. Although each method has weights on a slightly different scale, there is a tendency for aberrant log-ratios to be assigned relatively lower weights, particularly for the signal based methods. Area weights are least consistent at assigning more variable log-ratios less weight, with a handful of ratio controls receiving maximum weight when the $M$ values are higher/lower than the desired level. These figures give a raw view of where the reduction in variability is coming from.

Using GenePix image analysis output

The analysis was repeated using the GenePix image analysis output. Figure 5.20 shows the ordinary $t$ statistics for the LMS controls calculated from 100 arrays for each weighting method. As was the case for the Spot data, the $t$ statistics for the DE controls are
Figure 5.15: $M$ values of D03 versus relative spot weights for different weighting schemes for the QC data set (Spot data). The dashed lines represent the predicted $M$ values for the D03 ($-\log_2 3$).

Figure 5.16: $M$ values of D10 versus relative spot weights for different weighting schemes for the QC data set (Spot data). The dashed lines represent the predicted $M$ values for the D10 ($-\log_2 10$).
Figure 5.17: $M$ values of U03 versus relative spot weights for different weighting schemes for the QC data set (Spot data). The dashed lines represent the predicted $M$ values for the U03 ($\log_2 3$).

Figure 5.18: $M$ values of U10 versus relative spot weights for different weighting schemes for the QC data set (Spot data). The dashed lines represent the predicted $M$ values for the U10 ($\log_2 10$).
Figure 5.19: $M$ values of DR versus relative spot weights for different weighting schemes for the QC data set (Spot data). The dashed lines represent the predicted $M$ values for the DR (0).

Figure 5.20: Plot of $t$ statistics for LMS controls for each spot weighting scheme, using 100 arrays from the QC data set, and GenePix image analysis output.
more extreme after applying weights, with the biggest changes evident for combined (area, intensity and foreground variability) weights. The median increases in the $t$ statistics for area weighting range from 5%-23%, intensity weights add 18%-120%, for average foreground variability increases were in the order of 3%-38%, foreground variability divided by area gave increases of 5%-64%, combined weights 17%-127% and flag filtering increased the median $t$ statistics by 17%-116%.

This ordering holds when the equivalent $Z$-scores shown in figure 5.21 are examined. The null distributions of the DR are fairly constant across the 7 alternatives, which is desirable given that these controls are not DE.

Figure 5.22 plots the ratios of the SEs with:without weights for the genes and LMS controls. The reduction in SEs offered by weighting for the genes ranges from 2% for area weights through to 6% for (average foreground variability)/area weights, with intensity weights having no effect, and flag filtering increasing the SEs by 7% on average. For the controls, the decreases in SEs range from 2% when filtering is applied to 18% when
5.3.2 Spot weights applied to other data sets

While combined weights incorporating area, intensity and foreground pixel variability appeared to be the most useful on the QC data set for Spot and GenePix data, the weighting functions applied were learnt from this data set, and in the case of the intensity weights which were empirically derived, tuned to it. Along similar lines to machine learning techniques, it is important to look at examples outside of the ‘training’ data set to assess the general utility of these weighting methods. To this end, data from several smaller, more typical experiments were examined, to see if similar gains in the ability to detect DE genes could be made.

Pax5 experiment

The Pax5 experiment is a medium sized data set made up of 9 arrays comparing 4 sources of mRNA (refer to section 2.2 for details on this experiment). Data from Spot and GenePix
were available on all arrays, and normalisation was performed using intensity-based print-tip loess for the genes and a global intensity-based loess fit (with span=0.7) for the LUS controls.

Spot quality weights were derived using the weight functions described in section 5.2.4, with an ideal spot size of between 80 and 100 pixels specified for the area weight function. The coefficients, \((\beta_{W1-M11}, \beta_{W1-M12}, \beta_{W1-M13})\) were estimated for each gene, and \(\beta_{LUS}\) for each control using linear models, both with and without spot quality weights. The availability of duplicate spots, printed side-by-side for each gene and control, allowed estimation and specification of the inter-duplicate correlation in the linear models (Smyth et al. [2005]). Spot quality weights were used in the `duplicateCorrelation` function in `limma` to estimate these correlations, which varied slightly for each weighting scheme. Values of around 0.90 for Spot and 0.87 for GenePix were estimated for the controls, and 0.75 for Spot and 0.74 for GenePix were the correlations for the genes.

Figures 5.23 and 5.24 show the \(Z\)-scores calculated using the ordinary \(t\) statistics and residual degrees of freedom obtained using Spot and GenePix data respectively. For the DE controls, the weighting methods which use average foreground variability tend to improve the \(Z\)-scores for both sources of data, although not to the same extent as observed for the QC experiment. Combining the weights gives the best results, producing the largest increases in \(|Z|\). The weights derived from area and intensity for both Spot and GenePix along with flag filtering the GenePix data had little effect in improving the detection of the DE controls in this data set. The \(Z\) scores for the calibration controls were slightly more variable for all weighting methods for Spot data, and for all methods except area and intensity weighting and flag filtering for the GenePix data. This is undesirable since these spots are not DE.

An examination of the ratios of SEs with:without weights (see figures B.3 and B.4 in Appendix B) reveals that the use of area or intensity weights for Spot or GenePix offer no net reduction in the SEs of the genes or the controls. For the spot quality weighting approaches involving foreground variability, the trend is for a very small median reduction in the SEs in the order of 8% to 14% for the controls and 0% to 4% for the genes. Applying flag filtering to the data decreases the precision of the estimates for the genes by 2% and has no net effect on the SEs for the controls.
DC experiment

The DC experiment compares 3 samples of mRNA in a dye-swap fashion on 6 arrays (refer to section 2.5 for details on this experiment). Using GenePix data, area weights based on an ideal spot size of 120 pixels were assigned, along with the other spot weights described in section 5.2.4, and used in the linear model analysis of the normalised log-ratios. Gene expression coefficients, \((\beta_{\text{CD4}^+\text{CD4}^-}, \beta_{\text{CD4}^-\text{CD4}^+})\), and LUS control coefficients, \(\beta_{\text{LUS}}\) were estimated with and without weights; the specification of an inter-duplicate correlation parameter made use of the extra information available from the side-by-side duplicate prints of each clone/control. Average correlations of around 0.87 for the clones and 0.93 for the LUS controls were estimated.

Figure 5.25 shows the equivalent Z-scores for each weighting scheme. Most of the DE spike-in controls have a tendency to become more extreme when weights which use average foreground variability are used in the linear model fits. GenePix flag weighting recovered Z-scores very similar to those obtained without weights, indicating that very
few of these controls received a ‘not found’ flag. Spot weights based on intensity and area also made little difference, and the variation of the $Z$ scores of the non-DE calibration controls increased slightly for most weighting methods.

The ratios of SEs with:without weights (see figure B.5 in Appendix B) indicates that flag filtering inflates the SEs of the coefficients for the majority of the clones and controls by 3-4%, and area and intensity weights have no net effect. Weights which use foreground variability are most consistent at reducing the SEs in the clones and controls, by 4% to 7% on average, with combined weights faring the best.

**QDC experiment**

The final data set examined was a much smaller dye-swap experiment involving only 2 arrays comparing mRNA from quiescent CD4$^+$8$^-$ and CD4$^-$8$^+$ splenic DC directly (refer to section 2.6 for further details on this experiment). This experiment was examined to see whether spot weights make a noticeable difference when there is minimal data on
each gene. In this experiment, there are duplicate spots on the arrays, which gives 4 observations on each probe.

Normalisation was carried out separately for the controls (using a global loess fit, with span=0.66) and clones (using intensity based print-tip loess). The normalised log-ratios were used in the linear model fits to estimate the coefficients, $\beta_{(CD_{4+8^-})-(CD_{4-8^+})}$ for each gene, and $\beta_{LUS}$ for each control, with and without spot quality weights. The weights were derived in the manner described in section 5.2.4, using an ideal spot size range of 150 to 160 pixels for the area weights. The duplicate correlation methodology of Smyth et al. [2005] was used to estimate the average inter-array duplicate correlation, which was quite high, at around 0.81 for the clones and 0.86 for the controls.

Using GenePix data, the results in terms of the equivalent Z-scores (see figure 5.26) show that filtering the data using flags is not useful, reducing the median $|Z|$ scores of the DE controls in most instances (D03Low, U03Low, D03High, U03High and U10High). The other weighting methods have little effect, giving similar Z scores to those obtained

Figure 5.25: Plot of Z-scores for LUS controls calculated using GenePix data from the DC experiment.
using equal weights. The null distribution of the calibration controls is also fairly constant across all methods.

The ratio of SEs of the coefficients estimated with:without weights (see figure B.6 in Appendix B) indicate that area or intensity weighting offers no median reduction in the SEs for the genes, methods involving average foreground variability reduce the SEs by between 4% and 6%, and flag filtering increases the SEs slightly (by $\approx 2\%$). For the LUS controls, intensity weights improve the precision with which the coefficients are estimated (median reduction in SEs of 13%) and weights which involve average foreground variability reduce the SEs by between 3% and 5%, while area weights and flag filtering increase the SEs by an average of 3% and 8% respectively.

### 5.3.3 Are spot and array quality weights complementary?

A final question of interest is whether array weights as described in chapter 4 can be used in conjunction with spot quality weights to any benefit. To assess this, the Spot data from
the QC data set was used, with the gene-by-gene update procedure outlined in section 4.2.4 applied to the log-ratios from the normalised LMS controls. Spot weights obtained using area, intensity, average foreground variability, (average foreground variability)/area and combined weights were used in the calculation of the array variances. The final array weights were then multiplied by the spot weights and used in the gene-wise regressions to estimate the gene expression coefficients and their SEs.

Looking at the ordinary $t$ statistics in figure 5.27, it can be seen that combining spot weights with array weights increases the $|t|$ for the known DE ratio controls more than can be obtained using a single spot weighting method alone. This is the case for all spot quality weighting schemes, which indicates that the two different approaches of obtaining weights can work effectively in combination.

5.3.4 Summary

The use of spot quality weights in the linear model analysis of the log-ratios does hold some promise. The best results were observed for the QC data set, with large changes in the $|t|$ statistics for the known DE controls achieved with little change in the null distributions of the DR. It is anticipated that applying spot weights would have a similar effect on the genes, with the test statistics of the truly DE genes also increasing in absolute terms, when poor spots are down-weighted in the analysis. This would make such genes easier to detect, which represents a gain in statistical power.

To express this in another way, given a cut-off $C$, where genes with $|t| > C$ are regarded as likely to be DE, increasing the $|t|$ values of the highly DE genes should give more truly DE genes for a given $C$ than would be obtained without weighting. If the null distribution isn’t affected, the proportion of false positive genes in this selection shouldn’t increase which is also important.

The QC data set, being quite large and gathered over an extended period of time was useful for establishing that spot weighting can be effective, however demonstration on other more typical data sets was necessary to assess the wider applicability of the method.

The weighting schemes proposed in this chapter have been demonstrated to offer modest improvements in the $|t|$ (or equivalent $|Z|$ scores) for known DE controls in the Pax5

115
Figure 5.27: Plot of \( t \) statistics before and after combining array and spot weights for the LMS controls from the QC data set (using Spot data). The array weights were estimated for the controls only, using the relevant spot weights. Improvements in the \( |t| \) of the DE ratio controls can be seen when array and spot weights are combined, compared to using spot weights alone, for all spot weighting methods.

and DC experiments, and were shown to have little effect in the smaller QDC experiment. Reductions in the SEs for the genes weren’t always observed though, especially when filtering using the GenePix ‘not found’ flags was applied. Filtering systematically added a small amount of variation to the estimated coefficients for the genes in all 3 data sets. For this reason filtering the data using the negative ‘not found’ \((-50)\) flags prior to summarisation using linear models is not recommended when no background correction has been performed.

Intensity weights, assigned using an empirical weight function derived from the QC data set were also found to offer little improvement in the \( |Z| \) scores for the DE ratio controls in the smaller data sets. On average intensity weights did not reduce the variability
of the estimated gene expression coefficients in any of the 3 data sets. Area weights also had little effect for GenePix data. This is due to limited variation in the spot sizes for these data, which is partly due to the fixed circle segmentation algorithm, but mostly a result of good printing of these arrays by the AGRF. For the other weighting methods, which all involved average foreground variability, average reductions in the SEs (relative to the SEs obtained using equal weights) of the controls and genes of between 0% to 18% were achieved. Combined (area, intensity and foreground variability) weights typically gave the best outcome.

Based on these results, combined weights would be generally recommended when spot quality weights are to be used in the analysis. However, as has been observed in the previous sections, the gains from applying spot quality weighting can be quite small.

Finally, it has been demonstrated that combining spot and array quality weights gives better results in terms of identifying true DE genes compared with using either spot or array weights in isolation. This indicates that the empirical approach for deriving array quality weights described in chapter 4 can complement the predictive spot quality weights described in this chapter.

5.4 Discussion

A central question of this chapter is whether dealing with spot quality in a graduated way using relative weights offers any improvement over using a sharp accept or reject filtering method. This can be answered by comparing the results obtained using the ‘not found’ flags from GenePix with the other alternatives. Filtering based on these flags generally improved the $t$ statistics or the equivalent $Z$-scores of the known DE controls in absolute terms compared with using equal weights. However, removing these flagged log-ratios always increased the SEs of the estimated gene expression coefficients (on average) for the genes, which is undesirable. Some of the relative weighting schemes, such as the combined (area, intensity and average foreground variability) weights do marginally better in both respects for all data sets examined, which strengthens the argument for a graduated approach to quality. A more satisfactory use of the ‘not found’ GenePix flags would be to down-weight negatively flagged spots in the analysis, using non-zero weights. Weighting
offers other benefits over filtering, such as avoiding the need to choose cut-off values for the filtering criterion, and eliminating the risk of throwing away good data, which was observed to make the task of selecting the DE controls from the QDC experiment more difficult.

The empirically derived spot weighting function for intensity, which was obtained from the QC data set, was used to assign intensity weights to Spot and GenePix data in other experiments. The benefits of intensity weighting in these experiments were negligible, particularly in terms of improving the precision of the estimates of the gene expression coefficients. This indicates that the standard curve approach is probably inadequate, with some calibration required to tailor the intensity weights to a particular data set, to reflect variations in scanner settings or other effects which may be local to an experiment.

It is worth noting that the Pax5, DC and QDC experiments are newer data sets (hybridised early-mid 2003), carried out using better protocols and with greater care than was taken with the QC arrays, which date back to 2001-2002. The images from these later arrays are much cleaner, and would be regarded as higher quality based on visual assessment. This may provide an explanation as to why the use of spot quality weights in these experiments wasn’t as effective as for the QC data; with high quality data to begin with there is less room for improvement. Testing of the spot weighting methods was however restricted to these data sets which featured some truth in gene expression through the spike-in controls, to allow systematic evaluation of the weights.

Another point worth noting is that spot weights will have a negligible impact when the characteristic chosen to weight by doesn’t vary. For instance if most spots are of the ideal spot size due to good printing (such as for the Pax5 data), using weights based on area will have little or no effect on the results. This isn’t necessarily a negative feature of the method, as the issue of quality is likely to be less of a problem in such experiments. Due to fixed circle segmentation used by GenePix Pro 4.0, the application of weights based on spot size tend not to be as relevant as they are for Spot data, where spot size varies over a considerable range. In the most recent version of the software (GenePix Pro 6.0), adaptive shape segmentation is implemented, and it would be anticipated that an area weighting scheme would be more useful on data from this version of the software.

The ideal situation for using spot weights would be in the analysis of experiments
carried out in a microarray core facility. If arrays from each print run were sampled, and routinely hybridised with mRNA from the same two sources (as per the QC data set) using a standard protocol, quality trends could be learnt using the routine hybridisations for each print-run, or between print-runs. Tailored spot weighting schemes could be devised and applied to designed experiments involving slides from these batches to improve the detection of DE genes. Process variables may change over time, and the weighting schemes could evolve and adapt to accommodate fluctuations which have a measurable effect on log-ratio variability and can be characterised by different spot specific quality measures.

As an aside, attempts to use other measures suggested in the literature as spot quality weights have also been made. For instance, the $q_{com}$ measure proposed in Wang et al. [2001] which is on a $(0,1)$ scale where lower values indicate lesser relative quality than higher values, were calculated using GenePix data from the QC experiment and used as spot weights in the linear models. The results were quite poor in terms of both criteria used to assess the other weighting schemes (i.e. $|t|$ for known DE controls tended to decrease, and SEs for the estimated coefficients increased, data not shown). Randomly assigning the combined weights to spots was also tried to answer the question of whether intelligent assignment of spot weights is important. The answer was a clear ‘yes’, with random weights tending to reduce the $|t|$ for the DE controls, and doing little in terms of reducing the SEs of the coefficients estimated for the genes (data not shown).

At the WEHI, area weights have been used most widely in the analyses of other data sets (data not shown) due to their ease of calibration on new experiments.
Concluding Comments

In this thesis, several fundamental low-level analysis issues for two-colour microarrays have been considered. The comparison of methods available in the literature for background correction, by either subtracting various estimates from the foreground signal or using adaptive model-based correction methods on real experimental data, represents the first study of its kind. Subtraction of low morphological background estimates or not background correcting the foreground signals before forming log-ratios were chosen as the best alternatives for giving low variance gene expression estimates and the highest rates of true DE genes in a data set where differential expression could be independently determined. These methods also gave the most biased estimates, and in situations where accurate log-ratios are required, model-based background correction methods are recommended, such as the normal-exponential convolution model proposed in this thesis. This model is widely applicable to data from any image analysis package, and was shown to offer very good results in terms of selecting DE genes. The standard practice of subtracting high local background estimates, such as GenePix median background, performed worst in several criteria, and is not recommended. Future work on this topic would extend the comparison to include some of the variance stabilising data transforms suggested in the literature, to answer the question of whether the different approaches to the same problem give similar results (which might be anticipated a priori).

A quantitative approach to quality based on weights which give less influence to low-precision measurements in the linear model analysis of the log-ratios was also described. This method is quite different to the popular filtering based approaches cited in the literature, which are often ad-hoc and can result in throwing away good data. For array weights, the precisions of different microarrays from a designed experiment were obtained
empirically by looking for agreement between replicate arrays. Dispersion models, fitted with array variance parameters were used for this purpose, and a novel algorithm to estimate these parameters was necessary. The array variances were used as inverse weights, and these weights tended to be lower for suspect arrays with artifacts, and higher for good quality arrays.

Predictive spot specific weights based on area, intensity, foreground variability and combinations of these three measures were learned from a large data set and then applied to new experiments. Varying degrees of improvement in detecting true differential expression were observed using different spot weighting schemes, with combined weights tending to give the greatest improvements. In some situations, applying array or spot weights was shown to have no effect, which was usually because of the high quality of the data to begin with. Although the two methods for deriving array weights and spot morphology weights are quite different, they have been shown to be able to make a difference, in terms of giving slightly more power to detect differential expression than using equal weights, both individually and when combined. They have also been shown to perform better than particular filtering methods, and are less ad-hoc as they avoid the need to choose a cut-off value. Weights are however only useful in replicated microarray experiments where the data needs to be summarised across multiple arrays. Importantly, all of the methods described in this thesis are easily implemented in existing statistical software, which allows them to be routinely used in the analysis of two-colour microarray data. This thesis has demonstrated the usefulness of different statistical models, which can offer benefits over other less systematic approaches, in microarray data analysis, and also highlights the need for careful pre-processing of the data.
Bibliography


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


128


Appendix A

R Code from section 4.2.4, chapter 4.

```r
estimate.gammas <- function(M, design = NULL, weights = NULL)
{
  ngenes <- dim(M)[1]
narrays <- dim(M)[2]
  if (is.null(design)) {
    design <- matrix(1, ncol(M), 1)
  }
  # Set up design matrix for array variance model
  arrays <- as.factor(1:narrays)
  contrasts(arrays) <- contr.sum(levels(arrays))
  Z <- model.matrix(~as.factor(arrays))
  Z <- Z[,-1]
  # Intialise array variances to zero
  arraygammas <- rep(0, (narrays-1))
  # Estimate array variances via one-step update
  Zinfo <- 10*crossprod(Z)
  if(!is.null(weights)) {
    for(i in 1:ngen)
      if(max(weights[i,], na.rm=TRUE) > 1) {
        weights[i,] <- weights[i,]/max(weights[i,])
      }
  }
}
```

133
for(i in 1:ngenes) {
    vary <- exp(Z%*%arraygammas)
    if(!is.null(weights)) {  # combine spot & running weights
        w <- as.vector(1/vary*weights[i,])
    } else {
        w <- as.vector(1/vary)
    }
    y <- as.vector(M[i,])
    obs <- is.finite(y) & w!=0
    if (sum(obs) > 0) {
        if(sum(obs) == narrays) {
            X <- design
            Z2 <- Z
        } else {  # remove missing/infinite values
            X <- design[obs, , drop = FALSE]
            y <- y[obs]
            vary <- vary[obs]
            w <- w[obs]
            Z2 <- Z[obs, , drop = FALSE]
        }
        out <- lm.wfit(X, y, w)
        d <- w*out$residuals^2
        Q <- qr.Q(out$qr)
        nparms <- dim(Q)[2]
        h <- as.vector(Q^2 %*% array(1, c(nparams, 1)))
        Zinfo <- Zinfo + crossprod(Z2, (1-h)*Z2)
        R <- chol(Zinfo)
        zd <- (d-(1-h)*vary)/vary
    }
}

134
Zzd <- crossprod(Z2, zd)

gammas.iter <- backsolve(R,
  backsolve(R, Zzd, transpose = TRUE))

arraygammas <- arraygammas + gammas.iter

} } list(weights=1/exp(Z*arraygammas), gammas=Z*arraygammas) }
Appendix B

![Graph showing Perimeter and Circularit](image)

Figure B.1: Reproducibility trends for perimeter and circularity measures for Spot image analysis output.
Figure B.2: Average Signal-to-noise ratio (SNR) versus spot intensity \((A)\).

Figure B.3: Ratio of SEs of coefficients with:without spot weights for the genes (left) and LUS controls (right) from the Pax5 experiment (Spot data).
Figure B.4: Ratio of SEs of coefficients with:without spot weights for the genes (left) and LUS controls (right) from the Pax5 experiment (GenePix data).

Figure B.5: Ratio of SEs of coefficients with:without spot weights for the genes (left) and LUS controls (right) from the DC experiment.
Figure B.6: Ratio of SEs of coefficients with:without spot weights for the genes (left) and LUS controls (right) from the QDC experiment.
Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Ritchie, Matthew Edward

Title:
Quantitative quality control and background correction for two-colour microarray data

Date:
2004-12

Citation:

Publication Status:
Unpublished

Persistent Link:
http://hdl.handle.net/11343/38930

File Description:
Quantitative quality control and background correction for two-colour microarray data

Terms and Conditions:
Terms and Conditions: Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.