Exploring statistical methods for analysis of microarray data

by

Tanzy Mae Tallapoosa Paz Love

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Alicia L. Carriquiry, Major Professor
Dan Nettleton
Karin Dorman
Stephen Vardeman
Stephen Howell

Iowa State University

Ames, Iowa

2005

Copyright © Tanzy Mae Tallapoosa Paz Love, 2005. All rights reserved.

UMI Number: 3184637

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI[®]

UMI Microform 3184637

Copyright 2005 by ProQuest Information and Learning Company.

All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Graduate College
Iowa State University

This is to certify that the Doctoral dissertation of
Tanzy Mae Tallapoosa Paz Love
has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

~~Committee~~ Member

Signature was redacted for privacy.

~~Committee~~ Member

Signature was redacted for privacy.

~~Committee~~ Member

Signature was redacted for privacy.

~~Committee~~ Member

Signature was redacted for privacy.

Major Professor

Signature was redacted for privacy.

For the Major Program

DEDICATION

I would like to dedicate this paper to my mother, Dr. Julia Norton, whose example has been all things to me in my work. I would also like to thank my father, Dr. John Lovell, brothers, Robert and John Norton, and grandparents, Betty Hawkins and Robert Norton, for their love and support during the writing of this work.

TABLE OF CONTENTS

LIST OF TABLES	viii
LIST OF FIGURES	ix
ABSTRACT	xiii
CHAPTER 1. GENERAL INTRODUCTION	1
1.1 Introduction	1
1.2 Microarrays and gene expression	2
1.2.1 Various array technologies	3
1.2.2 Image analysis	4
1.2.3 Normalization	8
1.3 Literature Review	11
1.3.1 Log-Normal Models	11
1.3.2 Hierarchical Mixture Models	13
1.3.3 Nonparametric Models	17
1.4 Thesis organization	17
1.4.1 Embryogenic maize tissue	17
1.4.2 Bayesian model for combining scans	18
1.4.3 Clustering posterior distributions	19
1.5 References	20

CHAPTER 2. GENE EXPRESSION PATTERNS DURING SOMATIC EMBRYO DEVELOPMENT AND GERMINATION IN MAIZE

Hi II CALLUS CULTURES	23
2.1 Abstract	23
2.2 Introduction	24
2.3 Materials and methods	26
2.3.1 Materials and tissue culture methods	26
2.3.2 RNA extraction and microarray analysis	27
2.3.3 Microarray data analysis	30
2.3.4 Data normalization	31
2.3.5 Estimation of treatment effects	33
2.4 Results	34
2.4.1 Somatic embryogenesis and expression profiling	34
2.4.2 Gene expression patterns following induction of somatic embryo maturation	35
2.4.3 Line variation	37
2.4.4 Comparison of embryoid enriched and total callus	38
2.5 Discussion	38
2.6 Acknowledgements	41
2.7 References	41
2.8 Figure captions	45

CHAPTER 3. INCORPORATING MULTIPLE cDNA MICROARRAY SLIDE SCANS - APPLICATION TO SOMATIC EMBRYOGENESIS IN MAIZE

3.1 Abstract	54
------------------------	----

3.2	Introduction - An experiment to assess gene expression changes during maize embryogenesis	55
3.3	Microarray analysis	60
3.3.1	Multiple laser and sensor settings	62
3.4	Bayesian Hierarchical Gamma Model	64
3.4.1	Likelihood Function	65
3.4.2	Prior Distributions	69
3.4.3	Estimating the hyperparameters	71
3.4.4	Posterior distributions	76
3.5	Performance assessment via a simulation experiment	77
3.5.1	Estimation of scaling parameters	78
3.5.2	Estimation of gene expression	78
3.5.3	Improvements over estimates that rely on a single reading	81
3.6	Maize embryogenesis experiment	82
3.6.1	Single gene-level inferences	82
3.6.2	Gene expression profiles	83
3.6.3	Inferences about time and line effects	84
3.7	Discussion	92
3.8	References	97
3.9	Appendix	99
3.9.1	Marginal likelihood of (a, a_0, ν)	99
3.9.2	Marginal likelihood of (a, a_0) conditional only on largest reading	99
3.9.3	Expectation-Maximization to estimate (a, a_0)	100
 CHAPTER 4. CLUSTERING POSTERIOR DISTRIBUTIONS -		
	APPLICATION TO SOMATIC EMBRYOGENESIS IN MAIZE	102
4.1	Introduction	102

4.2	Proposed clustering methods	106
4.3	Model and posterior clustering for two treatments	108
4.3.1	Bayesian Hierarchical Gamma Model	108
4.3.2	Clustering the ρ 's with standard methods	112
4.3.3	Clustering the posterior distributions of the ρ_i 's	114
4.4	Model and posterior clustering for m treatments	116
4.4.1	Bayesian Hierarchical Gamma Model	117
4.4.2	Clustering the Ψ vectors	123
4.5	Maize embryogenesis	123
4.5.1	Example using two treatments	124
4.5.2	Example using three treatments	131
4.6	Discussion	136
4.7	References	141
4.8	Appendix	145
4.8.1	Posterior distribution of z in the two treatment case	145
4.8.2	Posterior distribution of z in the m treatment case	146
CHAPTER 5. GENERAL CONCLUSIONS		147
5.1	General Discussion	147
5.2	Recommendations for future research	151
5.2.1	Related technologies	151
5.2.2	Improved and consistent preprocessing of expression values	152
5.2.3	Intensity-dependent scaling of multiple scans	152
5.2.4	Blocked or covariate effects in the hierarchical model	152
5.2.5	Choosing the appropriate number of clusters using distribution divergences	153
ACKNOWLEDGEMENTS		154

LIST OF TABLES

Table 3.1	Values of the hyperparameters used in simulation experiment. .	78
Table 3.2	Comparing expression estimation methods. Simulation results for 1000 genes after 100 simulations.	79
Table 3.3	Average range in simulation.	81
Table 3.4	Comparing hierarchical models. Simulation results for 1000 genes after 100 simulations.	95
Table 4.1	Mean of the posterior distributions of the member of each cluster and membership (out of 713 total) in parentheses.	130
Table 4.2	Marginal means of the posterior distributions of the member of each cluster and membership size (out of 866 total).	137

LIST OF FIGURES

Figure 1.1	Censored expression reading from a cDNA microarray slide. (a) A medium level scan compared to a low scan shows the low scan is censored for low expression values. (b) A medium level scan compared to a high scan shows the high scan is censored for high expression values.	5
Figure 1.2	The measures M and A from a cDNA microarray, before and after normalization.	9
Figure 1.3	The two dye channels on a cDNA microarray, before and after normalization.	10
Figure 1.4	Boxplots of the 32 print tip blocks on a cDNA microarray, arranged by the four rows. Note that the groups do not have the same median and that the fourth and eighth columns are elevated in every row.	10
Figure 2.1	Somatic embryo development in maize Hi II callus.	47
Figure 2.2	Loop design for microarray hybridization experiments.	48
Figure 2.3	Summary of gene expression during somatic embryo maturation and germination in maize.	49
Figure 2.4	Genes with different expression profiles.	50
Figure 2.5	Expression profiles for four genes belonging to classes of genes expressed during zygotic embryo development or germination. . .	51

Figure 2.6	Line variation in gene expression.	52
Figure 2.7	Trends in expression of genes in various functional categories show significant time variation during embryo maturation and germination.	53
Figure 3.1	The time course experiment for somatic embryogenesis in maize.	57
Figure 3.2	The microarray double loop design with dye swap for the maize embryogenesis experiment. Each box represents one of the 30 slides created. The arrows show the direction of the loop as each time point is compared to its neighbors.	58
Figure 3.3	(a) Many spots are censored above in the higher reading. (b) Many spots are censored below in the lower reading.	63
Figure 3.4	Log true expression by average bias over 100 simulations for expression values less than 65535.	80
Figure 3.5	Errorbar plots for some example genes from Che et al. (2004) (Fig 4).	85
Figure 3.6	Errorbar plots for the example genes from Che et al. (2004) (Fig 4) after using all readings.	86
Figure 3.7	Errorbar plots for some example genes from Che et al. (2004)(Fig 6). The top two genes have significant differences in gene expression between the three lines and the line difference are not significant in the bottom two genes.	87
Figure 3.8	Errorbar plots for the example genes from Che et al. (2004)(Fig 6) after using all readings. The top two genes have significant differences in gene expression between the three lines and the line difference are not significant in the bottom two genes. . . .	88

Figure 3.9	Numbers of genes with fold changes in expression from Che et al. (2004)(Fig 3).	90
Figure 3.10	Numbers of genes with fold changes in expression from Che et al. (2004)(Fig 3) after using all readings.	91
Figure 4.1	Posterior distributions of ρ for four genes. These have been calculated analytically. The top two panels show the posterior distributions of ρ for two genes whose expressions do not change significantly between day 7 and day 14 and the bottom two panels correspond to genes which are differentially expressed. . . .	125
Figure 4.2	The distribution of samples from the posterior distribution of ρ (on log scale) for genes in the five clusters created by clustering the log posterior mean of ρ	126
Figure 4.3	The distribution of samples from the posterior distribution of ρ (on log scale) for genes in the five clusters created by clustering each $\log(\rho_b)$, the 1000 MCMC samples from the posterior distribution of ρ	127
Figure 4.4	The distribution of samples from the posterior distribution of ρ (on log scale) for genes in the five clusters created by clustering the posterior distributions of ρ , evaluated analytically.	128
Figure 4.5	The distribution of samples from the posterior distribution of ρ (on log scale) for genes in the five clusters created by clustering the posterior distributions of ρ , evaluated empirically.	129

- Figure 4.6 Posterior distributions of Ψ for two genes. These have been approximated from the 1000 MCMC samples from the posterior distribution. The first panel shows the posterior distribution of Ψ for a gene which is relatively continuous and the second panel corresponds to a gene with a noncontinuous posterior distribution for Ψ 132
- Figure 4.7 The distribution of samples from the posterior distribution of Ψ (on log scale) for genes in the four smallest clusters created by clustering the log posterior mean of Ψ 133
- Figure 4.8 The distribution of samples from the posterior distribution of Ψ (on log scale) for genes in the four smallest clusters created by clustering the log of samples from the posterior of Ψ 134
- Figure 4.9 The distribution of samples from the posterior distribution of Ψ (on log scale) for genes in the four smallest clusters created by clustering the posterior distributions of Ψ , evaluated empirically. 135

ABSTRACT

The expansion of molecular biology in recent years has created an increasing amount of data and interest in specific tools to analyze them. Much of these data come from a class of high-throughput technology that measures hundreds or thousands of variables at the same time. One such high-throughput technology currently in use is microarray technology. The three major objectives in expression analysis are data preprocessing, identifying differential expression, and grouping genes by common behavior. Extracting the useful information on gene expression from the available output is not trivial. The data collection process is quite noisy in that non-biological bias may be introduced at a number of points by the operators or the technology. Identifying differential expression is an important step in reducing the number of variables, p , of interest to a reasonable scale. It requires distinguishing random variation in expression measurements from signal of interest. Most statistical research so far has focused on this problem and many methods exist for making the determination. Finally, grouping genes has biological importance in identifying the purpose of unidentified genes and the interconnections between biological systems. We focus on achieving the first and last of these objectives while using relatively standard methods for the second one.

CHAPTER 1. GENERAL INTRODUCTION

1.1 Introduction

The development of high-throughput technologies in recent years has created an increasing amount of data and interest in specific tools to analyze them. Much of these data measure hundreds or thousands of variables at the same time. Often, the expense incurred in implementing these technologies is such that the number of subjects that are measured is much smaller than the number of variables. This is known as the “large p , small n ” problem because most statistical methods rely at least in theory on having more experimental units than variables measured on each unit.

One such high-throughput technology currently being used is microarray technology. It shares many of the attributes of other molecular biology technologies but also has several specific characteristics. By developing methods for preprocessing and analyzing microarrays, we hope not only to aid analysis of these data, but also to create tools that are more widely useful in similar problems.

The three major objectives in expression analysis are data preprocessing, identifying differential expression, and grouping genes by common behavior. Extracting the useful information on gene expression from the available output is not trivial. The data collection process is quite noisy in that non-biological bias may be introduced at a number of points by the operators or by the technology. Identifying differential expression is an important step in reducing the number of variables of interest, p , to a reasonable scale.

It requires distinguishing random variation in expression measurements from signal of interest. Most statistical research so far has focused on this problem and many methods exist for making the determination. Finally, grouping genes has biological importance; it helps in identifying the purpose of unidentified genes and the interconnections between biological systems. We focus on achieving the first and last of these objectives while using relatively standard methods for the second one.

This chapter is organized as follows. In Section 1.2 we give an overview of microarray technology. In Section 1.3 we review the literature on gene expression analysis. The organization of this thesis is described in Section 1.4.

1.2 Microarrays and gene expression

Microarray studies are carried out to investigate the complex processes related to gene expression. Typically, transcripts from mRNA extracted from biological materials subjected to different treatments, derived from different tissues, or obtained from the same tissue at different stages of development are used to probe arrays. The objective is to draw inferences about differential gene expression levels across treatments, tissues or developmental stages. Gene expression data from microarrays can only be compared across different arrays after the appropriate background cleaning and normalization procedures have been performed on the data. There are several important sources of variation in gene expression measurement that must be accounted for in statistical analyses, and much of this variation is array specific.

Although all non-reproductive cells in an organism have the same DNA, how they utilize DNA is different. Genes are used by being transcribed into mRNA within the nucleus. Then the mRNA is translated to proteins on ribosomes in the cytoplasm. The complex functions of these proteins is a major area of current research in many

laboratories.

Gene expression can be determined by measuring mRNA levels in cells or tissues under study. Protein expression can likewise be measured in a similar way. These are not equivalent measures because not all mRNA molecules are translated into proteins after they are produced. Also, some proteins may build up over time and remain after all corresponding mRNA has degraded.

1.2.1 Various array technologies

There are a number of technologies that have been created to estimate levels of specific mRNA's in a high-throughput manner. Generally, large replicates of selected cDNA sequences are affixed to a medium and labeled cDNA created from *in vivo* mRNA samples are hybridized to their complementary sequences. Variations include labeling the samples with radioactive labels or cyanine dyes, and hybridizing them to cDNA affixed to nylon membranes, glass slides, or printed chips. Depending on the label and medium, each array unit (membrane, slide, or chip) is capable of measuring mRNA levels in one or two samples. cDNA slides and oligonucleotide chips are currently the most popular methods.

The slides used in cDNA microarrays are printed by a computer using a robot with a set of print-tips at the end of an arm. All of the printing tips on the arm are dipped into vials of cDNA's and pressed to the surface to spot cDNA's onto the slide. The arm then goes back to pick up cDNA from the next set of vials and prints these adjacent to the last spot. In this way, each tip on the arm prints a block of spots adjacent to each other on the slide. The spots in one block are created by the same tip and may have similarity related to their spatial proximity so we are suspicious of correlation between these expression measurements.

Affymetrix corporation produces oligonucleotide arrays called gene chips for many organisms. On a chip, each spot (probe cell) is generated by photolithography with replicates of a gene probe 25 bases long. A partner spot for each probe is generated with a sequence of bases, called the mismatch, identical to the probe except at the 13th (middle) base, which is changed to its complement. For each gene selected for representation, a number (usually 16-20) probe spots are generated along with their mismatch spots and this group is called the probe set. Since the probes are generated a base at a time (the first nucleotide of the sequence, then the second, ...) we expect no print tip or printing time effects as one might see in cDNA microarrays.

Depending on the technology used, the complexity and importance of the experimental design will vary. Long standing methods of randomization of treatments to units should always be implemented to prevent confounding of treatment effects with production order and other experimental effects. It is also generally the case that true (biological) replication will improve the ability to generalize conclusions drawn in an experiment. A special situation exists when two samples can be applied to the same array and further thought must be applied in these situations, see Kerr and Churchill (2001) and Dobbin et al. (2003).

1.2.2 Image analysis

The data produced by microarrays are generally in the form of a picture file (such as TIFF format) with an intensity value for each pixel. Radio labeled microarrays are read with a phosphorimaging instrument. Microarrays labeled with cyan dye probes, such as cDNA or oligonucleotide microarrays, are excited with lasers so that the dye fluoresces. In arrays with two different dyes different laser frequencies are used to excite the two different fluorescent dyes. A fluorescence scanner reads the array and records an intensity value for each pixel.

Figure 1.1 Censored expression reading from a cDNA microarray slide. (a) A medium level scan compared to a low scan shows the low scan is censored for low expression values. (b) A medium level scan compared to a high scan shows the high scan is censored for high expression values.

Generally, the instrument used to record intensity values has a smaller dynamic range than the actual range of intensities on the microarray. Scans taken with different scanner settings measure different parts of the full range. The relatively smaller size of the measurable range results in censoring of data in the recorded pixels. All pixel readings over the upper threshold will be censored at the threshold value and all readings below the lower threshold will be censored at 0. This censoring leads to a loss of information about differential expression. Figure 1.1 shows how information is lost for lowly expressed genes in low level scans and for highly expressed genes in high level scans.

Converting the images into numerical observations for further analysis usually consists of several parts.

- Segmentation - assignment of pixels to gene spots
- Signal Summary - numerical summary of all pixels in a spot
- Background Selection - identification of background

- Background Correction - correction of signal observations

Segmentation refers to the process of determining which pixels belong to which spots. The method used to create the array gives us an estimate of spot locations and shapes. A simple method involves fitting a circle of uniform size to each spot. More complicated methods include circles of varied sizes and seeded region growing. In all cases, a computer algorithm chooses the pixels most likely to be in the spot by contrasting the (high) intensity of pixels in the spot with the (low) intensity of pixels outside of the spot. Research has shown that statistical results are not sensitive to segmentation methods (Yang et al. 2002).

Given a collection of pixels within one spot, each will have a different intensity value; a single summary statistic is needed for each spot. The mean pixel intensity is usually used as the estimate of signal because the spot selection algorithms create fairly similar pixels within a spot. The mean has less measurement variability than the individual pixel values. However, using the mean pixel intensity masks the censoring (above or below) of individual pixels; these censored pixels bias the mean.

Dust, imperfections in the surface, or labeled cDNA sticking to the medium can produce some intensity on the array where no cDNA was spotted. This fluorescent intensity is not considered part of the signal we wish to measure; it is background fluorescent intensity. Topological variation in the array and other artifacts cause the background to vary across the surface of the medium. A local region of background pixels is picked for each spot. This can be done by assigning to the spot background those pixels in a band around the circle or a square of pixels farthest from the neighboring spots. There is some evidence that statistical conclusions are sensitive to background selection methods (Yang et al. 2002). One possible reason may be that pixels with fluorescence from dyed cDNA, which are relatively intense, are sometimes assigned to

the background. These might cause a misleading background estimate.

The background fluorescence is assumed to be additive. The median pixel intensity for each group of local background pixels is frequently used as the measure of the background intensity because this estimate is robust against a few pixels with fluorescence from dyed cDNA included in the background pixels. These pixels would over influence the mean of the background pixel intensities. The local background value is subtracted from signal before any analysis takes place. This can result in negative expression values which are routinely set to zero or removed from the analysis.

For gene chip analysis, the raw data are intensity measurements for each spot and these are grouped into perfect match/mismatch pairs and probe sets. The proprietary Affymetrix GeneChip software generates an estimate of relative gene expression for each probe set called the Microarray Analysis Suite (MAS) 5.0 Signal. First, the intensities of the perfect match (PM) and mismatch (MM) cells are calculated by subtracting a position-specific background value from the raw cell intensities, giving (PM_p, MM_p) for $p = 1, \dots, n_p$, where n_p is the number of probe pairs. The ideal mismatch (IM) value is computed for each cell in such a way that it is always less than the perfect match value. The probe value, V_j , for the j^{th} probe pair is the maximum of $\log(PM_j - IM_j)$ and -20 . The signal log value (SLV) of the probe set is the one-step Tukey BiWeight of the n_p probe values $(PV_1, \dots, PV_{n_p})$. The MAS 5.0 Signal for the probe set is the SLV scaled by a constant times the trimmed mean of the SLV's of all probe sets on the chip (Affymetrix Inc. 2002). Other methods for summarizing probe sets have been proposed such as $X = \frac{1}{n_p} \sum_{p=1}^{n_p} (\log P_p - \log M_p)$ (Efron et al. 2001).

These background corrected gene expression intensities are the measures of gene expression we shall analyze.

1.2.3 Normalization

Normalization refers to procedures that allow for comparison between expression measurements taken using different labels or arrays. The appropriate normalization procedures are different for different technologies. We will briefly describe the more popular methods.

There are several reasons why normalization is necessary in cDNA microarray experiments. Various sources (climate, operator, time) contribute to the experimental variation between slides (Richardson et al. 1997); the differences between the dyes (the Cy3 dye tends to be more excitable than Cy5) is yet another source (Yang et al. 2002). Normalization can also remove artifacts from spatial and experimental sources such as print tips and uneven washing of slides. We will denote the expression intensities measured for the Cy5 and Cy3 dyed samples of the i th gene on a slide by R_i and G_i , respectively, for the $i = 1, \dots, n$.

We assume that most of the genes are not differentially expressed, i.e. $R_i = G_i$ for most i or $M_i = \log(R_i/G_i)$ should be centered at 0. The M_i 's are biased away from 0 (as in Figure 1.2(a)) because the dyes fluoresce differently. Also, M_i is empirically dependent on intensity ($R_i * G_i$). It is convenient to compare M_i to $A_i = \log \sqrt{R_i * G_i}$ to examine this relationship; (M_i, A_i) is a 45° rotation of $(\log(R_i), \log(G_i))$. In general, the relationship will be non-linear and of no interest to the biological questions at hand. To remove the intensity-dependent dye bias of expression values Smyth, Yang, and Speed (2002) adjust M_i by the loess fit of A_i , see Figure 1.2(b). Locally weighted polynomial regression, called loess, is a smoothing method that estimates a line of trend through a dataset by combining polynomial regressions from small subsets of the data (Cleveland 1979). Unfortunately, this method produces normalized values of M_i (which are used in the further analysis of Smyth et al. (2002)), but not of R_i and G_i .

Figure 1.2 The measures M and A from a cDNA microarray, before and after normalization.

The loess-fit normalization can be modified to create normalized values of R_i and G_i for all i . First, R and G are standardized by the total dye channel intensity, $R_i^* = VR_i / \sum R_i$ and $G_i^* = VG_i / \sum G_i$, where V is a chosen constant value. This has often been used as the only normalization step (as in Newton et al. 2001), but it does not correct for dye bias or intensity dependence. We then multiply the G_i^* by $\exp(c(A_i))$ for all i , where $c(\cdot)$ is the loess fit of M on A . This is equivalent to subtracting $c(A_i)$ from M_i . The resulting corrected R and G are corrected for intensity dependent dye bias and are comparable between slides. Notice in Figure 1.3(a) that the ranges of R and G are different and the observed data are biased off the diagonal line. After normalization, the data are centered around the diagonal line and have the same range of values. Another proposed method is simply correcting R and G each by half of the correction for M (Wit and McClure 2004).

As mentioned earlier, the production of cDNA microarray slides leads us to believe that there may be effects on expression measurements from print tips, and print tip groups are also surrogates for spatial effects such as uneven washing. It is clear in Figure 1.4 that the 32 print tip groups on this slide do not have the same median. However, we assume that most of the genes in any print tip group are not differentially

Figure 1.3 The two dye channels on a cDNA microarray, before and after normalization.

Figure 1.4 Boxplots of the 32 print tip blocks on a cDNA microarray, arranged by the four rows. Note that the groups do not have the same median and that the fourth and eighth columns are elevated in every row.

expressed, i.e. $R_i = G_i$ for most i as before, and that there is no biological reason for differences in print tip groups. Thus a possible normalization approach is to perform the loess fit separately for each print tip group (Smyth et al. 2002). The normalization procedure is then

1. Standardize R and G by channel intensity.
2. Replace G_i by $G_i * \exp(c_j(A_i))$ for all i , where $c_j()$ is the loess fit for the j th print tip group.

The normalized R and G are corrected for dye, slide, print tip group and intensity effects.

The data obtained from image analysis of photolithography oligonucleotide chips

are not intended to be analyzed without preprocessing and normalization. Many chip-specific non-biological effects (label quantity, sample quantity, operator, ...) can alter expression readings from printed gene chips (Ibrahim et al. 2002). Only one sample is hybridized to each chip, so the data from each chip must be normalized before being compared to data on a different chip.

The simplest method of normalization is the same as that for two color arrays, standardizing by total intensity (as in Ibrahim et al. (2002)). Other methods require a set of invariant (constitutively expressed) genes that can be used as “anchors.” Given that the value of these genes should be similar across all arrays in the experiment, all expression measurements are scaled to reduce the variation in the invariant set. However, this requires accurate knowledge of constitutively expressed genes which is may not available.

1.3 Literature Review

We provide a summary of the recent literature on Bayesian methods for analyzing microarray data. To standardize notation, we denote observed expression values by y_{ijk} for the i^{th} gene under the j^{th} treatment on the k^{th} replication where $i = 1, \dots, n, j = 1, \dots, m$, and $k = 1, \dots, r$.

1.3.1 Log-Normal Models

Baldi and Long (2001) and Ibrahim et al. (2002) both use the parametric model that assumes that expression values are approximately normally distributed after a log transformation. Observed gene expressions are generally right skewed and their variance tends to increase with the mean; thus, the log normal model may be a justifiable choice. To carry out a Bayesian analysis, a prior distribution for model parameters is needed.

In all cases, the authors choose to model the log expression for each treated gene, $x_{ijk} = \log(y_{ijk})$, as having its own mean, μ_{ij} , and variance, σ_{ij}^2 , for all i, j , and k . Here, the parameter μ_{ij} represents the true log expression level of gene i subjected to the j th treatment. The amount of variation in our observation is measured by σ_{ij}^2 and varies for genes and treatments. Baldi and Long (2001) and Ibrahim et al. (2002) also choose to model the n distinct $(\mu_{ij}, \sigma_{ij}^2)$ as generated from a common distribution. This assumption equates with assuming that there is an underlying population of possible expression levels that generates the true mean expression of each experimental condition. Thus, the model proposed by Baldi and Long (2001) and by Ibrahim et al. (2002) is hierarchical.

Baldi and Long (2001) assume that the parameters $(\mu_{ij}, \sigma_{ij}^2)$ come from the conjugate prior family. That is $\mu_{ij} | \sigma_{ij}^2 \sim N(\mu_{0j}, \sigma_{ij}^2 / \lambda_{0j})$ and $\sigma_{ij}^{-2} \sim \Gamma(\nu_{0j}, \sigma_{0j}^2)$ for all i and j . This structure implies *a priori* dependence between μ_{ij} and σ_{ij}^2 which is reasonable in microarray data. They use this model to derive the posterior mean values $\hat{\mu}_{ij} = \bar{x}_{ij}$ and $\hat{\sigma}_{ij}^2 = \frac{\nu_{0j}\sigma_{0j}^2 + (r-1)s_{ij}^2}{\nu_{0j} + r - 2}$ where s_{ij}^2 is the variance of x for gene i at treatment j . Restricting attention to two treatment experiments, $m = 2$, they then perform t -tests on each gene using $\hat{\mu}_{ij}$ as the mean estimates, $\hat{\sigma}_{ij}^2$ as the variance estimate, and $r + \nu_0 - 2$ as the degrees of freedom. In practice, therefore, the hyperparameters ν_{0j} and σ_{0j}^2 must be specified. They implement this procedure in the Cyber-T package with $\nu_{0ij} = 10 - r_i$ and σ_{0ij}^2 is the pooled variance of the 101 genes centered (when ranked by mean expression level) at gene i . This results in n p-values that can be used to rank the genes in order of evidence of differential expression or used to pick the subset of significant genes for a given α level.

Ibrahim et al. (2002) consider the y_{ijk} to be a mixture of a discrete component (data censored below) and a continuous component. That is $y_{ijk} = c_0$ with probability p_{ij} and $y_{ijk} = c_0 + y_{ijk}^*$ with probability $1 - p_{ij}$. The variable $\delta_{ijk} = 1$ if $y_{ijk} = c_0$ and 0

otherwise. They then assume $x_{ijk} = \log(y_{ijk}^*)$ has a Normal distribution. Using the same conjugate priors with different hyperparameters, they have $\mu_{ij}|\sigma_{ij}^2 \sim N(\mu_{0j}, \tau_0\sigma_{ij}^2/\bar{n}_j)$ and $\sigma_{ij}^{-2} \sim \Gamma(\nu_{0j}, \sigma_{0j}^2)$ for all i and j , where $\bar{n}_j = \frac{1}{n} \sum_{i=1}^n (r_j - \sum_{k=1}^{r_j} \delta_{ijk})$. Additionally, a prior must be placed on p_{ij} , namely $e_{ij} = \text{logit}(p_{ij}) \sim N(u_{0j}, k_{0j}w_{0j}^2)$. One more level of priors is introduced where $\mu_{0j} \sim N(m_{0j}, \nu_{0j}^2)$ introduces prior correlation between genes for a given individual and treatment and $\sigma_{0j}^2 \sim \Gamma(q_{0j}, t_{0j})$ $u_{0j} \sim N(\hat{u}_{0j}, h_{0j}w_{0j}^2)$ increase flexibility in the model. They suggest the values for the hyperparameters which can be used to complete the model. Restricting attention again to the two treatment case, the model focuses on $\xi_i = \frac{E(y_{i1\bullet})}{E(y_{i2\bullet})}$, the ratio of the expected expression values. The posterior distribution of ξ_i can be calculated from the model and all genes with $P(\xi_i > 1|y, \delta) \geq \gamma_0$ or $P(\xi_i > 1|y, \delta) \leq 1 - \gamma_0$ are declared differentially expressed, i.e. $\mu_{i1} \neq \mu_{i2}$, otherwise $\mu_{i1} = \mu_{i2} = \mu_i$. Different values of γ_0 will create different submodels. These are compared using the L measure, where the model with the smallest L measure is deemed the best-fitting model. The L measure compares the data, y , to a future observation, z , with the sampling density of the model, $L = E[(z - y)'(z - y)]$. In practice, L can be computed as posterior expectations; these can be evaluated by using MCMC methods to sample the parameters from their posterior distributions. The $P(\xi_i > 1|y, \delta)$ can be used to rank the genes in order of evidence of differential expression or the model with the smallest L measure will identify a subset of significant genes.

1.3.2 Hierarchical Mixture Models

Broët et al. (2002), Kendzierski et al. (2003), Smyth (2004), and Tai and Speed (2004) propose the use of hierarchical mixtures to capture the fact that some genes have constant expression across treatments while others have different expression over time. Letting $y_{i\bullet\bullet}$ denote the observed expression values for gene i measured for all treatments and replications, its distribution can be represented as a weighted average of G distribution functions; $f(y_{i\bullet\bullet}) = \sum_{g=1}^G w_g f_g(y_{i\bullet\bullet}|\theta_{ig})$ where θ_{ig} are the parameters of

the distribution $f_g(\cdot)$.

Broët et al. (2002) restrict their attention to two treatments and build their model on $d_i = \text{mean}(x_{i1k} - \bar{x}_{\bullet 1k}) - \text{mean}(x_{i2k} - \bar{x}_{\bullet 2k})$ where $x_{ijk} = \log(y_{ijk})$ as before and $\bar{x}_{\bullet jk} = \frac{1}{n} \sum_{i=1}^n x_{ijk}$ denotes the average of log expression over the genes. They model the values of d_i as coming from a mixture of distributions; $f(d_i) = \sum_{g=1}^G w_g f_g(d_i | \theta_{ig})$, where G is unknown. These transformed data are assumed to arise from a mixture of normal distributions, so f_g is a normal distribution and $\theta_{ig} = (\mu_g, \sigma_g^2)$. They place the following priors on the model: $\mu_g \sim U(\min(d_i), \max(d_i))$, $\sigma_g^{-2} \sim \Gamma(2, \beta)$, $w \sim \text{Dirichlet}(\delta, \dots, \delta)$, and $G \sim DUnif(1, g_{\max})$. One more level of the model is added where $\beta \sim \Gamma\left(0.2, \frac{10}{(\min(d_i) - \max(d_i))^2}\right)$. Sophisticated reversible-jump Metropolis-Hastings methods are used to explore the support given by the data to different numbers of subgroups, G , and parameter values. A large value of G is chosen from among those fitting the data well and the posterior probability of membership in each of these $\hat{G}$ clusters is calculated and each gene is assigned to the component in which it has the highest posterior probability of membership. It is to be expected that a large number of the genes will not be differentially expressed between the two treatments and this will result in components with mean close to zero. Genes assigned to component distributions with means distant from zero can be considered differentially expressed.

Kendzierski et al. (2003) expand their earlier work using mixture modeling of either gamma or log-normal distributions to describe the observed expression values. The use of a gamma distribution for expression values is supported by some experimental evidence that biological intensities fit gamma distributions. As both distributional assumptions seem reasonable, the authors suggest that different datasets may be more suited to one model or the other. In either case, for gene i the m treatments are assumed to be partitioned into $t(g)$ groups in mixture component g . The number of components is chosen

from the total possible (the Bell exponential number of possible partitions for a set of size m). For two treatments, there are only two possible partitions, both treatments equal or each different. The number of possible partitions grows exponentially with m , but the experiment may limit the number of interesting partitions to a reasonable number. Each vector of observations for a particular gene, $y_{i\bullet\bullet}$, is assumed to come from a mixture of distributions where each component holds the means of the elements of its partition equal. Component $g = 1$ is taken to be the null model where all m treatments are in the same partition, i.e. $t(g) = 1$ group. That is, $f(y_{i\bullet\bullet}) = \sum_{g=1}^G w_g f_g(y_{i\bullet\bullet})$. When the $f_g(\cdot)$ are log-normal, $\theta_{ig} = (\mu_{ig}, \sigma^2)$, where $\mu_{ig} = (\mu_{ig1}, \dots, \mu_{igm})$ has particular elements equal, depending on g . A $N(\mu_0, \tau^2)$ conjugate prior is used for μ_{ig} . When the $f_g(\cdot)$ are gamma, $\theta_{ig} = (a, \lambda_{ig})$, where $\lambda_{ig} = (\lambda_{ig1}, \dots, \lambda_{igm})$ has particular elements equal, depending on g . A $\Gamma(a_0, \nu)$ conjugate prior is used for λ_{ig} . The three unknown parameters (in the log-normal-normal model $(\sigma^2, \tau^2, \mu_0)$ and in the gamma-gamma model (a, a_0, ν)) and the mixing proportions, $(w_1, \dots, w_G)$, are estimated by the marginal maximum likelihood method. The posterior probability of membership in each component can be calculated. They assign a gene to the component in which it has probability greater than 0.5 of membership. Genes assigned to component $g = 1$ are not differentially expressed and those assigned to other components are differentially expressed.

Smyth (2004) uses a mixture of only two components, the first component has no differential expression ($H_0 : \mu_{i1} = \dots = \mu_{im}$). For arrays with two samples (like cDNA microarrays) he analyzes $\log_2(\frac{R_i}{G_i})$, where (R_i, G_i) is the pair of expression measurements from one array and for single channel data he uses log-transformed values. The difference between the two types of data is in the form of the regression model X matrix; in both cases a linear model with coefficient vector α_i is proposed. To allow for questions of biological interest, regardless of the study design that determines the interpretation

of α_i , the focus is on contrasts of the coefficients, $\beta_i = C^T \alpha_i$. In effect, this is a transformation of the data from $x_{i\bullet\bullet} = \log(y_{i\bullet\bullet})$ to $(\hat{\beta}_i, s_i^2)$, the fitted contrasts and the residual variance. The linear model is not necessarily fit with least-squares and the residuals are not necessarily assumed to have a normal distribution. The linear model assumes that $E(x_{i\bullet\bullet}) = X\alpha_i$ and $\text{var}(x_{i\bullet\bullet}) = W_i\sigma_i^2$, where W_i is a known weight matrix. Because of the relative nature of gene expression estimates, these contrasts will generally be of the form of the average difference between treatments; leaving $\beta_i \neq 0$ to imply differential expression. The likelihood on $\hat{\beta}_i$ is $MVN_c(\beta_i, C^T V_i C \sigma_i^2)$ and on s_i^2 is scaled- $\chi^2(d_i, \sigma_i^2)$, where V_i is a matrix of known constants, d_i is the residual degrees of freedom. The prior for σ_i^2 is a scaled-inverse- $\chi^2(d_0, s_0^2)$. However, the prior for $\beta_i = (\beta_{i1}, \dots, \beta_{ic})$ is a mixture for genes that are differentially expressed or not, where c is the number of contrasts. For the j^{th} contrast, $\Pr(\beta_{ij} = 0) = 1 - p_j$ and $\beta_{ij} | \beta_{ij} \neq 0 \sim N(0, v_{0j}\sigma_i^2)$. The hyperparameters $(d_0, s_0, v_{01}, \dots, v_{0c})$ can be estimated with empirical Bayes methods. The posterior mean of σ_j is $\tilde{s}_i = \frac{d_0 s_0^2 + d_j s_i^2}{d_0 + d_j}$. This is used to create the moderated t -statistic, $\tilde{t}_{ij} = \frac{\hat{\beta}_{ij}}{\tilde{s}_i \sqrt{v_{ij}}}$, where v_{ij} is the j^{th} diagonal element of $C^T V_i C$. When $\beta_{ij} = 0$, $\tilde{t}_{ij}$ has a t distribution with $d_0 + d_g$ degrees of freedom. These statistics are used to test for genes with differential expression (non-zero β_i).

Tai and Speed (2004) expand the two treatment case shown in Lönnstedt and Speed (2001). Assuming a multivariate normal distribution for the log transformed expression data, they also use a mixture of two components, the first component has no differential expression ($H_0 : \mu_{i1} = \dots = \mu_{im}$). They use a variable I_i to distinguish differentially expressed genes; $I_i = 0$ if $\mu_{i1} = \dots = \mu_{im}$ and $I_i = 1$ otherwise. Therefore, $x_{i\bullet\bullet} \sim \prod_{k=1}^m MVN_m(x_{i\bullet k}; \mu_i, \Sigma_i)$ and $\mu_i = \mu_0 \mathbf{1}$ if $I_i = 0$ and not otherwise, where MVN_m is the multivariate normal distribution with m dimensions and $\mathbf{1}$ is a $m \times 1$ vector of ones. A simplifying assumption about Σ is made and conjugate priors are added to the model (MVN for μ_i and inverse-Wishart for Σ_i). Empirical Bayes methods (calculating the

marginal maximum likelihood estimates) are used to estimate the hyperparameters of the prior distributions. From this model, the posterior odds of differential expression, $\frac{\Pr(I_i=1|x_{i\bullet\bullet})}{\Pr(I_i=0|x_{i\bullet\bullet})}$, can be computed. They report the *MB* statistic which is the log base 10 odds. This statistic ranks genes in order of their evidence of differential expression (higher values show more evidence).

1.3.3 Nonparametric Models

Efron et al. (2001) propose a nonparametric model for gene expression. Restricting their attention to the two treatment case and assuming that there are paired samples of the two treatments, they examine the average differences of log transformed gene expression measurements after standardization, called z_i . They propose that observed differences come from a mixture of the null distribution, $f_0(\cdot)$, with probability p_0 and the distribution of differentially expressed genes, $f_1(\cdot)$, with probability $p_1 = 1 - p_0$. The null distribution is estimated using differences between pairs of samples in the same treatment. The posterior probability of differential expression for gene i is then $p_1(z_i) = 1 - \frac{p_0 f_0(z_i)}{p_0 f_0(z_i) + p_1 f_1(z_i)}$. The ratio $\frac{f_0(z_i)}{p_0 f_0(z_i) + p_1 f_1(z_i)}$ can be estimated by comparing the empirical distributions of the real and null data and the upper bound for $p_0 = \min_i \frac{p_0 f_0(z_i) + p_1 f_1(z_i)}{f_0(z_i)}$ is used as an the estimate. Genes with $p_1(z_i) > 0.9$ can be classified as differentially expressed.

1.4 Thesis organization

1.4.1 Embryogenic maize tissue

Embryogenesis is a potentially important genetic engineering technique for maize. Though the process does not occur naturally, some lines of maize are more embryogenic than others. Therefore, we are interested in the differences in gene expression in embry-

onic cells over embryo development to identify genes that have an important role in the process.

We compared embryogenic cells from three two-line pools of maize plants of the HiII hybrid line at 7, 14, 21, 24, and 28 days after culling of embryogenic tissue from callus and transfer to an embryogenic medium. We used cDNA microarrays spotted on glass slides printed with 12160 genes. These were hybridized with two samples each from the mRNA batches. Image analysis and normalization were performed, resulting in corrected intensities for analysis.

The paper in Chapter 2 describes the experiment in which these data were collected. We carried out a standard statistical analysis to identify genes with differential expression over the five time points. For the 2,000 genes most likely to be differentially expressed, a standard clustering algorithm was used to group them into 15 clusters of genes and the characteristics of these clusters are discussed.

1.4.2 Bayesian model for combining scans

Generally, different laser and sensor settings can be used to read a cDNA microarray slide. Stronger laser settings create more fluorescence and stronger sensor settings pick up more signal. There is a balance to be struck between picking up signal from the lowly fluorescing spots and over-exposing the highly expressing genes. For our cDNA scanners, there is an upper limit to fluorescence intensity values of 65535; readings of spots which are brighter are censored. Over-exposing the high intensity spots will cause them to be artificially near other high expression values. Correspondingly, low signals will be artificially assigned to 0 if the laser and sensor settings are too low.

In Chapter 3, we incorporate three separate scans of a slide into one estimate of gene expression per gene. We assume that if the replicate scans were properly scaled, then

for each gene the scans would be independent draws from a common distribution. We propose a model relating the scans and the genes to each other to share information between genes. Using Markov chain Monte Carlo (MCMC) methods, we can generate draws from the joint posterior distribution of the true mean gene expression.

The posterior mean expression value for each gene is used as the estimated expression. These values are then normalized using the procedure described earlier and the resulting expression estimates are taken as the observed values of gene expression for each experiment. We show that by using these estimates of expression instead of the values from a single scan, we identify more differentially expressed genes than in Chapter 2.

1.4.3 Clustering posterior distributions

We propose approaches for clustering the 12,160 genes in the maize embryogenesis experiment into groups with similar expression patterns. Using a hierarchical model for gene expression, we can identify the set of non-differentially expressed genes. We assume that all these genes can be clustered together into a class of genes with the null (constant) expression pattern. We then wish to group differentially expressed genes according to the pattern of differential expression over time.

In Chapter 4 we compare several methods for clustering the differentially expressed genes. Since the characteristic of interest for these genes is expression change over time, the variable we use in clustering is the expression ratios for sequential time pairs. In all methods, we first cluster those genes with high probability of constant expression into one group. This reduces the computational burden by reducing the number of genes to be clustered with more sophisticated techniques. In Chapter 2, we clustered the ratios of observed expression means for the significantly differentially expressed genes. In Chapter 4, probability of differential expression given the observed data was determined

using a hierarchical model for expression measurements. This also generated posterior distributions for the expression ratios of each gene. The standard method uses the posterior means of the expression ratios over time to cluster genes, similar the the method in Chapter 2. The three other methods use the full posterior distributions to determine the difference between genes for clustering purposes.

1.5 References

Affymetrix Inc. (2002). *Statistical Algorithms Description Document*

Baldi, P. and Long, A.D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics* 17(6) 509–519

Broët, P., Richardson, S., and Radvanyi, F. (2002). Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. *Journal of Computational Biology* 9(4) 671–683

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B* 57, 289–300

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*

Dobbin, K., Shih, J.H., and Simon, R. (2003). Statistical design of reverse dye microarrays. *Bioinformatics* 19(7), 803–810

Efron, B., Tibshirani, R., Storey, J.D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* 96(456) 1151–1160

- Ibrahim, J.G., Chen, M., and Gray, R.J. (2002). Bayesian models for gene expression with DNA microarray data. *Journal of the American Statistical Association* 97(457) 88–99
- Kendzierski, C.M., Newton, M.A., Lan, H., and Gould, M.N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* 22 3899–3914
- Kerr, M.K., and Churchill, G.A. (2001). Experimental Design for Gene Expression Microarrays. *Biostatistics* 2, 183–201
- Lönnstedt, I. and Speed, T.P. (2001). Replicated microarray data. *Statistica Sinica* 12 31–46
- Richmond, C.S., Glasner, J.D., Mau, R., Jin, H., and Blattner, F.R. (1999). Genome-wide expression profiling in *Escherichia coli* K-12 *Nucleic Acids Research*, 27(19), 3821–3835.
- Skibbe, D.S., Nettleton, D., and Schnable, P.S. (2004). Scanning microarrays at multiple intensities increases the number of statistically significant differences detected. *Unpublished Manuscript*
- Smyth, G.K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* 3(1) Article 3
- Smyth, G.K., Yang, Y.H., and Speed, T. (2002). Statistical Issues in cDNA Microarray Data Analysis. *Functional Genomics: Methods and Protocols* Humana Press, Totowa, NJ.

Tai, Y.C. and Speed, T.P. (2004). A multivariate empirical Bayes statistic for replicated microarray time course data. *Technical Report 667*

Wit, E. and McClure, J. (2004). *Statistics for Microarrays : design, analysis, and inference*, John Wiley & Sons, West Sussex, England.

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

CHAPTER 2. GENE EXPRESSION PATTERNS DURING SOMATIC EMBRYO DEVELOPMENT AND GERMINATION IN MAIZE Hi II CALLUS CULTURES

A paper submitted to *Plant Molecular Biology*

Ping Che¹, Tanzy M. Love^{2,3}, Bronwyn R. Frame⁴, Kan Wang⁴,
Alicia L. Carriquiry² and Stephen H. Howell^{1,5}

2.1 Abstract

Gene expression patterns were profiled during somatic embryogenesis in a regeneration-proficient maize hybrid line, Hi II, in an effort to identify genes that might be used as developmental markers or targets to optimize regeneration steps for recovering maize plants from tissue culture. Gene expression profiles were generated from embryogenic calli induced to undergo embryo maturation and germination. Over 1,000 genes in the 12,060 element arrays showed significant time variation during somatic embryo development. A substantial number of genes were downregulated during embryo maturation, largely histone and ribosomal protein genes, which may result from a slowdown in cell proliferation and growth during embryo maturation. The expres-

¹Plant Sciences Institute, Iowa State University

²Department of Statistics, Iowa State University

³Responsible for the statistical analysis and writing in this paper

⁴Center for Plant Transformation and the Department of Agronomy, Iowa State University

⁵Author for correspondence

sion of these genes dramatically recovered at germination. Other genes up-regulated during embryo maturation included genes encoding hydrolytic enzymes (nucleases, glucosidases and proteases) and a few storage genes (an α -zein and caleosin), which are good candidates for developmental marker genes. Germination is accompanied by the up-regulation of a number of stress response and membrane transporter genes, and, as expected, greening is associated with the up-regulation of many genes encoding photosynthetic and chloroplast components. Thus, some, but not all genes typically associated with zygotic embryogenesis are significantly up or down-regulated during somatic embryogenesis in Hi II maize line regeneration. Although many genes varied in expression throughout somatic embryo development in this study, no statistically significant gene expression changes were detected between total embryogenic callus and callus enriched for transition stage somatic embryos.

2.2 Introduction

The regeneration of maize in tissue culture is important for the production of transgenic maize and for crop improvement using genetic engineering approaches. The first somatic embryos in maize tissue culture were produced by Green and Phillips (1975). Reports of fertile maize plants regenerated from protoplasts (Prioli and Sondahl, 1989, Shillito et al., 1989) were closely followed by the production of transgenic, fertile maize from transformed suspension cell cultures of the hybrid A188 x B73 line (Gordon-Kamm et al., 1990).

Maize cell lines derived from transformation competent sources such as immature embryos are heterogeneous for cells with differing embryogenic potential. Friable (Type II) callus (Armstrong and Green, 1985) was found to be highly embryogenic and readily produced plants. Induction of embryogenic callus is genotype-specific in many plant

species, including maize. Most maize elite lines remain inaccessible to improvement using standard transformation techniques either because they fail to produce embryogenic callus from transformation competent tissues, or they fail to regenerate efficiently after embryogenic callus induction.

Some of the early attempts to find indicators for embryogenic competence relied on biochemical markers. Isozyme differences between embryogenic and non-embryogenic cultures were demonstrated for glutamate dehydrogenase, isoperoxidase, esterase and malate dehydrogenase isozymes (Fransz et al., 1989, Rao et al., 1990). Schmidt et al. (1997) employed differential display to identify genes specifically expressed in embryogenic carrot cells. One such gene encoded a leucine-repeat receptor protein kinase and was dubbed as a somatic embryogenesis receptor kinase (SERK) (Schmidt et al., 1997). In Arabidopsis, five members of the SERK family have been identified (AtSERK1-5). AtSERK1 was expressed during somatic embryogenesis, and the embryogenic competence of callus derived from seedlings over-expressing AtSERK1 (driven by the CaMV35S promoter) was elevated 3-4 fold when compared with the wild-type callus (Hecht et al., 2001). At least two related genes have been identified in maize, ZmSERK1 and 2 (Baudino et al., 2001). ZmSERK1 was preferentially expressed in reproductive tissues with the strongest expression in microspores, while ZmSERK2 expression was fairly uniform in all tissues investigated. Both genes were expressed in callus cultures whether they were embryogenic or not, which suggested that the genes might not be good markers for embryogenesis in maize (Baudino et al., 2001).

LEAFY COTYLEDON 1 (LEC1) in Arabidopsis is up-regulated during zygotic embryogenesis and promoted somatic embryogenesis when ectopically expressed in vegetative cells (Lotan et al., 1998). *LEC1* encodes a transcription factor, and *lec1* mutants prematurely germinate producing cotyledons with characteristics of later postgermina-

tive development (Meinke, 1992, Meinke et al., 1994, West et al., 1994). *lec1* affects the expression of certain maturation phase genes including those encoding storage proteins (Meinke et al., 1994, West et al., 1994, Parcy et al., 1997, Vicient et al., 2000). Maize genes with sequences similar to *LEC1* have been identified, and the expression pattern of ZmLec1 has been profiled during somatic embryogenesis (Zhang et al., 2002). The expression of ZmLec1 during maize somatic embryogenesis was similar to *LEC1* during Arabidopsis zygotic embryogenesis with general expression throughout the embryo up to the globular stage of development (Zhang et al., 2002). Lowe et al. (2000) reported that ectopic expression of the ZmLec1 greatly improved the recovery of transformants in maize tissue culture.

In this study, we profiled gene expression patterns during somatic embryo maturation and germination in a regeneration-proficient maize line, Hi II. We found significant gene expression changes during somatic embryo maturation after removal from auxin-containing medium. However, no significant changes in gene expression were evident when comparing embryogenic callus enriched with transition stage somatic embryos and total callus on auxin-containing medium. The genes regulated during these later stages of somatic embryogenesis may serve as developmental markers or for genetically improving the regeneration of more recalcitrant lines.

2.3 Materials and methods

2.3.1 Materials and tissue culture methods

Somatic embryos were generated in embryogenic callus lines developed independently from immature Hi II zygotic embryo explants using protocols described at the Plant Transformation Facility website for the production of transgenic corn (<http://www.agron.iastate.edu/ptf/web/system.htm>).

Briefly, greenhouse-grown ears from the Hi-II hybrid line (Armstrong et al., 1991) were dehusked and surface sterilized for 20 min (50% commercial bleach in water plus 1 drop/L of Tween 20) then rinsed three times with sterilized water. Immature zygotic embryos were excised and cultured embryo-axis side down (scutellum side up) on N6E media (N6 salts and vitamins (Chu, 1975), 2 mg/L 2,4-D, 100 mg/L myo-inositol, 2.76 g/L proline, 30 g/L sucrose, 100 mg/L casein hydrolysate, 2.5 g/L gelrite, pH 5.8 after Songstad et al., (1996). Silver nitrate (25 μ M) was added after autoclaving. The plates were wrapped with vent tape and incubated at 28°C in the dark for 2 weeks.

Friable Type II callus was bulked up from 6 separate embryo explants over 8 weeks by sub-culturing every two weeks on the same medium. Callus was then subjected to regeneration conditions by transferring about 15 small pieces (approximately 4 mm) of embryoid-enriched embryogenic callus to Regeneration Medium I (MS salts and vitamins (Murashige and Skoog, 1962), 100 mg/L myo-inositol, 60 g/L sucrose, 3 g/L gelrite, pH 5.8) and incubating for 3 weeks at 25°C in the dark (McCain and Hodges, 1986). Petri-plates (100x25 mm) were wrapped with vent tape. After 3 weeks, matured somatic embryos were identified using a light microscope, transferred to Regeneration Medium II (as for Regeneration Medium I but with 3% sucrose), and placed in the light (80 μ E/m²/s) for germination. Plantlets sprouted leaves and roots on this medium.

2.3.2 RNA extraction and microarray analysis

RNA was extracted using a TRIzol method modified from Chomzynski and Sacchi (1987) and described in TAIR protocols (<http://www.arabidopsis.org/servlets/TairObject?type=protocol&id=501683718>). In this procedure 1 g of maize callus tissue was ground with liquid nitrogen in a mortar and pestle. The ground powder was mixed with 15 ml TRIzol reagent (Life Technologies) and incubated at 60° for 5 min. The mixture was centrifuged at 12,000 g at 4° for 10 min and to the supernatant was

added 3 ml of chloroform. The mixture was vortexed for 15 sec and allowed to sit at room temperature for 2-3 min. The mixture was centrifuged at 10,000 g at 4° for 15 min, and RNA was precipitated from the upper phase by adding 1/2 volume each of isopropanol and 0.8M sodium citrate/1.2M NaCl. The mixture was allowed to sit at room temperature for 10 min and centrifuged at 10,000g at 4° for 10 min. The pellet was washed with 70% EtOH, vortexed briefly and centrifuged again at 10,000g at 4° for 10 min. The pellet was air dried for 5 min and dissolved in 250 μ l of DEPC-treated water. The RNA sample was centrifuged in a microcentrifuge for 5 min at room temperature and the insoluble pellet discarded. The RNA sample was cleaned up by passing through an RNeasy column (Qiagen) according to manufacturers instructions.

cDNA was synthesized and labeled according to procedures described by Hegde et al. (2000). The procedure is an indirect labeling method in which first-strand cDNA is synthesized in the presence of amino-allyl labeled dUTP, and then NHS-esters of the appropriate cyanine fluor are covalently coupled to the substituted cDNA strand. The reaction mix for first strand synthesis consisted of Superscript II buffer (Life Technologies), 10 mM DTT, 5 mM dATP, dCTP and dGTP, 3 mM dTTP, 2mM aminoallyl-dUTP, 0.3 mg/ml oligo dT (Invitrogen) and 400 units of Superscript II reverse transcriptase (Invitrogen). The reaction was incubated overnight at 42° followed by base hydrolysis of RNA in 200 mM NaOH, 10 mM EDTA and incubation for 15 min at 65°.

The aminoallyl-label cDNA was purified using a modified QIAquick (Qiagen) PCR purification procedure. The cDNA reaction was mixed with 5X volume of 5 mM potassium phosphate (PB, pH 8.0) and transferred to a QIAquick column. The column was centrifuged for 1 min in a collection tube at 14,000 rpm in a microcentrifuge, washed twice with 750 μ l of 5 mM PB (pH 8.0) and 80% EtOH and centrifuged each time. 30 μ l of 4 mM potassium phosphate (pH 8.5) were added to the column, incubated for 1 min,

and RNA was eluted by centrifugation at 14,000 rpm for 1 min. The elution step was repeated once more with another 30 μ l of 4 mM PB (pH 8.5). The sample was dried in a SpeedVac.

The aminoallyl-label cDNA was coupled to the Cy dyes by dissolving the dried cDNA in 4.5 μ l of freshly prepared 0.1 M sodium carbonate buffer (pH 9.0). Cy3- or Cy5-esters (AmershamPharmacia) were dissolved in 73 μ l DMSO, and 4.5 μ l of the appropriate NHS-Cy were added to the labeled cDNA. The mixture was incubated in the dark at room temperature for 1 hr. Following the reaction, uncoupled dye was removed using a QIAquick PCR purification kit (Qiagen). 35 μ l of sodium acetate buffer (pH 5.2) and 250 μ l 5 mM PB (pH 8.0) were added to the reaction and transferred to a QIAquick column. The dye-coupled cDNA was eluted with 2 aliquots of 30 μ l of elution buffer (Qiagen) and dried in a SpeedVac.

Maize cDNA chips were prepared in the Iowa State University microarray facility by spotting aminosilane coated slides with a Cartesian PixSys 5500 Arrayer. The maize chips contained over 12,000 spotted cDNA inserts obtained from the NSF Plant Genome EST projects led by Virginia Walbot (Stanford) and Patrick Schnable (Iowa State). The cDNAs included in the chip (Gen II, Version B) are listed at <http://www.plantgenomics.iastate.edu/maizechip/>. The slides to be hybridized were placed in Coplin jar with prehybridization buffer (5XSSC, 0.1% SDS and 1% bovine serum albumin) and incubated at 42° for 45 min. The slides were washed 5X by dipping in MilliQ water (Millipore) at room temperature, followed by dipping in isopropanol and air drying.

For hybridization, each labeled probe was resuspended in 19 μ l of hybridization buffer (50% formamide, 5X SSC and 0.1% SDS) to which was added 1 μ l of 20 μ g / μ l human COT1 DNA (LifeTechnologies) and 1 μ l of 20 μ g / μ l poly A DNA (Invitrogen) to block

non-specific hybridization. The sample was heated at 95° for 3 min to denature the probe and centrifuged at 13,000 rpm for 1 min in a microcentrifuge at room temperature. The probe was applied to a microarray slide, covered with a 22 X 60 mm glass coverslip and placed in a sealed hybridization chamber with 20 μ l of water added to the chamber at the end of the slide. The chamber was incubated overnight at 42°. Following incubation the slide was carefully removed from the chamber and placed in a staining dish with wash buffer containing 1X SSC and 0.2% SDS at 42°. The coverslip was gently removed, and the slide was agitated in the wash buffer for 4 min. The slide was further washed with 0.1X SSC and 0.2% SDS at room temperature for 4 min and then in 0.1X SSC for another 4 min. The slides were allowed to air dry.

2.3.3 Microarray data analysis

Imagene software (Biodiscovery) was used to read image files from the General Scanning ScanArray 5000 scanner. Imagene employs a fixed circle method to segment spots by positioning a circle of fixed diameter for the greatest difference between pixels inside and outside the spot. The mean signal pixel intensity computed from approximately 120 pixel intensity measurements was obtained for each spot. Background was selected using a concentric-circle-band method in which a second circle is placed around the first and pixels within the halo are designated as background. The intensity of each background pixel was recorded, and the median background pixel intensity was used to estimate the background effect. (The median was used rather than the mean because some pixels designated as background may actually have fluorescent probe in them. These pixels, therefore, have much higher intensity values than the pixels from empty regions of the slide.)

In the time course study, all of the slides from each line pool were prepared in order and read in the same batch. This was done for job scheduling reasons and is

not recommended for an experimental setup. It would have been better to randomize the slides with respect to experimental order, because time effects (learning, machine calibration) may be present and confound with treatment effects. Substantially more effort to randomize preparation between line pools would allow more precise estimates of the line variation.

Different laser and sensor settings were used to scan each slide to adjust the dynamic range of the scanner to the overall fluorescence intensity of the slide. Higher laser settings create more fluorescence and higher photomultiplier settings amplify the light signal. However, low range settings miss spots with low signals and in high range settings, high intensity signals are saturated. (There is an upper limit of 65535 to the measurement of fluorescence so that signal from spots that are brighter will be censored.) Preliminary work indicates that a significant reduction in the variability of expression estimates can be obtained when analyzing the data from multiple readings with the appropriate statistical model. However, for speed and simplicity, we included only one reading for each slide by choosing the one with the highest median intensity among readings with the fewest intensities reported as 65535 (the maximum).

2.3.4 Data normalization

We assume that there is a systematic bias in the gene expression measurements between the two dyes. For gene j which is not differentially expressed, we do not expect $R_j = G_j$ on average, where (R_j, G_j) are the expression estimates of gene j measured on the Cy5 and Cy3 channels of a slide. Instead, we expect $R_j = k_j G_j$ for some k_j . The total signal intensity for each gene on a single slide is the sum of the fluorescent intensity in both R and G channels. The dye bias has been shown to be dependent on the intensity level (Yang et al., 2002). An alternative measure of intensity defined as $A_j = \log(\sqrt{R_j G_j})$ can be plotted against the log ratio, $M_j = \log(R_j/G_j)$, and this

shows the intensity dependence more clearly because both measures are defined on the log scale. Additionally, each print tip (32 on these slides) has characteristics which can result in spots printed by the same print tip to be correlated. As a consequence, spots in the same print tip group (metarow and metacolumn combination) appear in spatially similar groups within the slide. Thus print tip groups may account for bias due to print tips and act as a surrogate for spatial effects (Yang et al., 2002). The effects of intensity dependent, print tip group related dye bias should be removed in normalization.

Print tip group-intensity dependent normalization assumes that the normalizing constant is a function of intensity for each print tip group i , $k_j = f_i(R_j + G_j)$. We assume that only a small proportion of genes in our experiment are differentially expressed and use a robust estimator of $\log(k_j)$, the loess curve of M against A using only the middle range of the data in each print tip group (Yang et al., 2002). Print tip group-intensity dependent normalization has the following characteristics:

- I functions of intensity per slide where I is the number of print tip groups; each gene takes its own value within a group.
- The factor k_j is interpreted as the dye bias against Cy5 at intensity $R_j + G_j$.
- Accounts for intensity dependent effects.
- Includes some spatial effects.
- Does not rescale the data to have a similar variability on different slides.

We used print tip group-intensity dependent normalization to remove the systematic bias related to dye and print tip group. We fitted the loess curves of intensity for each print tip group and corrected each pair of expression values on a slide for the curve. These background corrected and normalized values were the ones used in our analysis.

2.3.5 Estimation of treatment effects

Analysis of variance (ANOVA) was used to determine whether the gene expressions from different groups (or treatments) have equal mean expression. Under the null hypothesis, all groups have a common mean and standard deviation. ANOVA is used to test whether any of the groups violate the assumption of common means. In the design of the time course experiment, each group corresponds to a different time point in embryonic development, and there were 12 observations for each gene at each time point. Therefore, we can, in principle, conduct a test of the null hypothesis for each of the genes, to investigate whether mean expression varies across treatments (or time points).

Because there are a large number of elements (12,060) in the arrays, conducting so many hypothesis tests would likely result in a large proportion of false positive conclusions. A false positive occurs when we erroneously conclude that a gene exhibits different expression levels at different time points. In experiments such as this, it is very important to control the experiment-wise error rate at a predetermined level by carrying out an adjustment that accounts for the erosion in confidence levels in multiple comparisons. We do so using the p-values generated by the ANOVA test, P_j , for each gene.

We use an adjustment proposed by Benjamini and Hochberg (1995). This adjustment attempts to control the expected proportion of false positives out of genes concluded to be differentially expressed. This proportion is called the false discovery rate. The multiple comparisons adjustment used assumes that the test statistics generating the p-values are independent. The j th gene is considered significantly differentially expressed over time if $P_j \leq P_{(k)}$ where $P_{(k)}$ is the k th ordered p-value for the genes, $k = \max\{j : P_{(j)} \leq j*\alpha/t\}$, t is the number of tests being performed, and $\alpha > 0$ is the predetermined target error rate. Using this rule, the expected false discovery rate will be less than α .

2.4 Results

2.4.1 Somatic embryogenesis and expression profiling

Callus derived from maize Hi II immature zygotic embryos can be propagated *in vitro* as Type II callus (Armstrong et al., 1991). During the growth of this callus on auxin (2,4-D)-containing medium, some of the callus cells form embryogenic cell clusters, which eventually differentiate into globular and transitional stage somatic embryos (Jimenez 2001), or so-called embryoids (globular-like embryos with conspicuous suspensor-like structures, Fig. 1A) (Armstrong and Green, 1985). For routine maize regeneration, highly embryogenic callus, rich in its content of embryoids, is transferred onto Regeneration Medium I (no 2,4-D, 6% sucrose) to induce somatic embryo maturation (Fig. 1B). After 7 days on this medium, tissue destined to form mature somatic embryos appears milky or less translucent. Embryo development and maturation continues for 21 days, and when mature somatic embryos are transferred to light on Regeneration Medium II (no 2,4-D, 3% sucrose), the embryos germinate.

Two independent experiments were conducted to examine gene expression patterns during somatic embryogenesis in maize (Fig. 2A and B). The goal of the first experiment (Fig. 2A) was to profile gene expression patterns during somatic embryo maturation and germination with the aim of understanding the gene expression events underlying somatic embryogenesis and possibly identifying developmental markers. The second (Fig. 2B) was designed to determine whether gene expression differences could be detected between embryogenic callus enriched with embryoids and total embryogenic callus growing on 2,4-D-containing (N6E) medium.

Six independent, embryogenic callus lines (A-F) were sampled, and two lines were pooled (creating 3 line pools) to obtain sufficient amounts of RNA for microarray analy-

sis (without amplification). Gene expression patterns were profiled using maize cDNA microarrays. Thirty-six microarray chips were each spotted with 12,060 maize cDNAs. Thirty were used for the time course analysis following induction of somatic embryo maturation (Fig. 2A) and the remaining six arrays were used to compare embryoid-enriched and total callus prior to removal from auxin-containing medium (Fig. 2B). The chips were hybridized with cy3 and cy5 cDNAs using a loop design strategy (Dobbin and Simon, 2002) in which samples were compared to each other and not to a single reference, such as a zero time sample. In the time course experiment, the strategy allows for more repetition of time points with the same number of chips. In each line pool, each time point is sampled 4 times – twice with a cy3 labeled probe and twice with a cy5 labeled probe. Thus, across all three line pools, a time point sample is repeated 12 times.

Such a scheme permits analysis of both time and line pool variation. However, it should be noted that of the 12 repeated measurements on each time point, only three are true biological replications (3 line pools). Thus, the power of our conclusions is lessened by the fact that the four replications at each time point within a line pool are technical replications using the same biological material.

2.4.2 Gene expression patterns following induction of somatic embryo maturation

Following induction of embryo maturation, somewhat more than a 1000 genes out of 12,060 in the study showed significant time variation (at the $\alpha=0.05$ level, considering multiple comparisons, see supplementary Table I). During maturation and germination, increasing numbers of genes were up-regulated by 2-fold or more (Fig. 3). Likewise, an increasing number of genes were down regulated 2-fold or more during maturation, but that trend reversed itself during embryo germination.

The overall trends were made up of individual genes with varied expression patterns, and patterns for ~1000 genes with significant time variation were organized into groups. The patterns were clustered into 12 groups with model-based clustering of the sequential expression ratios. Each gene was assigned to one of 19 functional categories. The largest category was for genes with unknown function and usually the second largest was for genes involved in primary metabolism. The functional distribution for four groups of genes in which a category, other than unknown or primary metabolism, emerged or dominated the distributions are shown.

The first pattern group was characterized by genes down-regulated during embryo maturation, which then recovered during germination (Fig 4A). Compared to other groups, this group had a larger number of genes encoding nuclear proteins, such as a gene encoding proliferating cell nuclear antigen, and histone genes. The decline in expression of these genes during maturation likely indicates a reduction in cell proliferation and growth during maturation. The recovery in expression of these genes later on accompanied the growth spurt during germination. In another group of genes, dominated by a category of glucosidases, nucleases and proteases, expression rose during maturation but then dropped off during germination (Fig. 4B).

Early in germination, a group of stress-related genes, such as a gene encoding a heat shock protein, were transiently up-regulated (Fig 4C). Other genes that were up-regulated during germination (with expression levels generally higher at both time points) included a large group encoding channel proteins or membrane transporters, such as a gene encoding a water channel (data not shown). Finally, genes encoding photosynthetic and other chloroplast components, such as a gene encoding a chlorophyll a/b binding protein, were up-regulated as shoots began to green (Fig. 4D). Thus, from a gene expression perspective, germination first involved the activation of expression of

stress-related and transporter/channel-encoding genes followed by the up-regulation of photosynthetic/chloroplast genes.

We also examined the expression patterns of genes typically associated with zygotic embryogenesis and germination to determine if they would be good markers for somatic embryogenesis. Some showed expected expression patterns – others did not. Of those that did, a gene encoding an α -zein, an embryo storage protein, was up-regulated as expected during embryo maturation, and then its expression levels fell during germination (Fig. 5A). Another gene encoding an embryo-specific Ca^{++} -binding protein (ATS1, caleosin) showed a somewhat similar pattern of expression (Fig. 5B). A gene encoding a late embryogenesis abundant protein was expressed at increasing levels during maturation, but transcript levels continued to rise during germination (Fig. 5C). Finally, a gene encoding a protein related to germans was up-regulated, as expected, during germination (Fig. 5D). Surprisingly, most other genes encoding zeins and late embryogenesis abundant proteins did not show significant time variations in expression during embryo maturation and germination.

2.4.3 Line variation

We also looked for genes with significant variation across time points and either with considerable or little variation across lines. Genes with expression patterns that vary considerably across lines might be useful if the variation correlates with regeneration competence or with other traits that vary from line to line. For example, two genes, one encoding a lipid transfer protein and the other a bZIP family transcription factor, showed significant time variation in one line pool, but not in the other two line pools (Fig 6, upper panel). Such genes might be useful developmental markers for distinguishing cultures that show line variation in later regeneration steps. Other genes, such as one encoding a putative disease response protein and a photosystem I assembly protein,

showed significant variation across time points but little variation between lines (Fig. 6, lower panel).

2.4.4 Comparison of embryoid enriched and total callus

In our maize regeneration procedure (<http://www.agron.iastate.edu/ptf/Web/mainframe.htm>), embryoid-enriched callus is selected (using a dissecting scope) from friable embryogenic callus for transfer onto Regeneration Medium I. This regeneration approach was reported to produce over 30 times more plants per gram fresh weight of callus than did the indiscriminate transfer of total embryogenic callus to the same medium (McCain and Hodges, 1986). However, when we compared gene expression levels in embryoid-enriched callus to those in total callus, we found that none of the 12060 genes in this study showed significant differences (at the $\alpha=0.05$ level, considering multiple comparisons, see supplementary Table II).

2.5 Discussion

Gene expression patterns change extensively during somatic embryo maturation and germination. Following transfer to medium lacking 2,4-D and throughout embryo maturation, there is a progressive decline in the expression of genes involved in cell proliferation and growth, such as genes encoding histones and ribosomal proteins (Fig. 7). Strikingly, the expression levels of these genes recover at the onset of germination. The changes in expression may reflect a slowdown in cell proliferation and growth during somatic embryo maturation and a resurgence in expression of these genes at germination. During maturation, expression rises for a group of genes encoding hydrolytic enzymes, such as nucleases, glucosidases and proteases, suggesting a breakdown, and perhaps a retooling, of cell components during this stage of somatic embryo development. Unlike zygotic embryogenesis (Lending and Larkins, 1989), we did not observe the large-scale

up-regulation in expression of storage protein genes. Whether maize Hi II somatic embryos accumulate fewer storage proteins than their zygotic counterparts is a matter that deserves exploration. In any case, only a few storage protein genes for α -zein and a caleosin, a lipid body protein (Naested et al., 2000), appear to be good markers for somatic embryo maturation in these Hi II lines.

Some stress response genes, such as heat shock genes are up-regulated at the onset of germination (Fig. 7). Their up-regulation may be a normal developmental event or a response to the transfer of tissue to new culture medium. Of interest is the up-regulation of a group of genes that encodes various transporters and membrane channels. Finally, as expected, germination and shoot greening are accompanied by the activation in expression of a myriad of genes encoding photosynthetic and chloroplast components.

Some of the gene expression patterns we observed were significantly line-pool dependent and others were not. Highly regulated genes with expression patterns that are line independent may be good developmental markers across multiple lines. Line-dependent genes, on the other hand, may be useful if their expression patterns correlate with traits such as efficient regeneration of fertile plants. We also looked for gene expression differences between embryoid-enriched and total callus and found none that were statistically significant. Two possible reasons for this may be: 1) total callus from the regeneration-proficient Hi II line is replete with viable embryoids but also contains embryogenic cells clusters and globular embryos (without suspensors), both of which also represent early stages of somatic embryogenesis. The developmental difference between embryoid enriched and total callus may therefore be modest. 2) Genes that differ in expression between embryoid enriched and total callus may not be present on the cDNA chips or may not be reflected by differences in the transcriptome. Gene expression differences between embryogenic and pre- or non-embryogenic callus might be more effectively de-

tected in less regeneration-proficient lines.

In early attempts to identify markers for embryogenic competence, translation products of RNA from cultured carrot cells and somatic embryo were compared by 2D gel electrophoresis. With the exception of two polypeptides, called E1 and E2, Sung and Okimoto (1981) found few differences, which led Choi et al. (1987) to suggest that the similarities in gene expression patterns may reflect the fact that pro-embryonic masses (PEMs) in cultured cells may already be "committed to the embryogenic program." Wilde et al. (1988) also used 2D gel electrophoresis of translation products to arrive at similar conclusions.

Gene expression markers have been used more widely in recent years to characterize embryogenic lines and to describe embryo development. Chugh and Khurana (2002) reviewed the state of knowledge on gene expression in somatic embryogenesis in higher plants prior to the extensive use of global gene profiling technologies. A recent microarray study by Thibaud-Nissen et al. (2003) profiled gene expression patterns during somatic embryogenesis in soybean. Soybean somatic embryos are formed on the adaxial surface of immature cotyledons placed on high levels of 2,4-D. Thibaud-Nissen et al. (2003) compared gene expression during embryo development on the adaxial side of cotyledons to callus formation on the abaxial side. Their results suggest that cotyledons dedifferentiate for two weeks prior to the development of somatic embryos. Genes involved in oxidative stress responses and cell division change in expression on the adaxial side of the cotyledons indicating that events involving cell proliferation and cell death are played out during somatic embryo development (Thibaud-Nissen et al., 2003).

Some of the general features of the gene expression program in soybean were also observed in the course of maize somatic embryogenesis such as the increase in expression of certain storage protein genes, the fall and subsequent rise in cell division gene expres-

sion and the mid-course expression of stress response genes. Because we measured gene expression patterns during the late stages of somatic embryo development in this study, we would not expect to observe expression of genes associated with oxidative burst, detoxification and cell wall modification that Thibaud-Nissen et al. (2003) attributed to the earlier, dedifferentiation stage of somatic embryogenesis from soybean cotyledon tissue.

2.6 Acknowledgements

PC is supported by the Plant Sciences Institute, Iowa State University; TML is supported by NSF DMS-0091953. BRF is supported by the Department of Agronomy, the Plant Sciences Institute, and the Biotechnology Program all at Iowa State University.

2.7 References

- Armstrong, C.L., Green, C.E. and Phillips, R.L. 1991. Development and availability of germplasm with high Type II culture formation response. *Maize Gen. Coop. Newsletter* 65: 92-93.
- Baudino, S., Hansen, S., Brettschneider, R., Hecht, V.F., Dresselhaus, T., Lorz, H., Dumas, C. and Rogowsky, P.M. 2001. Molecular characterisation of two novel maize LRR receptor-like kinases, which belong to the SERK gene family. *Planta* 213: 1-10.
- Benjamini, Y. and Hochberg, Y. 1995. Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *J Royal Stat. Soc. Ser. B* 57: 289-300.
- Choi, J.H., Liu, L.S., Borkird, C. and Sung, Z.R. 1987. Cloning of genes developmentally regulated during plant embryogenesis. *Proceedings of the National Academy of Sciences of the United States of America* 84: 1906-1910.

Chomczynski, P. and Sacchi, N. 1987. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem* 162: 156-159.

Chu, C.C. 1975. Establishment of an efficient medium for anther culture of rice through comparative experiments on the nitrogen sources. *Sci. Sinica* 18: 659-668.

Chugh, A. and Khurana, P. 2002. Gene expression during somatic embryogenesis – recent advances. *Curr. Science* 83: 715-730.

Dobbin, K. and Simon, R. 2002. Comparison of microarray designs for class comparison and class discovery. *Bioinformatics* 18: 1438-1445.

Fransz, P.F., de Ruijter, N.C.A. and Schel, J.H.N. 1989. Isozymes as biochemical and cytogenetic markers in embryogenic callus cultures of maize (*Zea mays* L.). *Plant Cell Rep.* 8: 67-70.

Gordon-Kamm, W., Spencer, T., Mangano, M.L., Adams, T., Daines, R., Start, W., O'Brien, J., Chambers, S., Adams Jr., W., Willetts, N., Rice, T., Mackey, C., Krueger, R., Kausch, A. and Lemaux, P.G. 1990. Transformation of maize cells and regeneration of fertile transgenic plants. *Plant Cell* 2: 603-618.

Green, C.E. and Phillips, R.L. 1975. Plant regeneration from tissue cultures of maize. *Crop Science* 15: 417-420.

Hecht, V., Vielle-Calzada, J.P., Hartog, M.V., Schmidt, E.D., Boutilier, K., Grossniklaus, U. and de Vries, S.C. 2001. The Arabidopsis SOMATIC EMBRYOGENESIS RECEPTOR KINASE 1 gene is expressed in developing ovules and embryos and enhances embryogenic competence in culture. *Plant Physiol.* 127: 803-816.

Hegde, P., Qi, R., Abernathy, K., Gay, C., Dharap, S., Gaspard, R., Hughes, J.E., Sniesrud, E., Lee, N. and Quackenbush, J. 2000. A concise guide to cDNA microarray analysis. *Biotechniques* 29: 548-50, 552-4, 556 passim.

Lending, C.R. and Larkins, B.A. 1989. Changes in the zein composition of protein bodies during maize endosperm development. *Plant Cell* 1: 1011-1023.

Lotan, T., Ohto, M.A., Yee, K.M., West, M.A.L., Lo, R., Kwong, R.W., Yamagishi,

K., Fischer, R., L., Goldberg, R., B. and Harada, J., J. 1998. Arabidopsis LEAFY COTYLEDON1 is sufficient to induce embryo development in vegetative cells. *Cell* 93: 1195-1205.

Lowe, K., Abbit, S., Glassman, K., Gregory, C., Hoerster, G., Rasco-Gaunt, S., Sun, X., Lazzeri, P. and Gordon-Kamm, W. 2000. Use of maize *Lec1* to improve transformation. *In Vitro Cell. Dev. Biol.* 36: W15.

McCain, J.W. and Hodges, T.K. 1986. Anatomy of somatic embryos from maize embryo cultures. *Bot. Gaz.* 147: 453-460.

Meinke, D.W. 1992. A homeotic mutant of *Arabidopsis thaliana* with leafy cotyledons. *Science* 258: 1647-1650.

Meinke, D.W., Franzmann, L.H., Nickle, T.C. and Yeung, E.C. 1994. Leafy cotyledon mutants of *Arabidopsis*. *Plant Cell* 6: 1049-1064.

Murashige, T. and Skoog, F. 1962. A revised medium for rapid growth and bioassay with tobacco tissue cultures. *Physiol. Plant.* 15: 437-497.

Naested, H., Frandsen, G.I., Jauh, G.Y., Hernandez-Pinzon, I., Nielsen, H.B., Murphy, D.J., Rogers, J.C. and Mundy, J. 2000. Caleosins: Ca^{2+} -binding proteins associated with lipid bodies. *Plant Mol. Biol.* 44: 463-476.

Parcy, F., Valon, C., Kohara, A., Misera, S. and Giraudat, J. 1997. The ABSCISIC ACID-INSENSITIVE3, FUSCA3, and LEAFY COTYLEDON1 loci act in concert to control multiple aspects of *Arabidopsis* seed development. *Plant Cell* 9: 1265-1277.

Prioli, L.M. and Sondahl, M.R. 1989. Plant regeneration and recovery of fertile plants from protoplasts of maize (*Zea mays* L.). *Bio/Technology* 7: 589-594.

Rao, K.V., Suprasanna, P. and Reddy, G.M. 1990. Biochemical changes in embryogenic and non-embryogenic calli of *Zea mays* L. *Plant Sci.* 66: 127-130.

Schmidt, E.D.L., Guzzo, F., Toonen, M.A.J. and de Vries, S.C. 1997. A leucine-rich repeat containing receptor-like kinase marks somatic plant cells competent to form embryos. *Devel.* 124: 2049-2062.

Shillito, R.D., Carswell, G.K., Johnson, C.M., Di Maio, J.J. and Harms, C.T. 1989. Regeneration of fertile plants from protoplasts of elite inbred maize. *Bio/Technology* 7: 581-587.

Songstad, D., Armstrong, C.L., Petersen, W.L., Hairston, B. and Hinchey, M.A.W. 1996. Production of transgenic maize plants and progeny by bombardment of Hi II immature embryos. *In Vitro Cell Dev. Biol.-Plant* 32: 179-183.

Sung, Z.R. and Okimoto, R. 1981. Embryonic proteins in somatic embryos of carrot. *Proceedings of the National Academy of Sciences of the United States of America* 78: 3683-3687.

Thibaud-Nissen, F., Shealy, R.T., Khanna, A. and Vodkin, L.O. 2003. Clustering of microarray data reveals transcript patterns associated with somatic embryogenesis in soybean. *Plant Physiol.* 132: 118-36.

Vicient, C.M., Bies-Etheve, N. and Delseny, M. 2000. Changes in gene expression in the leafy cotyledon1 (*lec1*) and *fusca3* (*fus3*) mutants of *Arabidopsis thaliana* L. *J. Exp. Bot.* 51: 995-1003.

West, M.A.L.W., Yee, K.M., Danao, J., Zimmerman, J.L., Fischer, R.L., Goldberg, R.B. and Harada, J.J. 1994. *LEAFY COTYLEDON1* is an essential regulator of late embryogenesis and cotyledon identity in *Arabidopsis*. *Plant Cell* 6: 1731-1745.

Wilde, H.D., Nelson, W.S., Booij, H., de Vries, S.C. and Thomas, T.L. 1988. Gene expression programs in embryonic and non embryonic carrot cultures. *Planta* 176: 205-211.

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

Zhang, S., Wong, L., Meng, L. and Lemaux, P.G. 2002. Similarity of expression patterns of *knotted1* and *ZmLEC1* during somatic and zygotic embryogenesis in maize (*Zea mays* L.). *Planta* 215: 191-4.

2.8 Figure captions

Figure 1. Somatic embryo development in maize Hi II callus. (A) Examples of total and embryoid enriched callus growing on N6E medium. Arrow points out one of many transition stage somatic embryos in embryogenic callus. (B) Time course of somatic embryo development, maturation and germination. Somatic embryo maturation was initiated by transferring embryoid-enriched callus to Regeneration Medium I (-2,4-D, 6% sucrose). Embryos were germinated by transfer to the light on Regeneration Medium II (2,4-D, 3% sucrose). Samples were taken at time points as indicated during embryo maturation and germination for profiling gene expression patterns. Bars = 1 mm.

Figure 2. Loop design for microarray hybridization experiments. Six independent callus lines (A-F) were initiated, and the lines were pooled into 2 lines per pool to obtain enough RNA in each pool for the microarray analysis. Each rectangle represents 1 chip. (A) Time course analysis. Time point and probe dye type (either cy3 or cy5) are indicated for each chip. (B) Comparison between embryoid-enriched (E) and total callus (T). Sample source (E or T), line pool # (1, 2 or 3) and the probe dye type are indicated for each chip.

Figure 3. Summary of gene expression during somatic embryo maturation and germination in maize. Number of genes out of 1026 genes are shown that vary significantly with time and are either up or down-regulated more than 2-, 3- or 4-fold during embryo development. Shaded bar represents the period of somatic embryo development and maturation. Unshaded bar is the time of germination.

Figure 4. Genes with different expression profiles. Expression profiles of the 1026 genes with the greatest variation across time points cluster into 12 pattern groups. Examples of genes from four different pattern groups are shown here. The distributions of gene functions in the pattern group are shown in the pie charts. Genes were categorized into 19 functional groups. Means and standard errors (SEs) for 12 repeats at each time

point are shown in the line graphs. Period of embryo maturation (stippled bar), embryo germination (unshaded bar).

Figure 5. Expression profiles for four genes belonging to classes of genes expressed during zygotic embryo development or germination.

Figure 6. Line variation in gene expression. Examples of genes with significant variation across time points and that (upper panel) show significant line pool variation or (lower panel) show little line pool variation. Period of embryo maturation (stippled bar), embryo germination (unshaded bar). Line pool 1 , line pool 2 , line pool 3 .

Figure 7. Trends in expression of genes in various functional categories show significant time variation during embryo maturation and germination. Based on the expression profiles of genes that are typical of the pattern group, most of which are shown in Figs. 4 and 5. In descending order in the diagram: water channel, BM079333; alpha zein, AL795292; chlorophyll a/b binding protein, BG841274; beta-glucosidase, AW352489; heat shock protein, AI901570; proliferating cell nuclear antigen, AL734348.

Figure 2.1 Somatic embryo development in maize Hi II callus.

Figure 2.2 Loop design for microarray hybridization experiments.

Figure 2.3 Summary of gene expression during somatic embryo maturation and germination in maize.

Figure 2.4 Genes with different expression profiles.

Figure 2.5 Expression profiles for four genes belonging to classes of genes expressed during zygotic embryo development or germination.

Figure 2.6 Line variation in gene expression.

Figure 2.7 Trends in expression of genes in various functional categories show significant time variation during embryo maturation and germination.

CHAPTER 3. INCORPORATING MULTIPLE cDNA MICROARRAY SLIDE SCANS - APPLICATION TO SOMATIC EMBRYOGENESIS IN MAIZE

A paper submitted to the *Journal of the American Statistical Association*

Tanzy M. Love^{1,2,3}, Alicia L. Carriquiry¹

3.1 Abstract

Microarray data are subject to multiple sources of measurement error. One source of potentially significant error is the settings of the instruments (laser and sensor) that are used to obtain the measurements of gene expression. Because ‘optimal’ settings may vary from slide to slide, operators typically scan each slide multiple times and then choose the reading with the fewest over-exposed and under-exposed spots. We discuss a somatic embryogenesis experiment carried out on *Zea mays* at Iowa State University. The main objective of the study was to identify the set of genes in maize that actively participate in embryo development and to do so, embryo tissue was sampled and analyzed at various time periods and under different light conditions. We propose a hierarchical modeling approach to estimating gene expression that combines all available readings on each spot. The basic premise is that all readings contribute some information about

¹Department of Statistics, Iowa State University

²Primary researcher and author

³Author for correspondence, supported by National Science Foundation grant number DMS 0091953

gene expression and that after appropriate re-scaling, it would be possible to combine all readings into a single estimate. We assess the statistical properties of the proposed expression estimates using a simulation experiment. As expected, combining all available scans using a reasonable approach results in expression estimates with noticeably lower bias and root mean squared error relative to other approaches that have been proposed in the literature. We then revisit the maize experiment and present results obtained using a standard and the proposed approaches. We argue that more precise inferences on gene expression patterns are obtained when all available scans on each spot are used in the statistical analyses thus resulting in increased power of tests.

3.2 Introduction - An experiment to assess gene expression changes during maize embryogenesis

Somatic embryogenesis in *Zea mays* is an experimental important tool for genetic engineering. Natural plant development from a fertilized egg cell follows zygotic embryogenesis development into a seed and eventually a mature plant. Somatic embryos begin as callus (undifferentiated cells) and are induced to develop into embryos by immersion in an embryogenic medium. Callus can be generated from existing plants by transfer to a callus-generating medium. Mature plants can be grown from existing plant material through another experimental process called organogenesis. Somatic embryogenesis creates embryos that are similar to those arising from sexual reproduction and which have the same genotype as the explant from which they were created.

The first somatic embryos in maize tissue culture were produced by Green and Phillips (1975). Armstrong and Green (1985) found that cell lines derived from sources such as immature embryos are heterogeneous for cells with different embryogenic competence and that certain types of callus tend to be more embryogenic. Unfortunately,

embryogenic competence is genotype-specific in many plant species including *Zea Mays*, and often the most desirable or economically important lines are recalcitrant to regeneration. Since genetic transformation of the plant in the embryonic stage has enormous potential for development of high yielding varieties, it is important to recognize embryonic cells or tissues and to identify genetic markers for them. In this way we hope to gain tools for improving embryogenesis in recalcitrant maize lines.

In order to identify the genetic traits responsible for highly embryogenic lines, we examined gene expression changes during maize somatic embryo development. Somatic embryos were generated identically in six callus lines (labeled A, B, C, D, E, and F) developed from immature Hi II embryo explants. These lines are assumed to be random samples from the population of Hi II lines. Hi II is a regeneration-proficient hybrid of *Zea mays* which also produces high crop yields and thus is of economic importance. After callus populations were generated from the six lines in a callus-generating medium (N6E medium +2, 4-D, 3% sucrose), embryogenic calli (identifiable by shape) were selected from the total callus for each of the lines. These selected calli were matured into somatic embryos by transferring them to a sucrose-enhanced medium (Regen Medium I -2, 4-D, 6% sucrose). After 21 days, the embryos were exposed to light and transferred to a new medium (Regen Medium II -2, 4-D, 3% sucrose) to encourage germination. Material was sampled at five time points during the development and maturation of the embryos into seedlings, see Figure 3.1. To reduce the interplant variability without actually losing the opportunity to estimate the line to line variance in measurements, samples from pairs of lines were pooled at each time point, so the material used later in the microarray analysis comes from the pools rather than from individual plant lines. Pools were labeled AB, CD, and EF. The dataset used in this analysis can be obtained by contacting the authors.

Gene expression patterns in the AB, CD and EF lines were profiled using 12,060

Figure 3.1 The time course experiment for somatic embryogenesis in maize.

element maize cDNA arrays. The experimental design was a loop design with dye-swapping for line pools AB, CD, and EF (Kerr and Churchill 2001), which resulted in a total of 30 slides. The loop design is an efficient design in that each time point is compared to an earlier time period rather than to a reference or time zero. Thus, the design results in additional replicates of measurements at each time point without increasing the number of arrays. Under our design, gene expression is measured at each time point four times within each line pool, so that across line pools, time point samples are repeated 12 times. The design therefore allows for an analysis of the measurement variance across time and across line pools, see Figure 3.2.

We are interested in identifying the genes or groups of genes which actively participate in somatic embryogenesis. These genes will exhibit significant changes in their expression over the course of tissue development and maturation. While most of the genes partic-

Figure 3.2 The microarray double loop design with dye swap for the maize embryogenesis experiment. Each box represents one of the 30 slides created. The arrows show the direction of the loop as each time point is compared to its neighbors.

ipating in embryonic development are expected to be up-regulated as embryos mature, it is also possible that some genes active in embryogenesis will down-regulate over time or will exhibit some other expression profiles. Therefore, we seek to classify the 12,060 elements of the microarrays into those with constant expression over all time points and those with any other pattern of expression. Some initial results obtained using a subset of these data are reported in (Che et al. 2004).

As is often the case in microarray experiments, each slide was scanned multiple times using different laser and sensor settings. By varying the settings of the instruments, the operator can strike a balance between over-exposing the highly expressing genes while still picking up a signal from the lowly fluorescing spots. In a typical statistical analysis of gene expression data, only the 'best' scan of each slide is included in the analysis and the rest are discarded (Smyth et al. 2002). We propose an approach that permits estimating gene expression profiles using all available measurements for each spot and show later in this manuscript that by making use of the additional information we obtain estimates of quantities of interest that are better (in the minimum MSE-sense) than all other approaches recently reported in the literature. Importantly, the set of genes identified as embryogenic in our experiment changes if all scans, rather than just the best, are used in statistical analyses of the data.

This paper is organized as follows. In Section 3.3 we provide some background on the cDNA microarray technology and include a discussion on the effects of changing laser and sensor settings on the readings obtained for a slide. We propose a new method for estimating gene expression profiles using multiple array scans in Section 3.4. The performance of the approach is compared to that of other approaches proposed in the literature via simulation in Section 3.5. We finally revisit the original embryogenesis study and analyze the experimental data using the standard and the proposed approaches.

Results are presented in Section 3.6 and discussed in Section 3.7. Details of some of the derivations presented in the paper are given in an Appendix.

3.3 Microarray analysis

Large microarray studies are carried out to investigate many complex processes and behaviors. Typically, mRNA from biological materials that have been subjected to different treatments or that arise from different tissues or from the same tissue at different stages of development are applied to different arrays (or slides). The objective is to draw inferences about differential gene expression levels across treatments, tissues or developmental stages. Gene expression levels can only be compared across different arrays or slides after the appropriate background cleaning and normalization procedures have scaled the data to the same range. There are several important sources of variation in gene expression measurement that must be accounted for in statistical analyses, and much of this variation is array specific. In the following, we focus on the case of cDNA microarrays; however many of the same issues are pertinent in other technologies.

The data generated from cDNA microarray experiments are obtained by combining two types of images of the microarray slide. The two images are obtained while the slide is excited with a laser tuned to Cy5 and to Cy3 fluorescent dyes, respectively. Different laser strengths and the sensitivities of the camera result in different images. While a particular setting for laser and camera may produce a large number of saturated spots on the slide image, other settings may result in too many spots with measured expression below the minimum that can be captured by the instruments. The laser strength and the sensitivity of the photomultiplier to light can be adjusted by the operator to find a 'best' picture, one where most of the spots show some measurable expression and where very few of the spots reach saturation. Once the best image is obtained, the spots of cDNA

must be found, a process called segmentation, and the background of each spot must be calculated and removed from the signal. Finally, before statistical analysis of the data can proceed, each slide must be normalized within itself to recover from any systematic dye bias (usually Cy3 is stronger) and all the slides must be normalized jointly to make them comparable (Smyth et al. 2002).

In this work, we focus on the measurement error that is introduced when scientists vary the strength of the laser and the sensitivity of the photomultiplier used to amplify the expression signals and propose a modeling approach that allows incorporating multiple readings of each slide into the analysis. We show that under relatively lax model assumptions, expression levels can be estimated with significantly lower bias and higher precision when combining multiple readings for each gene into the statistical analysis than when choosing only the ‘best’ reading.

Methods for estimating gene expression that use the multiple slide scans that are typically produced in microarray experiments have been discussed in the literature in recent months. Lyng et al. (2004) and Skibbe, Nettleton, and Schnable (2004) investigated the effects of scanner settings on expression ratios and significant differential expression, respectively. Both found that scanner settings have an important impact on the quality of and conclusions from microarray data. Lyng et al. (2004) suggests using two scans at different settings to increase the usable range of expression values. Romualdi et al. (2003) suggest combining the pixel intensities over multiple slide readings by averaging them before segmentation to create more uniform spots. However, the same scanner settings must be used to make the pixels exchangeable between scans. This means that there was no improvement of the dynamic range of expression estimates by Romualdi et al. Dudley et al. (2002) and Garcia de la Nava, van Hijum, and Trelles (2004) used multiple slide readings at varying settings to extend the dynamic range and address the

censoring error. However, the former use only the estimate from one reading for each gene (possibly linearly transformed) while the latter accommodate only two scans.

3.3.1 Multiple laser and sensor settings

Different laser and sensor settings can be used to read a cDNA microarray slide. Stronger laser settings create more fluorescence and stronger sensor settings pick up more signal. There is a balance to be struck between picking up signal from the lowly fluorescing spots and over-exposing the highly expressing genes. There is an upper limit of 65535 to the measurement of fluorescence; readings of spots which are brighter are censored. Over-exposing the high intensity spots will cause them to be artificially near other high expression values. Correspondingly, low signals will be artificially assigned to 0 if the laser and sensor settings are too low.

Figure 3.3 illustrates the two scenarios. It is possible to have both overexposed and underexposed spots in the same scan. The expression estimates used are background corrected average pixel intensities (Smyth et al. 2002). Any spot will have variation in its pixels due to inconsistencies in spot printing and irregular spot shape. Further, background correction will reduce the measured expression value so that 65535 is no longer the point of censoring (see Figure 3.3 (a) where there is censoring at around 55000) and create expression estimates that are negative or near zero. The censored values in the figure include those genes that are not expressed in the sample on that particular slide, and those genes which may have exhibited measurable expression levels had the spot been more exposed. Negative expression measurements are routinely set to zero, however, often the true point of censoring is not zero (see Figure 3.3 (b) where there is censoring near ten).

Multiple readings of the microarray slides can be taken for both fluorescence channels.

Figure 3.3 (a) Many spots are censored above in the higher reading. (b) Many spots are censored below in the lower reading.

Since all of the readings at different settings attempt to capture true expression levels for the genes on the slide, it is reasonable to assume that all readings contribute useful information about true expression levels and to think of combining the multiple readings into one estimate of gene expression for each spot. If the readings at different settings contain information about the true expression of the gene, then the variance in estimated gene expression that is due to the measurement process should be reduced in the estimate that is based on all available readings.

Several aspects of the measurement process of gene expression create challenges for statistical modeling. As discussed earlier, many microarray experiments include pseudo-replicates, which we define as multiple readings of the slide under different laser and sensor settings. Generally, settings for different slides are very different because of the large experimental variation between slides. That is, one slide may result in a good reading at low laser and sensor settings while another may require higher settings to reduce the number of expression levels below threshold while keeping the number of overexposed spots to a minimum. Because of this practice, we are typically unable to assume that the settings act as blocks in a traditional experimental design. However, since the settings to read the two channels are almost always chosen separately across slides, we can model each slide/dye combination separately. In what follows, we consider an arbitrary slide and dye channel in the experiment and propose a hierarchical model for estimating gene expression levels that permits incorporating multiple measurements for each gene into a single analysis.

3.4 Bayesian Hierarchical Gamma Model

In order to estimate gene expression, we propose a Bayesian hierarchical model. This model incorporates all slide scans into one estimate of expression per spot. To formulate

the model, we rely on the natural ordering of slide readings. For instance, if we have two readings with the same sensor setting and different laser settings, the measurements on the reading with the higher laser setting will tend to be larger. Dudley et al. (2002) discuss gene expression and its dependence on changes in one of the experimental settings (laser and sensor). Here we consider changing both settings simultaneously and to do so order the slides from smallest to largest based on median reading. Clearly, the median-based ordering is subject to some uncertainty because of the measurement error in observed gene expressions.

3.4.1 Likelihood Function

Suppose that there are $m + 1$ readings taken at each of n spots on a particular slide and a dye. In the maize embryogenesis experiment that we discuss here, $m + 1 = 3$ and $n = 12,060$ for all 60 slide/dye combinations, but the number of readings need not be constant over slides. For a given gene i , we use $S_{i1}, \dots, S_{i(m+1)}$ to denote the $m + 1$ ordered signal measurements after background correction. Here S_{i1} is the gene expression measurement from the scan having smallest median (of all spots on the scan) expression and $S_{i(m+1)}$ denotes the reading for gene i on the scan with the highest median expression.

We assume that all readings measure the same quantity – actual gene expression – with error. Therefore, under suitable scaling the readings would be identically distributed. We assume that the scaled readings (which are strictly positive) can be represented by a Gamma distribution. The Gamma has support on the positive real line and, depending on parameter values, exhibits noticeable skewness. Therefore, in the absence of censoring, we could model the background corrected signals for each gene i across the $m + 1$ readings in the following way:

$$S_{ij}\chi_j = S'_{ij} \sim \Gamma(a, \psi_i)$$

for all $i = 1, \dots, n$ and $j = 1, \dots, m + 1$, where the χ_j are constant for all genes in a given slide and dye combination. This assumes that the changes in laser and sensor settings increase or decrease each spot's fluorescence by the same amount. This model can also be written as

$$S_{ij} \sim \Gamma(a, \psi_i \chi_j)$$

for all $i = 1, \dots, n$ and $j = 1, \dots, m + 1$.

As formulated, the model is not identifiable in that there is no way to estimate all the parameters, a , ψ , and χ , directly. Thus, we do not attempt to estimate ψ_i and instead focus on estimating $\theta_i = \chi_{m+1} \psi_i$. We choose the highest of the $m + 1$ readings as a reference reading and scale all other readings to that level. By scaling all readings upwards to the highest one we are increasing the effective range of gene expression measurement. This does not limit the usefulness of the model in any way, because all measures of gene expression are relative and normalization is performed on the expression estimates.

We now have the following model, still assuming that no censoring occurs:

$$S_{ij} \sim \Gamma(a, \theta_i \delta_j)$$

for all $i = 1, \dots, n$ and $j = 1, \dots, m + 1$, where the δ_j are constant for all genes in a given slide and dye combination and $\delta_{m+1} \equiv 1$. We let $S = \{S_{ij}\}$ denote the set of measurements on a particular slide and dye. The unknown parameters in this model are a , $\theta_1, \dots, \theta_n$, and $\delta_1, \dots, \delta_m$.

However, this likelihood assumes that we observe all the values of S_{ij} . Yet we do not observe intensity of spots in readings where they are censored; however, we do know that they are censored and we also know that the measurement is larger (smaller) than a known value. We define an indicator variable, C_{ij} , where $C_{ij} = 0$ if observation S_{ij} is

not censored, $C_{ij} = 1$ if observation S_{ij} is censored below, and $C_{ij} = 2$ if observation S_{ij} is censored above. This variable and the subset of S , $S^{(o)}$, which includes non-censored measurements make up our observed data. The measurements that would have been observed in the absence of censoring are therefore taken to be missing. The set of missing data is denoted by $S^{(m)}$ and $S = S^{(o)} \cup S^{(m)}$. In a Bayesian framework, we can estimate missing values along with parameters.

In practice a spot can be designated as censored below if any of its pixels are less than the background median. A spot can be designated as censored above if any of its pixels are saturated. Alternatively, exploratory data analysis can be used to decide appropriate cut-off values for a particular slide/dye combination, such as 20 and 50,000. To ensure that no gene has all of its values missing, a spot censored below in the highest scan or above in the lowest scan is not recorded as censored for that scan. We will denote the lower and upper truncation points by L and U , respectively.

We now examine the conditional likelihood of S_{ij} , given the censoring indicator, C_{ij} . Let $f(\cdot|\lambda)$ be the density function of the $\text{Gamma}(a, \lambda)$ distribution and $F(\cdot|\lambda)$ be its cumulative distribution function. Then censoring implies that the likelihood for $S_{ij} \in S^{(o)}$, an uncensored point, should have the following form:

$$p(S_{ij}|C_{ij} = 0) = f(S_{ij}|\theta_i\delta_j) (F(U|\theta_i\delta_j) - F(L|\theta_i\delta_j))^{-1} I_{(L,U)}(S_{ij}),$$

where $I_A(\cdot)$ is the identity function on the set A . For a gene expression measurement $S_{ij} \in S^{(m)}$, which is censored below, the likelihood has the following form:

$$p(S_{ij}|C_{ij} = 1) = f(S_{ij}|\theta_i\delta_j) F(L|\theta_i\delta_j)^{-1} I_{[0,L]}(S_{ij}).$$

The likelihood of $S_{ij} \in S^{(m)}$ which is censored above is

$$p(S_{ij}|C_{ij} = 2) = f(S_{ij}|\theta_i\delta_j) (1 - F(U|\theta_i\delta_j))^{-1} I_{[U,\infty)}(S_{ij}).$$

The restriction on the support of the likelihood will remain in the posterior distributions of the $S_{ij} \in S^{(m)}$.

The joint probability of S_{ij} and C_{ij} can be written in the following form:

$$\begin{aligned}
p(S_{ij}, C_{ij}) &= p(S_{ij}|C_{ij}=0)p(C_{ij}=0) + p(S_{ij}|C_{ij}=1)p(C_{ij}=1) + p(S_{ij}|C_{ij}=2)p(C_{ij}=2) \\
&= p(S_{ij}|C_{ij}=0)(F(U|\theta_i\delta_j) - F(L|\theta_i\delta_j))I_{\{0\}}(C_{ij}) \\
&\quad + p(S_{ij}|C_{ij}=1)F(L|\theta_i\delta_j)I_{\{1\}}(C_{ij}) + p(S_{ij}|C_{ij}=2)(1 - F(U|\theta_i\delta_j))I_{\{2\}}(C_{ij}) \\
&= f(S_{ij}|\theta_i\delta_j)I_{(L,U)}(S_{ij})I_{\{0\}}(C_{ij}) + f(S_{ij}|\theta_i\delta_j)I_{[0,L]}(S_{ij})I_{\{1\}}(C_{ij}) \\
&\quad + f(S_{ij}|\theta_i\delta_j)I_{[U,\infty)}(S_{ij})I_{\{2\}}(C_{ij}).
\end{aligned}$$

The full data likelihood is then the following:

$$\begin{aligned}
p(S, C) &= \prod_{i=1}^n \prod_{j=1}^{m+1} \left(f(S_{ij}|\theta_i\delta_j)I_{(L,U)}(S_{ij})I_{\{0\}}(C_{ij}) + f(S_{ij}|\theta_i\delta_j)I_{[0,L]}(S_{ij})I_{\{1\}}(C_{ij}) \right. \\
&\quad \left. + f(S_{ij}|\theta_i\delta_j)I_{[U,\infty)}(S_{ij})I_{\{2\}}(C_{ij}) \right).
\end{aligned}$$

We will divide the set of indexes, (i, j) , into three groups, one for each value of C_{ij} . Let I_N be the set of all (i, j) such that $C_{ij} = 0$, I_L be the set of all (i, j) such that $C_{ij} = 1$, I_U be the set of all (i, j) such that $C_{ij} = 2$. The full data likelihood can then be written as

$$\begin{aligned}
p(S, C) &= \prod_{(i,j) \in I_N} \left(f(S_{ij}|\theta_i\delta_j)I_{(L,U)}(S_{ij}) \right) \times \prod_{(i,j) \in I_L} \left(f(S_{ij}|\theta_i\delta_j)I_{[0,L]}(S_{ij}) \right) \\
&\quad \times \prod_{(i,j) \in I_U} \left(f(S_{ij}|\theta_i\delta_j)I_{[U,\infty)}(S_{ij}) \right).
\end{aligned}$$

This leads to the following observed data likelihood:

$$\begin{aligned}
p(S^{(o)}, C) &= \int \prod_{(i,j) \in I_N} \left(f(S_{ij}|\theta_i\delta_j)I_{(L,U)}(S_{ij}) \right) \times \prod_{(i,j) \in I_L} \left(f(S_{ij}|\theta_i\delta_j)I_{[0,L]}(S_{ij}) \right) \\
&\quad \times \prod_{(i,j) \in I_U} \left(f(S_{ij}|\theta_i\delta_j)I_{[U,\infty)}(S_{ij}) \right) dS^{(m)} \\
&= \prod_{(i,j) \in I_N} \left(f(S_{ij}|\theta_i\delta_j)I_{(L,U)}(S_{ij}) \right) \times \prod_{(i,j) \in I_L} \int f(S_{ij}|\theta_i\delta_j)I_{[0,L]}(S_{ij}) dS_{ij}
\end{aligned}$$

$$\begin{aligned}
& \times \prod_{(i,j) \in I_U} \int f(S_{ij} | \theta_i \delta_j) I_{[U, \infty)}(S_{ij}) dS_{ij} \\
& = \prod_{(i,j) \in I_N} \left(f(S_{ij} | \theta_i \delta_j) I_{(L, U)}(S_{ij}) \right) \times \prod_{(i,j) \in I_L} F(L | \theta_i \delta_j) \times \prod_{(i,j) \in I_U} (1 - F(U | \theta_i \delta_j))
\end{aligned}$$

The mean of the Gamma distribution for the expression of gene i is a/θ_i . Within a classical framework, an estimate of expression level for the i th gene would be based on the corresponding mean. In a Bayesian framework, inference would be based on the posterior distribution of a/θ_i . In both cases, these estimates still require normalization so that expressions observed for different slide/dye combinations can be compared.

3.4.2 Prior Distributions

We adopt a Bayesian approach to estimating the parameters in the model. In order to do so, we must complete the specification of the model by assigning prior distributions to each parameter. We restrict our attention to proper prior distributions to guarantee integrability of the posterior, and within the family of proper distributions we focus on the conjugate or semi-conjugate families to attempt to simplify computations wherever possible. If the prior distribution for the parameters is conjugate, then the posterior will have the same form as the prior. If a prior distribution for a set of parameters (α, β) is semi-conjugate, then the conditional posterior distributions of $\alpha | \beta$ and $\beta | \alpha$ have the same form as the priors on α and β , respectively. However, in this case, the joint posterior of (α, β) is not of the same form as the joint prior on (α, β) .

We assume that the scale parameters, $\theta_1, \dots, \theta_n$ and the scaling parameters $\delta_1, \dots, \delta_m$ arise from a common population distribution. Let

$$p(\theta_1, \dots, \theta_n, \delta_1, \dots, \delta_m) = p(\theta, \delta)$$

represent a joint prior distribution that for now will remain unspecified. We derive a joint posterior distribution for the vectors $\theta = (\theta_1, \dots, \theta_n)$ and $\delta = (\delta_1, \dots, \delta_m)$ and then

determine the form of the prior distribution $p(\theta, \delta)$ that would be conjugate for the likelihood.

Conditional on the shape parameter a and on the data S , the joint posterior distribution of (θ, δ) is given by

$$\begin{aligned}
 p(\theta, \delta | S, a) &\propto p(\theta, \delta) * p(S | \theta, \delta, a) \\
 &= p(\theta, \delta) \prod_{i=1}^n \prod_{j=1}^{m+1} \text{Gamma}(S_{ij} \delta_j | a, \theta_i) \\
 &= p(\theta, \delta) \prod_{i=1}^n \prod_{j=1}^{m+1} \frac{\theta_i^a}{\Gamma(a)} (S_{ij} \delta_j)^{a-1} e^{-\theta_i (S_{ij} \delta_j)} \\
 &\propto p(\theta, \delta) \left(\prod_{i=1}^n \theta_i^{a(m+1)} \right) \left(\prod_{j=1}^m \delta_j^{n(a-1)} \right) \exp \left(- \sum_{i=1}^n \theta_i \sum_{j=1}^{m+1} \delta_j S_{ij} \right).
 \end{aligned}$$

A conjugate prior for θ and δ would have the form

$$p(\theta, \delta) \propto \prod_{i=1}^n \theta_i^{\phi_1} \prod_{i=1}^m \delta_j^{\phi_2} e^{-\phi_3 \sum_{i=1}^n \theta_i \sum_{j=1}^{m+1} \delta_j}.$$

This distribution is difficult to interpret from a biological viewpoint and further, implies a prior dependency between θ and δ which we cannot justify. Thus, the conjugate prior option, while convenient from a mathematical viewpoint appears to be unsuitable from a biological viewpoint. We consider instead independent Gamma prior distributions for each of the $n+m$ parameters. Gamma distributions can be justified from a biological point of view because typically genes spotted on a slide exhibit low expression levels and only some of them exhibit high levels of expression. The Gamma distribution would appear to be an appropriate model for the population distribution because the expression values of the genes, estimated by a/θ_i , will be skewed. Thus

$$\theta_i \sim \Gamma(a_0, \nu)$$

for $i = 1, \dots, n$. The Gamma model may also be reasonable for the strictly positive scaling parameters, so that

$$\delta_j \sim \Gamma(\alpha_1, \alpha_2)$$

for $j = 1, \dots, m$. The joint Gamma prior has the form

$$p(\theta, \delta) \propto \prod_{i=1}^n \theta_i^{a_0} \prod_{j=1}^m \delta_j^{\alpha_1} e^{-\nu \sum_{i=1}^n \theta_i - \alpha_2 \sum_{j=1}^{m+1} \delta_j}. \quad (3.1)$$

The conditional posterior distributions of $\theta|\delta$ and $\delta|\theta$ are Gamma distributions under this prior, but the joint posterior of (θ, δ) is not. Therefore, the prior in (3.1) is a semi-conjugate prior distribution.

3.4.3 Estimating the hyperparameters

The hyperparameters in the model are $\eta = (a, a_0, \nu, \alpha_1, \alpha_2)$. We must either specify prior distributions for these hyperparameters or fix the parameters at some appropriate value. The hyperparameters α_1 and α_2 are both chosen to be 10 to create a relatively noninformative prior on the δ 's. Specifying a value for the other hyperparameters a , a_0 and ν , however, requires some thought since these parameters can have a significant effect on the estimates of expression levels.

One approach to obtaining values for hyperparameters is to find the values $(\hat{a}, \hat{a}_0, \hat{\nu})$ that maximize the marginal likelihood of the parameters (MMLEs, e.g., Carlin and Louis, 2003). The marginal likelihood $p(a, a_0, \nu|S)$ is obtained by integrating (δ, θ) out of the joint likelihood function as follows (the complete derivation of $p(a, a_0, \nu|S)$ is presented in the Appendix):

$$p(a, a_0, \nu|S) = \int \int p(a, a_0, \nu, \theta, \delta|S) d\delta d\theta$$

$$\begin{aligned}
& \propto \Gamma(a)^{-n(m+1)} \Gamma(na + \alpha_1)^m \prod_{i=1}^n \prod_{j=1}^{m+1} S_{ij}^{a-1} \\
& \int \prod_{i=1}^n \theta_i^{a(m+1)+a_0-1} e^{-\nu \sum_{i=1}^n \theta_i} \prod_{j=1}^m \left(\sum_{i=1}^n \theta_i S_{ij} - \alpha_2 \right)^{-(na+\alpha_1)} d\theta. \quad (3.2)
\end{aligned}$$

This marginal distribution is not analytically tractable. However, we could integrate δ out analytically if instead of conditioning on S we were to derive the marginal likelihood given only expression values from the largest reading, $S_{\bullet(m+1)}$. Here, we have chosen to use the largest reading because it is not scaled. Any reading $S_{\bullet j}$ used as the standard with the model adjusted to estimate $\theta = \chi_j \psi$ would be used in the same way. The subsequent normalization that expression estimates must undergo makes any choice of standard reading equivalent. In this case,

$$\begin{aligned}
p(a, a_0, \nu | S_{\bullet(m+1)}) &= \int \int p(a, a_0, \nu, \theta, \delta | S_{\bullet(m+1)}) d\delta d\theta \\
&\propto \int \int p(S_{\bullet(m+1)} | a, a_0, \nu, \theta) p(\theta | a_0, \nu) p(\delta) d\delta d\theta \\
&= \int p(S_{\bullet(m+1)} | a, a_0, \nu, \theta) p(\theta | a_0, \nu) d\theta \\
&= \int \Gamma(a)^{-n} \prod_{i=1}^n \theta_i^a S_{i(m+1)}^{a-1} \exp(-\theta_i S_{i(m+1)}) \\
&\quad \Gamma(a_0)^{-n} \nu^{na_0} \prod_{i=1}^n \theta_i^{a_0-1} \exp(-\nu \theta_i) d\theta \\
&\propto (\Gamma(a)\Gamma(a_0))^{-n} \left(\prod_{i=1}^n S_{i(m+1)} \right)^{a-1} \nu^{na_0} \\
&\quad \int \left(\prod_{i=1}^n \theta_i^{a+a_0-1} \right) \exp \left(- \sum_{i=1}^n \theta_i (S_{i(m+1)} + \nu) \right) d\theta \\
&= (\Gamma(a)\Gamma(a_0))^{-n} \left(\prod_{i=1}^n S_{i(m+1)} \right)^{a-1} \nu^{na_0} \prod_{i=1}^n \Gamma(a+a_0) (S_{i(m+1)} + \nu)^{-a-a_0} \\
&= \left(\frac{\Gamma(a+a_0)}{\Gamma(a)\Gamma(a_0)} \right)^n \left(\prod_{i=1}^n S_{i(m+1)} \right)^{a-1} \nu^{na_0} \prod_{i=1}^n (S_{i(m+1)} + \nu)^{-a-a_0}. \quad (3.3)
\end{aligned}$$

The resulting expression can now be maximized with respect to a , a_0 and ν using standard nonlinear optimization techniques.

However, conditioning on $S_{\bullet(m+1)}$ may lead to poor estimates of the hyperparameters because $S_{\bullet(m+1)}$ may include censored spots which reduce the variability of the data. Since it is the highest reading, we expect it to have more such spots than moderate readings. With a bit more computation, we can estimate a , a_0 and ν using any reading, $S_{\bullet j}$ as follows

$$\begin{aligned}
p(a, a_0, \nu | S_{\bullet j}) &= \int \int p(a, a_0, \nu, \theta, \delta | S_{\bullet j}) d\delta d\theta \\
&\propto \int \int p(S_{\bullet j} | a, \theta, \delta_j) p(\theta | a_0, \nu) p(\delta) d\delta d\theta \\
&= \int \int \Gamma(a)^{-n} \prod_{i=1}^n \theta_i^a \delta_j^a S_{ij}^{a-1} \exp(-\theta_i \delta_j S_{ij}) \\
&\quad \frac{\nu^{na_0}}{\Gamma(a_0)^n} \prod_{i=1}^n \theta_i^{a_0-1} \exp(-\nu \theta_i) \frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} \delta_j^{\alpha_1-1} \exp(-\alpha_2 \delta_j) d\theta d\delta_j \\
&= \frac{\nu^{na_0} \alpha_2^{\alpha_1}}{\Gamma(a)^n \Gamma(a_0)^n \Gamma(\alpha_1)} \left(\prod_{i=1}^n S_{ij} \right)^{a-1} \int \delta_j^{na+\alpha_1-1} \exp(-\alpha_2 \delta_j) \\
&\quad \int \prod_{i=1}^n \theta_i^{a+a_0-1} \prod_{i=1}^n \exp(-\theta_i (\delta_j S_{ij} + \nu)) d\theta d\delta_j \\
&= \frac{\nu^{na_0} \alpha_2^{\alpha_1}}{\Gamma(a)^n \Gamma(a_0)^n \Gamma(\alpha_1)} \left(\prod_{i=1}^n S_{ij} \right)^{a-1} \int \delta_j^{na+\alpha_1-1} \exp(-\alpha_2 \delta_j) \\
&\quad \prod_{i=1}^n \left(\Gamma(a + a_0) (\delta_j S_{ij} + \nu)^{-a-a_0} \right) d\delta_j \\
&= \frac{\Gamma(a + a_0)^n \nu^{na_0} \alpha_2^{\alpha_1}}{\Gamma(a)^n \Gamma(a_0)^n \Gamma(\alpha_1)} \left(\prod_{i=1}^n S_{ij} \right)^{a-1} \\
&\quad \int \delta_j^{na+\alpha_1-1} \exp(-\alpha_2 \delta_j) \prod_{i=1}^n (\delta_j S_{ij} + \nu)^{-a-a_0} d\delta_j.
\end{aligned}$$

An efficient approach to finding the values $(\hat{a}, \hat{a}_0, \hat{\nu})$ that maximize (3.4) is the EM algorithm (Dempster, Laird, and Rubin 1977). In the E-step, we find the conditional posterior expectation of δ_j given (a, a_0, ν) :

$$\begin{aligned}
E(\delta_j | a, a_0, \nu, S_{\bullet j}) &= \int \delta_j p(\delta_j | a, a_0, \nu, S_{\bullet j}) d\delta_j \\
&= \int \delta_j \int p(\delta_j, \theta | a, a_0, \nu, S_{\bullet j}) d\theta d\delta_j
\end{aligned}$$

$$\begin{aligned}
&= C_1 \int \delta_j \int p(S_{\bullet j}, \delta_j, \theta | a, a_0, \nu) d\delta_j d\theta \\
&= C_2 \int \delta_j \int \delta_j^{na+\alpha_1-1} \exp(-\delta_j \alpha_2) \prod_{i=1}^n \theta_i^{a+a_0-1} \exp(-\theta_i(\delta_j S_{ij} + \nu)) d\theta d\delta_j \\
&= C_2 \int \delta_j^{na+\alpha_1} \exp(-\delta_j \alpha_2) \prod_{i=1}^n (\delta_j S_{ij} + \nu)^{-a-a_0} d\delta_j \\
&= \frac{\int \delta_j^{na+\alpha_1} \exp(-\delta_j \alpha_2) \prod_{i=1}^n (\delta_j S_{ij} + \nu)^{-a-a_0} d\delta_j}{\int \delta_j^{na+\alpha_1-1} \exp(-\delta_j \alpha_2) \prod_{i=1}^n (\delta_j S_{ij} + \nu)^{-a-a_0} d\delta_j}. \tag{3.4}
\end{aligned}$$

In the M-step, we compute the MLE for (a, a_0, ν) given δ_j . The conditional likelihood to be maximized is

$$\begin{aligned}
L(a, a_0, \nu | \delta_j, S_{\bullet j}) &= \int p(a, a_0, \nu, \theta | \delta_j, S_{\bullet j}) d\theta \\
&\propto \int p(S_{\bullet j} | a, \theta, \delta_j) p(\theta | a_0, \nu) d\theta \\
&= \int \frac{\prod_{i=1}^n S_{ij}^{a-1}}{\Gamma(a)^n} \frac{\nu^{na_0}}{\Gamma(a_0)^n} \prod_{i=1}^n \theta_i^{a+a_0-1} \exp(-\theta_i(\delta_j S_{ij} + \nu)) d\theta \\
&= \left(\frac{\Gamma(a+a_0) \nu^{a_0}}{\Gamma(a) \Gamma(a_0)} \right)^n \left(\prod_{i=1}^n S_{ij} \right)^{a-1} \prod_{i=1}^n (\delta_j S_{ij} + \nu)^{-a-a_0}. \tag{3.5}
\end{aligned}$$

The algorithm iterates between the E-step and the M-step until reaching convergence.

One disadvantage of proceeding within an empirical Bayes framework and fixing hyperparameters to point values is that we ignore the uncertainty about the hyperparameters when estimating other parameters in the model. To ameliorate this problem we consider placing a prior distribution on ν that again has the form of a $\text{Gamma}(\beta_1, \beta_2)$. Here, we specify $\beta_1 = 100$ and $\beta_2 = 1$ to create a relatively noninformative prior for ν . In the maize embryogenesis experiment that we describe later in this manuscript we checked sensitivity to the choice of values of β and found that inferences are robust to choices of the parameter value. If we now attempt to obtain the MMLE of (a, a_0) as before using data from all the readings, we again obtain an expression for the marginal likelihood that is computationally intractable. However, approximating the marginal likelihood by conditioning only on the highest reading $S_{\bullet(m+1)}$ again results in an ap-

proximate marginal distribution for a and a_0 that can be maximized. The derivation is presented in the Appendix. The expression to be maximized with respect to (a, a_0) is the following:

$$\begin{aligned}
 p(a, a_0 | S_{\bullet(m+1)}) &= \int \int \int p(a, a_0, \theta, \delta, \nu | S_{\bullet(m+1)}) d\delta d\theta d\nu \\
 &\propto \left(\frac{\Gamma(a + a_0)}{\Gamma(a)\Gamma(a_0)} \right)^n \prod_{i=1}^n S_{i(m+1)}^{a-1} \\
 &\quad \int \nu^{na_0+\beta_1-1} \exp(-\beta_2\nu) \prod_{i=1}^n (S_{i(m+1)} + \nu)^{-a-a_0} d\nu. \quad (3.6)
 \end{aligned}$$

Again, to find the values $(\hat{a}, \hat{a}_0)$ that maximize (3.6) we use the EM algorithm. In the E-step, we find the conditional posterior expectation of ν given a and a_0 :

$$\begin{aligned}
 E(\nu | a, a_0, S_{\bullet(m+1)}) &= \int \nu p(\nu | a, a_0, S_{\bullet(m+1)}) d\nu \\
 &= \frac{\int \nu^{na_0+\beta_1} \exp(-\beta_2\nu) \prod_{i=1}^n (S_{i(m+1)} + \nu)^{-(a+a_0)} d\nu}{\int \nu^{na_0+\beta_1-1} \exp(-\beta_2\nu) \prod_{i=1}^n (S_{i(m+1)} + \nu)^{-(a+a_0)} d\nu}. \quad (3.7)
 \end{aligned}$$

In the M-step, we compute the MLE for a and a_0 given ν . The conditional likelihood to be maximized is

$$\begin{aligned}
 L(a, a_0 | \nu, S_{\bullet(m+1)}) &= \int \int p(a, a_0, \theta, \delta | \nu, S_{\bullet(m+1)}) d\delta d\theta \\
 &\propto \left(\frac{\Gamma(a + a_0)}{\Gamma(a)\Gamma(a_0)} \right)^n \nu^{na_0} \\
 &\quad \left(\prod_{i=1}^n S_{i(m+1)} \right)^{a-1} \prod_{i=1}^n (S_{i(m+1)} + \nu)^{-(a+a_0)}. \quad (3.8)
 \end{aligned}$$

The algorithm iterates between the E-step and the M-step until reaching convergence. The derivation of expressions (3.7) and (3.8) is given in the Appendix.

The final alternative for estimating (a, a_0, ν) is the fully Bayesian method. We assign chosen priors to each parameter and estimate (a, a_0, ν) along with the rest of the

parameter values. All three parameters are restricted to be strictly positive. This led us to assign independent gamma or exponential priors to each of them. No differences in expression estimates were observed when we varied the values of the parameters of these priors. In practice, we use the following three priors:

$$a \sim \Gamma(5, 1)$$

$$a_0 \sim \Gamma(4, 1)$$

$$\nu \sim \Gamma(100, 1).$$

3.4.4 Posterior distributions

The joint posterior distribution of (δ, θ) is given by:

$$p(\delta, \theta | S^{(o)}, C, \eta) \propto \left(\prod_{j=1}^m \delta_j \right)^{na + \alpha_1 - 1} e^{-\sum_{j=1}^{m+1} \delta_j (\sum_{i=1}^n \theta_i S_{ij} - \alpha_2)} \left(\prod_{i=1}^n \theta_i \right)^{a(m+1) + a_0 - 1} e^{-\nu \sum_{i=1}^n \theta_i}.$$

The posterior distribution of $S^{(m)}$ is

$$p(S^{(m)} | S^{(o)}, C, \eta) = \int \int p(S^{(m)} | \theta, \delta, S^{(o)}, C, \eta) p(\theta, \delta | S^{(o)}, C, \eta) d\theta d\delta.$$

We use Markov chain Monte Carlo (MCMC) methods to approximate the joint posterior distribution of the parameters and missing values in the model. To do so, we first derive the full conditional distributions for each of them:

$$\delta_j | \eta, \delta_{-j}, \theta, S, C \sim \Gamma(na + \alpha_1, \sum_{i=1}^n \theta_i S_{ij} + \alpha_2)$$

for $j = 1, \dots, m$,

$$\theta_i | \eta, \delta, \theta_{-i}, S, C \sim \Gamma((m+1)a + a_0, S_{i(m+1)} + \sum_{j=1}^m \delta_j S_{ij} + \nu)$$

for $i = 1, \dots, n$,

$$S_{ij} | \eta, \delta, \theta, S^{(o)}, S_{-ij}^{(m)}, C \sim \Gamma(a, \theta_i \delta_j) I_{[0, L]}(S_{ij})$$

for $(i, j) \in I_L$, and

$$S_{ij} | \eta, \delta, \theta, S^{(o)}, S_{-ij}^{(m)}, C \sim \Gamma(a, \theta_i \delta_j) I_{[U, \infty)}(S_{ij})$$

for $(i, j) \in I_U$. Here, sampling from the last distribution is equivalent to drawing from $\Gamma(a, \theta_i \delta_j)$ and rejecting the draw if it is less than U .

Notice that all full conditional distributions have standard form, and thus the Gibbs sampler can be used to sequentially draw parameter values from the conditionals. If we have chosen a fully Bayesian estimation of (a, a_0, ν) , then the full conditionals of (a, a_0, ν) , are included in the Gibbs sampler. A point estimate for the expression of the i th gene is the posterior mean of a/θ_i . These estimates may be subsequently used as the expression values for further normalization.

3.5 Performance assessment via a simulation experiment

Before applying the proposed approach to the data collected in the maize embryogenesis experiment, we assessed its performance via simulation. We designed a simulation study to examine the differences in bias and root mean squared error (RMSE) of gene expression estimates between different approaches to estimate gene expression. A cDNA microarray dataset read at $m + 1$ reading levels was simulated from the hierarchical model. Gene expression was then estimated using the Bayesian hierarchical model we propose here, and also using the average gene expression over the $m + 1$ readings, the geometric mean of the gene expression over the $m + 1$ readings, and a linear extrapolation method that was recently proposed in the literature (Dudley et al. 2002). We replicated the experiment 100 times, and computed average bias and RMSE over the 100 replicates for each gene.

The values of the hyperparameters that we chose for the simulation are given in Table 3.1. These values were used in every replicate. We then generated values for the

Parameter	Value
n	1,000
$m + 1$	3
a	20
a_0	1
ν	12
$\alpha_1 = \alpha_2$	10

Table 3.1 Values of the hyperparameters used in simulation experiment.

θ 's and δ 's from the population distributions given the chosen hyperparameters. These θ 's and δ 's were in turn used to generate a set of observed expression values for $m + 1 = 3$ readings on $n = 1,000$ spots.

3.5.1 Estimation of scaling parameters

The method performs very well when estimating the scaling parameters $\delta_1, \dots, \delta_m$. The average bias over 100 replications was -0.00047 and the average root mean squared error (RMSE) was 0.00568. To put these values in context, the average value of δ in these simulations was 5.3. The linear extrapolation method proposed by Dudley et al. (2002) relies on the same assumption that the scaling between scans is constant across the slide. However, only one scan value (possibly scaled) is used to estimate gene expression. The average bias in estimating the scaling constants under the linear model was 0.07597 and the average RMSE was 0.10260. Both of the estimation methods use the same assumption that there is a linear relationship between the readings, so we expect their estimates to be close.

3.5.2 Estimation of gene expression

We fitted the hierarchical model we propose to each of the 100 simulated datasets using the $m + 1$ readings available for each spot in each replicate. The posterior means

Method	Average Absolute Bias	RMSE
Hierarchical model using $m + 1$ readings	257.45	499.71
Average observed gene expression	1065.94	1117.12
Geometric average observed expression	1217.33	1286.10
Linear extrapolation	441.89	1565.42
Hierarchical model using highest reading	622.10	864.14
Naive (highest reading)	619.11	863.15

Table 3.2 Comparing expression estimation methods. Simulation results for 1000 genes after 100 simulations.

of the a/θ 's (mean expression values) were used as estimates of the true expression values. We also calculated the average observed expression value from $m + 1$ readings, the geometric average observed expression value, the estimates obtained by linearly extrapolating between readings (Dudley et al. 2002) and the posterior mean expression values under a Bayesian hierarchical model similar to ours but that relies on only one reading per slide (Newton et al. 2001). All of the estimates were compared to a naive gene expression estimate obtained by simply using the value from the highest scan (by median ranking) as the estimate. This is close to the standard method; however it is ad hoc and there is no clear way to implement it in simulation.

The biases for each of the 1,000 expression values were calculated for each of the 100 simulated datasets. The average absolute bias and RMSE for each expression value were calculated over the simulations. The range of expression values for these simulations was 424891; of the 1000 expression values, three were saturated in all three readings and another three were saturated in one or two scans. Simulations were run with many more or no genes being always saturated and the results are the same; bias increases for all methods with more always-saturated genes and our method remains better in both the bias and RMSE sense. Results are presented in Table 3.2.

Figure 3.4 Log true expression by average bias over 100 simulations for expression values less than 65535.

Results suggest that gene expression estimates obtained by implementing the hierarchical model that we propose are better (in the minimum bias and RMSE sense) than estimates obtained as arithmetic or geometric averages and from the hierarchical model that relies on a single slide reading. Even though gene expression is estimated with similar bias by linearly interpolating between the multiple readings, the hierarchical model proposed here for all scans has a significantly lower RMSE. This means that the sampling variance of the estimates from the model is much smaller. For expression values below the censoring limit (65535), the biases are evenly spread above and below zero, see Figure 3.4. For values above the the censoring limit, biases are uniformly negative.

Method	Average Range	Average Bias	RMSE
Hierarchical model using $m + 1$ readings	240732	-184159	225372
Hierarchical model using one reading	65993	-358898	358898
Naive (highest reading)	65507	-359385	359385

Table 3.3 Average range in simulation.

3.5.3 Improvements over estimates that rely on a single reading

The model we propose here is an extension of the Gamma hierarchical model proposed by Newton et al. (2001) that relies on a single reading per slide and can be used to obtain the posterior probability of differential expression.

Results presented in Table 3.2 suggest that the sampling variance of gene expression estimates is reduced when all available readings are used for estimation. We now argue that by incorporating all available information about gene expression, it is also possible to increase the dynamic range of gene expression estimates in the microarray. Table 3.3 shows the average ranges that result from the application of three different methods. The average range (over the 100 simulated datasets) of the actual gene expressions in these simulations was 424891, so Table 3.3 shows that by appropriately combining the three available scans we manage to recover more of the range of the actual expressions than using one scan. Here, the bias is defined as the difference between the simulated range (maximum true expression minus minimum true expression) and the estimated range (maximum estimate minus minimum estimate), and is averaged over the 100 replicates. The RMSE is defined in a similar manner.

3.6 Maize embryogenesis experiment

We now revisit the maize embryogenesis experiment that was introduced earlier. The main objective of the experiment was to determine the subset of the 12,060 genes that vary significantly in the process of somatic embryogenesis in maize. Thirty cDNA microarray slides were spotted in the course of the experiment, which resulted in 60 slide/dye combinations on which to implement one of the standard approaches as well as the new method proposed here. We first analyzed the gene expression data using background-corrected expression as the gene expression estimate. In order to make results comparable, however, we also carried out some of the analyses applying the hierarchical modeling approach proposed by Newton et al. (2001). Each of the 60 slide/dye combinations were scanned three times at different laser and sensor settings. Unequal numbers of scans per slide would not have limited the application of the procedure.

3.6.1 Single gene-level inferences

We first present a direct comparison of the results that would be obtained at the single gene-level if one or all available readings are used to estimate expression. Two hierarchical models were fit to one dye channel of one slide: the hierarchical model proposed here, that incorporates all readings and the hierarchical model proposed by Newton et al. (2001) that incorporates information from only one scan. In this example, the Newton model was applied to the highest available reading for the gene. We present the results obtained for two arbitrarily chosen genes. Examination of those results suggests that by combining all readings for a spot we realize several improvements.

For the slide we used for illustration, the point estimates (posterior means) for the scaling parameters δ_1, δ_2 were $\hat{\delta}_1 = 2.181$ and $\hat{\delta}_2 = 1.142$ so that scaling of the three gene expression measurements results in $2.2S_{i1} \approx 1.1S_{i2} \approx S_{i3}$. The first gene on which

we focus is gene labeled #1 for which we obtained a posterior expression estimate of 276.2 based on its three measurements of (51.3, 211.0, 227.3). When using only its highest measurement (227.3), the estimate obtained was 593.5 with a 95% posterior credible set of (170, 1735). Notice that the gene expression estimate based on the three readings is within the 95% posterior probability interval. Consider now gene #1735, for which the highest reading was censored at 0 due to very large within-spot measurement error. The posterior point estimate of expression for this gene was 153.9, based on three measurements of (59.9, 77.2, 0). When using only the highest measurement, the estimate of gene expression was 368.7 with a 95% posterior credible set of (112, 1080). In this case, the estimate obtained by combining the three readings for the gene was also contained in the 95% posterior probability interval.

We argue that by combining multiple measurements into the estimate of gene expression for a single gene the resulting estimator has lower standard error. In fact, the posterior standard deviations of expression of genes #1 and #1735 in our example were 124 and 70, respectively when using the three measurements but increased to 919 and 307 when only one scan was used for estimation. Therefore, the posterior distributions of gene expression are much less concentrated around the mean when only one reading is used in estimation.

3.6.2 Gene expression profiles

In this section, we focus on inferences about gene expression profiles and compare results that are obtained by using a standard and the proposed approaches. Here, the standard approach consists of using the observed, background-corrected expression value in the 'best' scan as the estimate of gene expression. We grouped genes according to their expression profiles over time (Che et al. 2004). Figure 3.5 shows errorbar plots for example genes from six of the groups that were identified by Che et al. (2004).

When all readings are used, we get the errorbar plots in Figure 3.6. Though the patterns remain almost unchanged, it is clear from the figure that variation in gene expression estimates has been reduced.

We have also identified genes with significant difference in gene expression across the three line pools used in our experiment (Che et al. 2004). In Figure 3.7 we plot expression estimates of four genes that have significant expression variation over time. Two of these also exhibit significant variation across lines while two do not. The same four genes are shown in Figure 3.8 where now gene expression estimates are based on all available readings. Reduction in expression estimate variation is again shown here when examining between line variability over time.

3.6.3 Inferences about time and line effects

In order to identify line and time effects we fitted two-way analysis of variance models (ANOVA) to the estimated expression for each of the 12,060 genes. We are interested in identifying those genes for which time effects, line effects or both are statistically significant after controlling for multiple comparisons using the approach proposed by Benjamini and Hochberg (1995).

When gene expression is estimated using the observed background-corrected reading for each gene, 1,026 genes were identified as exhibiting significantly differential expression over time. When the three slide readings per slide/dye combination were used to estimate gene expression, we found that 3,840 genes appear to have significantly different expression levels at different time points, thus indicating that these genes are regulated during somatic embryogenesis. Note that the number of genes identified as differentially expressed during somatic embryogenesis approximately tripled when all available measurements on each spot were used for analysis. This result was to be expected given

Figure 3.5 Errorbar plots for some example genes from Che et al. (2004) (Fig 4).

Figure 3.6 Errorbar plots for the example genes from Che et al. (2004) (Fig 4) after using all readings.

Figure 3.7 Errorbar plots for some example genes from Che et al. (2004)(Fig 6). The top two genes have significant differences in gene expression between the three lines and the line difference are not significant in the bottom two genes.

Figure 3.8 Errorbar plots for the example genes from Che et al. (2004)(Fig 6) after using all readings. The top two genes have significant differences in gene expression between the three lines and the line difference are not significant in the bottom two genes.

the reduction in bias and RMSE in gene expression estimates that was achieved by implementing the procedure we propose.

Of the 1,026 genes that were identified as embryogenic using only one scan, 780 were also included in the longer list of differentially expressed genes identified when all readings were utilized. Given that the Benjamini and Hochberg method for correcting for multiple comparisons is not conservative, we expect that some of the genes labeled differentially expressed are actually false positives. If we adjust for multiple comparisons using a much more conservative correction such as a Bonferroni adjustment, only 187 genes are designated as differentially expressed when one reading per slide/dye is used. When all available readings are incorporated into the analysis, 843 genes are tagged as being differentially expressed. Almost every gene identified as differentially expressed using one reading is also identified as differentially expressed when the three readings are combined. Only eight out of the 187 genes are not also included in the longer list of 843.

In Che et al. (2004) we counted the number of genes with estimated two and three-fold changes in expression relative to the day seven measurement (Che et al. 2004). By so doing we obtain an overall assessment of gene activity during somatic embryogenesis. Figure 3.9 shows the number of genes with different expression values at each time point relative to day seven and was obtained using gene expression estimates based on only one scan per gene. We say that a gene is up-regulated if its expression increased and that it is down-regulated if its expression value decreased. From the figure we see that in the absence of light, more of the active genes are down-regulated than up-regulated. The light was turned on on day 23 of the experiment. At days 23 and 28, many more genes become active and exhibit two- and three-fold expression changes relative to day seven. This was an expected result; light triggers photomorphogenesis,

Figure 3.9 Numbers of genes with fold changes in expression from Che et al. (2004)(Fig 3).

a complex biological process that is known to involve many genes. Note too that about 3% to 5% of the genes are down-regulated at days 14 and 21 (relative to day seven). The number of down-regulated genes then falls off dramatically at days 23 and 28. Upon further investigation, these genes were found to be largely histone and ribosomal protein genes, which may be downregulated as a result of a slowing down in cell proliferation and growth during embryo maturation. We carried out the same analysis using the three readings available for each slide and the hierarchical modeling approach proposed in this manuscript. We did not find noticeable differences in the conclusions that would be drawn from the one-scan or the three-scan analyses. Using gene expression estimates generated from all available readings, we obtain a similar pattern of numbers of genes up- or down-regulated during the course of the experiment (see Figure 3.10).

Figure 3.10 Numbers of genes with fold changes in expression from Che et al. (2004)(Fig 3) after using all readings.

3.7 Discussion

Data collected in the course of microarray experimentation is subject to multiple sources of measurement error. Some of the measurement error may actually introduce biases and analysts typically attempt to reduce those biases by re-scaling and normalizing the data prior to analysis. One source of potentially significant measurement error is the settings of the instruments (laser and sensor) that are used to obtain the data. Because the 'optimal' settings may vary from slide to slide, operators often obtain multiple readings of each slide and then choose the 'best', meaning the reading that includes the fewest saturated spots and the fewest under-exposed spots.

The use of multiple scans obtained under the same laser and sensor settings have been proposed as a means to reduce the variability of gene expression estimates (Romualdi et al. 2003). Yet improving homogeneity of spots and accounting for the purely random measurement error should be possible using effective segmentation and background cleaning methods. It has been only recently that some attention has been focused on analytical methods that might permit incorporating multiple slide scans obtained under different measurement conditions into statistical analyses. Several approaches have been proposed in the literature for doing so (Dudley et al., 2002; Lyng et al., 2004; Garcia de la Nava et al., 2004). In this manuscript, we propose a general hierarchical modeling approach that allows incorporation of as many readings as may be available for each slide into the model, even if the number of readings per slide vary across slides. The basic premise is that each reading of a spot contains some information about the true expression of the gene and that if an appropriate scaling factor for each spot can be estimated, then all readings for a spot estimate the same quantity and can be combined. If so, then it is to be expected that the estimate of gene expression will have smaller variance than it would have if based on a single spot measurement.

We make several modeling assumptions in our work. For example, we assume that a single multiplicative factor is applicable to expression levels of all spots on a slide. That is, if a specific laser and sensor setting tends to increase expression levels, we assume that the multiplicative factor is uniform across all spots on a slide. This assumption may not hold in all situations, but modeling each spot within a slide individually makes the problem intractable from an analytical point of view. Simulation results show that the bias with which we can estimate gene expression is associated to expression levels, indicating that different spots on the slide might require different scalings to correct for the effect of the same laser and sensor settings.

To determine whether the modeling approach we propose results in estimators of gene expression with good statistical properties, we ran a simulation study and assessed bias and root mean squared error of the estimators over repeated sampling. The simulation experiment is described in some detail in Section 4. Using simulated gene expression data, we applied several of the approaches (including the approach proposed here) to estimate gene expression for 1,000 genes and compared the various methods on the basis of bias, RMSE and dynamic range of the estimates that were obtained. The hierarchical modeling approach we propose had smaller bias and smaller RMSE than all other estimators, suggesting that basing estimation on as many readings for each spot as might be available is probably a reasonable idea. As mentioned earlier, genes with very high and with very low true expression levels were subject to the larger biases; this is to be expected since these are the genes that are likely to have censored expression measurements under very high or very low laser and sensor settings. Thus, changes in instrument settings cause not only a shift but also a censoring of the expression measurements in those genes.

The gamma-gamma (GG) model is not the only plausible model for gene expression

data. A log-normal-normal (LNN) model has been proposed as an alternative. To assess the dependence of our results on the model choice, we implemented our method with the LNN model for the expression values.

In the LNN model, we have a log-normal likelihood function for the data as follows:

$$\begin{aligned}
S_{i1} * \delta_1 = S'_{i1} &\sim \log \text{Norm}(\mu_i, \sigma^2) \\
S_{i2} * \delta_2 = S'_{i2} &\sim \log \text{Norm}(\mu_i, \sigma^2) \\
&\vdots \\
S_{im} * \delta_m = S'_{im} &\sim \log \text{Norm}(\mu_i, \sigma^2) \\
S_{i(m+1)} &\sim \log \text{Norm}(\mu_i, \sigma^2)
\end{aligned} \tag{3.9}$$

where $\delta_1, \dots, \delta_m$ are constant for all genes in a given slide and dye combination and have the same interpretation as in the original model. The unknown parameters in this model are σ^2 , $\mu_1, \dots, \mu_n$, and $\delta_1, \dots, \delta_m$.

As in the GG model, we place independent priors on the scaling parameters, δ , and the expression means, μ . Using log-normal and normal distributions, respectively, gives us a semi-conjugate prior in this case. Thus

$$\delta_j \sim \log \text{Norm}(\delta_0, \kappa^2)$$

for $j = 1, \dots, m$ where we specify $\delta_0 = 0$ and $\kappa^2 = 100$ and

$$\mu_i \sim \text{Norm}(\mu_0, \tau^2)$$

for $i = 1, \dots, n$. The hyperparameters in the model are $\eta = (\mu_0, \tau^2, \sigma^2)$. We choose conjugate (normal and inverse gamma) distributions for the hyperparameters.

To compare the method proposed in this work using the two hierarchical models, we performed a simulation study. As described in Section 3.5, we simulated datasets

Data Generation	Model	Average Absolute Bias	Average RMSE	Average Bias of Range
GG	GG	256.68	499.58	-184208
GG	LNN	375.14	544.77	-222013
LNN	GG	79.47	244.90	1137
LNN	LNN	28.44	212.54	-3557

Table 3.4 Comparing hierarchical models. Simulation results for 1000 genes after 100 simulations.

with three readings from the GG model. The expression values of each observation were estimated as before using our method with the GG model and average bias, RMSE and range were computed for 100 simulations. Additionally, the same simulations were estimated using our method with the LNN model. Another simulation study was carried out with datasets generated from the LNN model and both implementations of our method (GG and LNN) were used to estimate the expression values from those data. Table 3.4 compares the results of these simulations. Note that with data generated under the GG model, both implementations of the method proposed in this work are better than the alternatives discussed here (see Table 3.2).

Our method has an obvious non-model-based corollary. The scaling factors, $\delta_1, \dots, \delta_m$, can be estimated by the ratios of the medians of each scan in the following way:

$$\begin{aligned}
\hat{\delta}_1 &= M_{(m+1)}/M_{(1)} \\
&\vdots \\
\hat{\delta}_m &= M_{(m+1)}/M_{(m)}
\end{aligned}$$

where $M_{(1)}, \dots, M_{(m+1)}$ are the ordered scan median values. Then, the expression of gene i can be estimated as the scaled mean of the $m+1$ scans, $\frac{1}{m+1} \sum_{j=1}^{m+1} S_{ij} \hat{\delta}_j$ where $\hat{\delta}_{m+1} \equiv 1$ for notational convenience. Once again, censored values are recorded as missing

data and not included in the estimates. This method performs almost identically to point estimates from the hierarchical model in simulation. The scaled mean estimates, however, do not provide a measure of the uncertainty around the δ 's nor do they result in posterior distributions of gene expression.

While promising, conclusions drawn from the simulation experiment may be overly optimistic. Since the model used to generate the data is similar to that used for analyzing the data, biases and uncertainties in the estimates that may result from actually fitting the wrong model cannot be assessed. Thus, data obtained through simulation, while often quite informative, must be cautiously interpreted. We have partially explored this issue by applying our model to data generated from a log normal-normal model (see Table 3.4) with promising results.

We implemented the proposed hierarchical modeling approach on a set of slides obtained in a maize embryogenesis experiment carried out by scientists in the Plant Sciences Institute at Iowa State University (Che et al., 2004). While we present only a subset of the results here, this serves to highlight some of the improvements that appear to be associated to the use of the three scans available for each slide. When comparing our results to those obtained from fitting the Newton (2001) hierarchical model using only one reading per slide, we note that the variance of expression estimates is lower when based on three readings, as would be expected. We also notice that expression levels are not as shrunken toward the mean expression (2594). Because of the smaller bias and RMSE in gene expression estimates, inferences about the set of genes involved in somatic embryogenesis in maize change drastically when statistical analyses are based on one or on three readings of each slide. As might be expected, the power of tests increases as the RMSE in gene expression estimates decreases which in turn results in more precise time and biological line pool effects. As Skibbe, Nettleton, and Schnable

(2004) pointed out, conclusions drawn about differential expression can be dependent on the slide scan used. Here we see that stronger conclusions are possible using all available scans than using only one.

3.8 References

Armstrong, C.L. and Green, C.E. (1985). Establishment and maintenance of friable embryogenic maize callus and the involvement of L-proline. *Planta* 164, 207–214

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B: Methodological* 57, 289–300

Carlin, B.P. and Louis, T. (1997). *Bayes and Empirical Bayes* Chapman & Hall, London

Che, P., Love, T.M., Frame, B.R., Wang, K., Carriquiry, A.L., and Howell, S.H. (2004). Gene expression program during maturation and germination of somatic embryos in maize cultures. submitted to *Plant Molecular Biology*.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, Methodological* 39, 1–22.

Dudley, A.M., Aach, J., Steffen, M.A., and Church, G.M. (2002). Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *PNAS* 99(11), 7554–7559

Garcia de la Nava, J., van Hijum, S., and Trelles, O. (2004). Saturation and Quantization Reduction in Microarray Experiments using Two Scans at Different Sensitivities. *Statistical Applications in Genetics and Molecular Biology* 3(1), article 11.

- Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis* Chapman & Hall, London
- Green, C.E., and Phillips, R.L. (1975). Plant regeneration from tissue cultures of maize. *Crop Science* 15, 417–420
- Kerr, M.K., and Churchill, G.A. (2001). Experimental Design for Gene Expression Microarrays. *Biostatistics* 2, 183–201
- Lyng, H., Badiie, A., Svendsrud, D.H., Hovig, E., Myklebost, O., and Stokke, T. (2004). Profound influence of microarray scanner characteristics on gene expression ratios: analysis and procedure for correction. *BMC Genomics* 5:10.
- Newton, M.A., Kendzierski, C.M., Richmond, C.S., Blattner, F.R., and Tsui, K.W. (2001). On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data. *Journal of Computational Biology*, 8(1), 37–52.
- Romualdi, C., Trevisan, S., Celegato, B., Costa, G., and Lanfranchi, G. (2003). Improved detection of differentially expressed genes in microarray experiments through multiple scanning and image integration. *Nucleic Acids Research* 31(23) e149
- Skibbe, D.S., Nettleton, D., and Schnable, P.S. (2004). Scanning microarrays at multiple intensities increases the number of statistically significant differences detected. *Unpublished Manuscript*
- Smyth, G.K., Yang, Y.H., and Speed, T. (2002). Statistical Issues in cDNA Microarray Data Analysis. *Functional Genomics: Methods and Protocols* Humana Press, Totowa, NJ.

3.9 Appendix

3.9.1 Marginal likelihood of (a, a_0, ν)

Consider the joint likelihood function $p(a, a_0, \nu, \theta, \delta | S)$. The marginal likelihood function $p(a, a_0, \nu | S)$ is obtained by integrating the likelihood function with respect to δ and θ as follows:

$$p(a, a_0, \nu | S) = \int \int p(a, a_0, \nu, \theta, \delta | S) d\delta d\theta \quad (3.10)$$

$$\propto \int \int \Gamma(a)^{-n(m+1)} \prod_{i=1}^n \theta_i^{a(m+1)+a_0-1} \prod_{j=1}^m \delta_j^{na+\alpha_1-1} \prod_{i=1}^n \prod_{j=1}^{m+1} S_{ij}^{a-1} e^{-\sum_{i=1}^n \theta_i \sum_{j=1}^{m+1} \delta_j S_{ij} - \nu \sum_{i=1}^n \theta_i - \alpha_2 \sum_{j=1}^{m+1} \delta_j} d\delta d\theta \quad (3.11)$$

$$= \Gamma(a)^{-n(m+1)} \prod_{i=1}^n \prod_{j=1}^{m+1} S_{ij}^{a-1} \int \prod_{i=1}^n \theta_i^{a(m+1)+a_0-1} e^{-\nu \sum_{i=1}^n \theta_i} \int \prod_{j=1}^m \delta_j^{na+\alpha_1-1} e^{-\sum_{j=1}^{m+1} \delta_j (\sum_{i=1}^n \theta_i S_{ij} - \alpha_2)} d\delta d\theta \quad (3.12)$$

$$= \Gamma(a)^{-n(m+1)} \prod_{i=1}^n \prod_{j=1}^{m+1} S_{ij}^{a-1} \int \prod_{i=1}^n \theta_i^{a(m+1)+a_0-1} e^{-\nu \sum_{i=1}^n \theta_i} \Gamma(na + \alpha_1)^m \prod_{j=1}^m (\sum_{i=1}^n \theta_i S_{ij} - \alpha_2)^{-(na+\alpha_1)} d\theta \quad (3.13)$$

$$= \Gamma(a)^{-n(m+1)} \Gamma(na + \alpha_1)^m \prod_{i=1}^n \prod_{j=1}^{m+1} S_{ij}^{a-1} \int \prod_{i=1}^n \theta_i^{a(m+1)+a_0-1} e^{-\nu \sum_{i=1}^n \theta_i} \prod_{j=1}^m (\sum_{i=1}^n \theta_i S_{ij} - \alpha_2)^{-(na+\alpha_1)} d\theta. \quad (3.14)$$

3.9.2 Marginal likelihood of (a, a_0) conditional only on largest reading

We can derive the marginal distribution of the hyperparameters (a, a_0) conditional on the highest gene expression reading $S_{\bullet(m+1)}$ as follows:

$$p(a, a_0 | S_{\bullet(m+1)}) = \int \int \int p(a, a_0, \theta, \delta, \nu | S_{\bullet(m+1)}) d\delta d\theta d\nu \quad (3.15)$$

$$\propto \int \int \int p(S_{\bullet(m+1)}|a, a_0, \theta, \nu) p(\theta|a_0, \nu) p(\delta) p(\nu) d\delta d\theta d\nu \quad (3.16)$$

$$= \int \int p(S_{\bullet(m+1)}|a, a_0, \theta, \nu) p(\theta|a_0, \nu) p(\nu) d\theta d\nu \quad (3.17)$$

$$= \int \int \Gamma(a)^{-n} \prod_{i=1}^n (\theta_i^a S_{i(m+1)}^{a-1} \exp(-\theta_i S_{i(m+1)})) \\ \Gamma(a_0)^{-n} \nu^{na_0} \prod_{i=1}^n (\theta_i^{a_0-1} \exp(-\nu \theta_i)) \\ \frac{\beta_2^{\beta_1}}{\Gamma(\beta_1)} \nu^{\beta_1-1} \exp(-\beta_2 \nu) d\theta d\nu \quad (3.18)$$

$$\propto (\Gamma(a)\Gamma(a_0))^{-n} \left(\prod_{i=1}^n S_{i(m+1)} \right)^{a-1} \int \nu^{na_0+\beta_1-1} \exp(-\beta_2 \nu) \\ \int \left(\prod_{i=1}^n \theta_i \right)^{a+a_0-1} \exp \left(- \sum_{i=1}^n \theta_i (S_{i(m+1)} + \nu) \right) d\theta d\nu \quad (3.19)$$

$$= (\Gamma(a)\Gamma(a_0))^{-n} \left(\prod_{i=1}^n S_{i(m+1)} \right)^{a-1} \int \nu^{na_0+\beta_1-1} \exp(-\beta_2 \nu) \\ \prod_{i=1}^n (\Gamma(a+a_0)(S_{i(m+1)} + \nu)^{-a-a_0}) d\nu \quad (3.20)$$

$$= \left(\frac{\Gamma(a+a_0)}{\Gamma(a)\Gamma(a_0)} \right)^n \left(\prod_{i=1}^n S_{i(m+1)} \right)^{a-1} \\ \int \nu^{na_0+\beta_1-1} \exp(-\beta_2 \nu) \prod_{i=1}^n (S_{i(m+1)} + \nu)^{-a-a_0} d\nu \quad (3.21)$$

3.9.3 Expectation-Maximization to estimate (a, a_0)

We wish to maximize expression (3.6) with respect to (a, a_0) using the E-M algorithm.

Below we derive expressions to be used in the E- and the M-steps in the algorithm.

E-step:

$$E(\nu|a, a_0, S_{\bullet(m+1)}) = \int \nu p(\nu|a, a_0, S_{\bullet(m+1)}) d\nu \quad (3.22)$$

$$= \int \nu \int p(\theta, \nu|a, a_0, S_{\bullet(m+1)}) d\theta d\nu \quad (3.23)$$

$$= C \int \nu \int p(S_{\bullet(m+1)}, \theta, \nu|a, a_0) d\theta d\nu \quad (3.24)$$

$$= C \frac{\left(\prod_{i=1}^n S_{i(m+1)}\right)^{a-1} \beta_2^{\beta_1} \Gamma(a+a_0)^n}{\Gamma(\beta_1) \Gamma(a)^n \Gamma(a_0)^n} \int \nu \nu^{na_0+\beta_1-1} \exp(-\beta_2 \nu) \prod_{i=1}^n (S_{i(m+1)} + \nu)^{-(a+a_0)} d\nu \quad (3.25)$$

$$= \frac{\left(\prod_{i=1}^n S_{i(m+1)}\right)^{a-1} \beta_2^{\beta_1} \Gamma(a+a_0)^n}{\Gamma(\beta_1) \Gamma(a)^n \Gamma(a_0)^n} \times \frac{\int \nu^{na_0+\beta_1} \exp(-\beta_2 \nu) \prod_{i=1}^n (S_{i(m+1)} + \nu)^{-(a+a_0)} d\nu}{\int \nu^{na_0+\beta_1-1} \exp(-\beta_2 \nu) \prod_{i=1}^n (S_{i(m+1)} + \nu)^{-(a+a_0)} d\nu} \quad (3.26)$$

$$= \frac{\int \nu^{na_0+\beta_1} \exp(-\beta_2 \nu) \prod_{i=1}^n (S_{i(m+1)} + \nu)^{-(a+a_0)} d\nu}{\int \nu^{na_0+\beta_1-1} \exp(-\beta_2 \nu) \prod_{i=1}^n (S_{i(m+1)} + \nu)^{-(a+a_0)} d\nu} \quad (3.27)$$

M-Step:

$$L(a, a_0 | \nu, S_{\bullet(m+1)}) = \int \int p(a, a_0, \theta, \delta | \nu, S_{\bullet(m+1)}) d\delta d\theta \quad (3.28)$$

$$\propto \int \int p(S_{\bullet(m+1)} | a, a_0, \nu, \theta) p(\theta | a_0, \nu) p(\delta) d\delta d\theta \quad (3.29)$$

$$= \int p(S_{\bullet(m+1)} | a, \theta) p(\theta | a_0, \nu) d\theta \quad (3.30)$$

$$= \int \Gamma(a)^{-n} \prod_{i=1}^n \left(\theta_i^a S_{i(m+1)}^{a-1} \exp(-\theta_i S_{i(m+1)}) \right) \Gamma(a_0)^{-n} \nu^{na_0} \prod_{i=1}^n \left(\theta_i^{a_0-1} \exp(-\nu \theta_i) \right) d\theta \quad (3.31)$$

$$= (\Gamma(a) \Gamma(a_0))^{-n} \left(\prod_{i=1}^n S_{i(m+1)} \right)^{a-1} \nu^{na_0} \int \left(\prod_{i=1}^n \theta_i \right)^{a+a_0-1} \exp \left(- \sum_{i=1}^n \theta_i (S_{i(m+1)} + \nu) \right) d\theta \quad (3.32)$$

$$= (\Gamma(a) \Gamma(a_0))^{-n} \left(\prod_{i=1}^n S_{i(m+1)} \right)^{a-1} \nu^{na_0} \prod_{i=1}^n \left(\Gamma(a+a_0) (S_{i(m+1)} + \nu)^{-(a+a_0)} \right) \quad (3.33)$$

$$= \left(\frac{\Gamma(a+a_0)}{\Gamma(a) \Gamma(a_0)} \right)^n \nu^{na_0} \left(\prod_{i=1}^n S_{i(m+1)} \right)^{a-1} \prod_{i=1}^n (S_{i(m+1)} + \nu)^{-(a+a_0)} \quad (3.34)$$

CHAPTER 4. CLUSTERING POSTERIOR DISTRIBUTIONS - APPLICATION TO SOMATIC EMBRYOGENESIS IN MAIZE

Tanzy M. Love^{1,2,3}, Alicia L. Carriquiry¹

4.1 Introduction

In large time series or multiple treatment microarray experiments, we are often interested in locating groups of genes which react together. Subject matter theory designates these groups of genes as coregulated by the same biological pathways. For example, genes responsible for photosynthetic processes may express together in an experiment covering time periods during which plant tissues are exposed to light and darkness. The statistical problem is then clustering genes based on their expression values over multiple treatments.

We do not observe true gene expression, however. Instead, we observe a single or perhaps a replicated noisy estimate of gene expression under each treatment condition. There has been extensive work on methods for clustering genes based on gene expression reported in the literature (e.g., Yeung et al. 2001, Lazzeroni and Owen 2002, Liang and Kachalo 2002, Bryan et al. 2002, and Goldstein et al. 2002). In most cases, the clustering

¹Department of Statistics, Iowa State University

²Primary researcher and author

³Author for correspondence, supported by National Science Foundation grant number DMS 0091953

algorithm was implemented using the estimated expression levels without accounting for the uncertainty about the true levels of gene expression. In other cases, simplifying assumptions such as assuming that expression data are normal on the log-scale are used. In our own work, we have used the mean of the gene expression estimates to cluster genes (Che et al. 2004). In a later manuscript (Love and Carriquiry 2004) we proposed a hierarchical modeling approach to estimating the posterior distribution of gene expression when multiple readings (or scans) are available for each gene under each treatment condition. By determining a distribution of likely values for expression for each gene, we explicitly recognize that true expression is unobservable. The method proposed in Love and Carriquiry (2004) provides posterior probability distributions for expression value. Extensions of that model generate posterior distributions for expression ratios for pairs or sets of treatments. We can also construct the joint posterior probability distribution of all expression values or all pairwise expression ratios for each gene.

Given that we only know expression up to a distribution that reflects the range of likely expression values for each gene, how should we go about constructing groups of 'similar' genes? Standard clustering techniques are not designed to group units on the basis of distributions. Heuristic clustering (k-means, e.g.) requires the use of a distance metric between the items being clustered. However, measuring the distance between two probability distributions is not straight forward. Further, calculation of distance between distributions becomes more computationally difficult if we are using joint posterior distributions because of their multidimensionality. There are many possible metrics to measure the distance between two distributions (Devroye 1987), but not all distance measures are suitable for clustering units into similar groups. For example, the Kullback-Leibler divergence does not satisfy symmetry conditions (Kullback and Leibler 1951). However, Jeffreys (1946) had earlier proposed a divergence (the sum of the two Kullback-Leibler divergences) which is symmetric.

One other shortcoming of standard clustering methods is that they do not provide a means for estimating the number of clusters in a sample. There are several ad hoc methods for choosing a number of clusters from the data, but they have no probabilistic justification. Yet in almost all applications, estimating the number of groups in the data is an important aspect of the analysis. While we do not address the issue here, we discuss a possible extension of our method in the Conclusion section to allow estimation of the number of clusters.

Model-based clustering (Banfield and Raftery 1993) solves some of these problems for multidimensional data. It creates a probabilistic model framework in which the most likely number of clusters can be determined using standard statistics for model choice. The probabilities of inclusion in each component are calculated. A mixture of these components (probability distributions) is assumed to generate the data and the mixing coefficients and parameters of the component distributions are estimated. Two challenges in applying this method to our problem are the following:

- The standard model-based clustering algorithm relies on the assumption of multivariate normality. Even if we could observe true gene expression levels, we would be unable to assume multivariate normality of the observations. As presented by Fraley and Raftery (1999) the population is assumed to be composed of a finite mixture of multivariate normal groups. Results presented in the literature (e.g. Richmond et al. 1999 and Smyth et al. 2002) and our own empirical results (Love and Carriquiry 2004) suggest that gene expression is far from normal.
- Since we do not observe expression levels, the objects to be clustered are themselves probability distributions. Describing the population of probability distributions as a mixture requires some very careful thought.

Also of concern is defining the variables on the basis of which to cluster the units.

Gene expression is a relative measurement; changes of a similar direction and fold size are more meaningfully grouped together than genes of similar expression magnitude. For example, all genes with nearly constant expression over time belong together regardless of the size of the expression. Therefore, the expression ratios are a natural candidate for clustering (Che et al. 2004). However, these are pairwise values and, except in time course experiments where sequential ratios are natural, it is not obvious which ratios to pick. Depending on the design of the experiment, the pairwise differences can be estimated with differing precision (Kerr and Churchill 2001). Note that the distributions of ratios are less apt to fit normal distributions than mean expression values.

The remainder of this paper is organized as follows. In the next section, we outline the proposed clustering methods to be developed and implemented. In Section 4.3 we describe the hierarchical model that can be implemented to obtain posterior distributions of expression levels for each gene. The model is based on the approach proposed by Newton et al. (2001) and is also described in Love and Carriquiry (2004). We then discuss the problem of clustering genes into groups with similar expression values. One simple approach is to summarize the posterior distributions of gene expression into one or more parameters that can then be used as the variables on the basis of which to group the genes. Note that in this case, we are back in the situation in which we wish to cluster units on the basis of one or more variables that we observe with error. In Sections 4.3.2 and 4.3.3, we discuss some alternative methods for clustering the posterior distributions of model parameters when there are two treatments. Finally, Section 4.4 briefly describes the generalization of these methods to experiments where the design includes m treatments. We do not include background information about the analysis of gene expression data; that information can be found in the manuscript by Love and Carriquiry (2004). Section 4.5 compares the techniques using data from an experiment in maize embryogenesis.

4.2 Proposed clustering methods

We first develop an empirical model-based clustering algorithm that can be implemented using posterior simulation. For each gene, we can obtain a posterior distribution of expression by applying the modeling approach described later in this proposal. Consider the following algorithm:

- Draw a value $\rho_i^{(j)}$ for the i th gene from $p_i(\rho|y_i)$, where $j = 1, \dots, M$ for large M and for $i = 1, \dots, n$. Here, ρ_i denotes the parameter of interest for clustering genes. We develop a model in which ρ_i is the ratio of expression between treatments for the i th gene.
- Given the n expression values, implement a model-based clustering algorithm to group genes into $C^{(j)}$ groups. We postulate mixtures of non-normal components.

For each sample, the most likely number of clusters can be identified and the most likely cluster membership for each gene can be estimated. Summarizing these results is challenging because of the complexity of the information generated. Note that this is a varying-dimensional parameter space problem, in that the number of clusters may well vary across samples from the posterior distributions of gene expressions. Thus, while the i th gene may be allocated to a cluster with the same label at every iteration, it is not clear that the cluster label has any meaning as the number of clusters is unlikely to be constant across iterations. A posterior distribution of the optimal (in some sense) number of clusters can be generated and has a straightforward interpretation, but all other parameters (including cluster membership) that depend on the number of clusters at each iteration are not easily interpretable. A matrix of pairwise probabilities of joint cluster membership for each pair of genes can be generated by simulation and is one means of summarizing results. Ideally, the $n - 1$ elements in the i th matrix row (not including the (i, i) element) will include some very small values and some large values,

indicating that over the M draws, the i th gene is almost always grouped with the same subset of the genes.

Further, we develop analytical results for clustering posterior distributions in the two treatment case; the actual implementation requires some numerical approximation. These posterior distributions can be analytically expressed in the two treatment case which allows us to derive analytic expressions for the distance between two posterior distributions. However, the posterior distributions can only be expressed up to a normalizing constant in the general m treatment case. Therefore, we use the empirical posterior distributions generated by MCMC methods as in the earlier approach to approximate the true posterior distributions. The distance between the posterior distributions of relative gene expression for two genes can be calculated with a distance metric for probability distributions and used to cluster the genes. The method used to cluster is Partitioning Around Medoids (PAM), a more robust version of K-means (Kaufman and Rousseeuw 1990). After finding a set of C medoids, C clusters are constructed by assigning each observation to the nearest medoid. The goal is to find C representative objects (genes here) which minimize the sum of the dissimilarities of the observations to their closest representative object (medoid gene). It searches for these objects (medoids) iteratively by first selecting medoids, then recalculating cluster membership and continuing these steps until the minimum sum of the dissimilarities of the observations (genes) to their closest medoid is found. The product is a method for clustering multivariate, non-normal distributions. The algorithm generates groups of random variables that have probability distributions that are 'close'.

4.3 Model and posterior clustering for two treatments

We describe a hierarchical model for comparing two treatments measured with R replications. The case of an experiment containing a single cDNA slide with two samples is a special case of this model with $R = 1$. Let there be n genes spotted on a batch of slides. For a given gene i , we use X_{ij} and Y_{ij} to denote the gene expression measurements from the j^{th} replicate of the two treatments. We assume that these values have undergone background correction and normalization so that we can assume that the replicates are exchangeable, given the treatment. While the functional form of the posterior distribution of the expression ratios will be different under a different model, the methods for generating and clustering the empirical posteriors will be the same.

4.3.1 Bayesian Hierarchical Gamma Model

In order to estimate gene expression, we use a Bayesian gamma-gamma hierarchical model (Newton et al. 2001).

$$X_{ij} \sim \Gamma(a, \theta_{Xi}) \quad (4.1)$$

$$Y_{ij} \sim \Gamma(a, \theta_{Yi}) \quad (4.2)$$

for $i = 1, \dots, n$ and $j = 1, \dots, R$ where a is constant for all genes on the slide. This assumption is based on a constant coefficient of variation argument.

First, to simplify computations, we introduce the variable z_i , that indicates whether a gene is differentially expressed. For the i th gene, $z_i = 1$ if the gene is differentially expressed and $z_i = 0$ if the gene is not differentially expressed. Therefore, z_i is a Bernoulli random variable with parameter p ; $p(z_i|p) = p^{z_i}(1-p)^{1-z_i}$. If gene i is differentially expressed ($z_i = 1$), $\theta_{Xi} \neq \theta_{Yi}$; otherwise, $\theta_{Xi} = \theta_{Yi}$ ($z_i = 0$). In the remainder of this section, we focus on an individual gene and omit the subscript i to simplify notation and let $X = (X_1, \dots, X_R)$ be the vector of R replicated measure of the gene's expression

under the first treatment. Since we are modeling X and Y as arising from gamma distributions, the joint likelihood function for the observed intensities X and Y when there is differential expression is given by

$$p(X, Y | \theta_X, \theta_Y, a, z = 1) = \prod_{j=1}^R \Gamma(a)^{-2} (\theta_X \theta_Y)^a (X_j Y_j)^{a-1} \exp(-X_j \theta_X - Y_j \theta_Y). \quad (4.3)$$

We choose conjugate gamma prior distributions for θ_X and θ_Y . Their joint prior distribution, given differential expression and hyperparameters a_0 and ν is

$$p(\theta_X, \theta_Y | a_0, \nu, z = 1) = \Gamma(a_0)^{-2} \nu^{2a_0} (\theta_X \theta_Y)^{a_0-1} \exp(-\nu(\theta_X + \theta_Y)). \quad (4.4)$$

The posterior distribution of θ_X and θ_Y when there is differential expression, given the observed intensities and the hyperparameters, is proportional to the product of the likelihood and the prior distributions:

$$p(\theta_X, \theta_Y | X, Y, a, a_0, \nu, z = 1) \propto p(X, Y | \theta_X, \theta_Y, a, z = 1) p(\theta_X, \theta_Y | a_0, \nu, z = 1). \quad (4.5)$$

Substituting the distributions from above and dropping all of the terms which are constant with respect to θ_X and θ_Y , we obtain

$$p(\theta_X, \theta_Y | X, Y, a, a_0, \nu, z = 1) \propto (\theta_X \theta_Y)^{Ra+a_0-1} \exp(-(\sum_{j=1}^R X_j + \nu)\theta_X - (\sum_{j=1}^R Y_j + \nu)\theta_Y), \quad (4.6)$$

which is the product of a $\Gamma(Ra+a_0, \sum_{j=1}^R X_j + \nu)$ distribution on θ_X and an independent $\Gamma(Ra+a_0, \sum_{j=1}^R Y_j + \nu)$ distribution on θ_Y .

Let $\rho = \theta_Y / \theta_X$ and note that

$$\rho = \frac{\theta_Y}{\theta_X} = \frac{a}{\theta_X} \frac{\theta_Y}{a} = \frac{E(X)}{E(Y)}.$$

We reparameterize $p(\theta_X, \theta_Y | X, Y, a, a_0, \nu, z = 1)$ in terms of ρ as follows, where the last term is the determinant of the Jacobian of the inverse transformation.

$$p(\theta_X, \rho | X, Y, a, a_0, \nu, z = 1)$$

$$\begin{aligned}
&= p(\theta_X, \rho \theta_X | X, Y, a, a_0, \nu, z = 1) \begin{vmatrix} 1 & 0 \\ 0 & \theta_X \end{vmatrix} \\
&\propto (\theta_X \rho \theta_X)^{Ra+a_0-1} \exp\left(-\left(\sum_{j=1}^R X_j + \nu\right)\theta_X - \left(\sum_{j=1}^R Y_j + \nu\right)\rho\theta_X\right) \\
&\propto \rho^{Ra+a_0-1} \theta_X^{2Ra+2a_0-1} \exp\left(-\theta_X\left(\sum_{j=1}^R X_j + \nu + \left(\sum_{j=1}^R Y_j + \nu\right)\rho\right)\right).
\end{aligned}$$

The marginal posterior distribution of ρ can then be obtained by integrating out θ_X from the joint posterior distribution:

$$\begin{aligned}
&p(\rho | X, Y, a, a_0, \nu, z = 1) \\
&\propto \rho^{Ra+a_0-1} \int_0^\infty \theta_X^{2Ra+2a_0-1} \exp\left(-\theta_X\left(\sum_{j=1}^R X_j + \nu + \left(\sum_{j=1}^R Y_j + \nu\right)\rho\right)\right) d\theta_X \\
&= \rho^{Ra+a_0-1} \Gamma(2Ra + 2a_0) \left(\sum_{j=1}^R X_j + \nu + \left(\sum_{j=1}^R Y_j + \nu\right)\rho\right)^{-2(Ra+a_0)} \\
&\propto \rho^{Ra+a_0-1} \left(1 + \rho \frac{Y + \nu}{X + \nu}\right)^{-2(Ra+a_0)}. \tag{4.7}
\end{aligned}$$

Thus, under the assumption of differential expression for the i th gene, the marginal posterior distribution of ρ conditional on the observed intensities and on the values of the hyperparameters is given by

$$p(\rho | X, Y, a, a_0, \nu, z = 1) = \frac{\Gamma(2Ra + 2a_0)}{\Gamma(Ra + a_0)^2} \left(\frac{Y + \nu}{X + \nu}\right)^{Ra+a_0} \rho^{Ra+a_0-1} \left(1 + \rho \frac{Y + \nu}{X + \nu}\right)^{-2(Ra+a_0)}. \tag{4.8}$$

If the gene is not differentially expressed, $z = 0$ and $\rho \equiv 1$. The unconditional posterior distribution of ρ is then

$$\begin{aligned}
p(\rho | X, Y, a, a_0, \nu, p) &= p(\rho | X, Y, a, a_0, \nu, p, z = 1) p(z = 1 | X, Y, a, a_0, \nu, p) \\
&\quad + I_{\{1\}}(\rho) p(z = 0 | X, Y, a, a_0, \nu, p),
\end{aligned}$$

where $I_{\{1\}}(\cdot)$ denotes the indicator function of the value one.

We now require the posterior distribution of z . We use the fact that $p(z|X, Y, a, a_0, \nu, p)$ can be obtained as the ratio of the joint conditional distribution $p(z, X, Y|a, a_0, \nu, p)$ to the distribution of (X, Y) conditional on (a, a_0, ν, p) (full derivation is given in the appendix):

$$\begin{aligned}
p(z|X, Y, a, a_0, \nu, p) &= \frac{p(z, X, Y|a, a_0, \nu, p)}{p(X, Y|a, a_0, \nu, p)} \\
&\propto p(z, X, Y|a, a_0, \nu, p) \\
&= p^z(1-p)^{1-z} \left(\frac{z(\prod_{j=1}^R X_j Y_j)^{a-1} \nu^{2a_0}}{(\sum_{j=1}^R X_j + \nu)^{Ra+a_0} (\sum_{j=1}^R Y_j + \nu)^{Ra+a_0}} \frac{\Gamma(Ra + a_0)^2}{\Gamma(a)^{2R} \Gamma(a_0)^2} \right. \\
&\quad \left. + \frac{(1-z)(\prod_{j=1}^R X_j Y_j)^{a-1} \nu^{a_0}}{(\sum_{j=1}^R X_j + \sum_{j=1}^R Y_j + \nu)^{2Ra+a_0}} \frac{\Gamma(2Ra + a_0)}{\Gamma(a)^{2R} \Gamma(a_0)} \right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
p(z|X, Y, a, a_0, \nu, p) &= K p^z(1-p)^{1-z} \left(\frac{z(\prod_{j=1}^R X_j Y_j)^{a-1} \nu^{2a_0}}{(\sum_{j=1}^R X_j + \nu)^{Ra+a_0} (\sum_{j=1}^R Y_j + \nu)^{Ra+a_0}} \frac{\Gamma(Ra + a_0)^2}{\Gamma(a)^{2R} \Gamma(a_0)^2} \right. \\
&\quad \left. + \frac{(1-z)(\prod_{j=1}^R X_j Y_j)^{a-1} \nu^{a_0}}{(\sum_{j=1}^R X_j + \sum_{j=1}^R Y_j + \nu)^{2Ra+a_0}} \frac{\Gamma(2Ra + a_0)}{\Gamma(a)^{2R} \Gamma(a_0)} \right),
\end{aligned}$$

where K is the normalizing constant and K^{-1} is given by

$$\begin{aligned}
K^{-1} &= \frac{p(\prod_{j=1}^R X_j Y_j)^{a-1} \nu^{2a_0}}{(\sum_{j=1}^R X_j + \nu)^{Ra+a_0} (\sum_{j=1}^R Y_j + \nu)^{Ra+a_0}} \frac{\Gamma(Ra + a_0)^2}{\Gamma(a)^{2R} \Gamma(a_0)^2} \\
&\quad + \frac{(1-p)(\prod_{j=1}^R X_j Y_j)^{a-1} \nu^{a_0}}{(\sum_{j=1}^R X_j + \sum_{j=1}^R Y_j + \nu)^{2Ra+a_0}} \frac{\Gamma(2Ra + a_0)}{\Gamma(a)^{2R} \Gamma(a_0)} \\
&= \frac{(\prod_{j=1}^R X_j Y_j)^{a-1} \nu^{a_0}}{\Gamma(a)^{2R} \Gamma(a_0)} \\
&\quad \left(\frac{p \nu^{a_0} \Gamma(Ra + a_0)^2}{(\sum_{j=1}^R X_j + \nu)^{Ra+a_0} (\sum_{j=1}^R Y_j + \nu)^{Ra+a_0} \Gamma(a_0)} + \frac{(1-p) \Gamma(2Ra + a_0)}{(\sum_{j=1}^R X_j + \sum_{j=1}^R Y_j + \nu)^{2Ra+a_0}} \right).
\end{aligned}$$

We can now compute the probability that $z = 1$ (differential expression of the gene) as follows:

$$\begin{aligned}
& p(z = 1|X, Y, a, a_0, \nu, p) \\
&= \left(\frac{p\nu^{a_0}\Gamma(Ra + a_0)^2}{(\sum_{j=1}^R X_j + \nu)^{Ra+a_0}(\sum_{j=1}^R Y_j + \nu)^{Ra+a_0}\Gamma(a_0)} + \frac{(1-p)\Gamma(2Ra + a_0)}{(\sum_{j=1}^R X_j + \sum_{j=1}^R Y_j + \nu)^{2Ra+a_0}} \right)^{-1} \\
&\quad \times \left(\frac{(\prod_{j=1}^R X_j Y_j)^{a-1} \nu^{a_0}}{\Gamma(a)^{2R}\Gamma(a_0)} \right)^{-1} p \left(\frac{(\prod_{j=1}^R X_j Y_j)^{a-1} \nu^{2a_0}}{(\sum_{j=1}^R X_j + \nu)^{Ra+a_0}(\sum_{j=1}^R Y_j + \nu)^{Ra+a_0}\Gamma(a)^{2R}\Gamma(a_0)^2} \right) \\
&= \left(1 + \frac{(1-p)\Gamma(2Ra + a_0)}{(\sum_{j=1}^R X_j + \sum_{j=1}^R Y_j + \nu)^{2Ra+a_0}} \frac{(\sum_{j=1}^R X_j + \nu)^{Ra+a_0}(\sum_{j=1}^R Y_j + \nu)^{Ra+a_0}\Gamma(a_0)}{p\nu^{a_0}\Gamma(Ra + a_0)^2} \right)^{-1} \\
&= \left(1 + \frac{(1-p)(\sum_{j=1}^R X_j + \nu)^{Ra+a_0}(\sum_{j=1}^R Y_j + \nu)^{Ra+a_0}\Gamma(2Ra + a_0)\Gamma(a_0)}{p\nu^{a_0}(\sum_{j=1}^R X_j + \sum_{j=1}^R Y_j + \nu)^{2Ra+a_0}\Gamma(Ra + a_0)^2} \right)^{-1}.
\end{aligned}$$

One drawback to this model is that it does not account for the uncertainty in our estimate of the hyperparameters, (a, a_0, ν, p) . This can be resolved by assigning a hyperprior to them and then estimating the distributions of the hyperparameters jointly with the other posterior distributions. However, these values are genome-wide and are estimated using all n genes. Since n is large, we get a good estimate of these values using marginal maximum likelihood methods. Closed form expressions for the posterior distributions of ρ_i are not possible without the simplifying assumption of known hyperparameters. We can place independent gamma priors on a, a_0 , and ν with parameters λ, ψ , and β , respectively, and a beta prior on p with parameters α . All of these parameters can be specified *a priori*. We have not found evidence of sensitivity to these prior choices. We fix the estimated values of the hyperparameters to simplify the model and allow calculation of the full posterior distributions.

4.3.2 Clustering the ρ 's with standard methods

We propose two methods for clustering the n ρ_i 's using standard methodology. In the first, we cluster using the posterior means as observed values and include only those genes for which $\Pr(z_i = 1|X_i, Y_i) > 0.5$. We do this because the genes with $\Pr(z_i = 1|X_i, Y_i) \leq 0.5$ are most likely not differentially expressed. This means that we can group all of them into a single cluster with the null (not changing) expression

pattern. Secondly, we can use the same method to iteratively cluster samples from the joint posterior distribution of ρ . From these results, we draw inferences about the most likely set of clusters, the most likely number of clusters among the n genes and also obtain a measure of the uncertainty of the grouping of the genes into clusters.

To implement model-based clustering, we assume that the n ρ_i 's come from a mixture of C distributions, so that

$$\rho_i \sim \sum_{c=1}^C \tau_c f_c(\rho_i | \lambda_c)$$

where τ_c is the probability that ρ_i belongs to the c^{th} component of the mixture for $c = 1, \dots, C$.

An alternative to the gamma-gamma hierarchical model for gene expression values proposed in Section 4.3.1 is a log-normal-normal model. Under this model, $\rho = \frac{E(X)}{E(Y)} = \frac{\exp(\mu_X)}{\exp(\mu_Y)} = \exp(\mu_X - \mu_Y)$ where μ_X and μ_Y have normal posterior distributions. Therefore $\log(\rho) = \mu_X - \mu_Y$ will have a normal distribution when $z = 1$ and $\log(\rho) \equiv 0$ when $z = 0$. While the ρ_i generated from our model will not have a log-normal distribution, we believe that using a mixture of log-normals to cluster the ρ_i is appropriate because they may have distributions similar to a log-normal.

When using the posterior means, $\hat{\rho}_i$, we can cluster genes by first log transforming the means and then implementing model-based clustering with normal distributions (Fraley and Raftery 1999). We force all genes with posterior probability of differential expression, $\Pr(z_i | X_i, Y_i)$, less than 0.5 into one cluster and cluster the remaining genes using $\hat{\rho}_i$. In this approach, we fix the number of clusters at a range of pre-determined values, and for each possible number of clusters estimate the parameters of the component distributions. Note that by so doing, we do not need to implement reversible jump MCMC (RJMCMC) since conditional on a fixed number of clusters, the dimension of the parameter space

is also fixed. We use Bayes Information Criterion (BIC) to estimate the most likely number of clusters.

To use more of the information available in the posterior distributions than just the mean, we cluster using samples from the posterior distribution. We generate the samples using Markov chain Monte Carlo (MCMC) methods. For a particular sample ρ_b from the posterior distribution of ρ we can cluster the n ρ_{ib} . We have found that including those genes with $\rho_{ib} = 1$ results in one cluster around 1 and $C_b - 1$ clusters each of size one. To prevent this, we force all genes with $\rho_{ib} = 1$ into one cluster and use model-based clustering with log-normal distributions to cluster the remaining genes. For each sample b , we estimate the ‘best’ estimate of the number of clusters $\hat{C}_b$ using BIC. The clustering step is repeated B times, once for each of B samples from the posterior. For any two genes i and j , the proportion of samples in which they belong to the same cluster is approximately the posterior probability that they are in the same cluster.

We now have a set of B clusterings and the joint probabilities that any subset of the genes were clustered together in those clusterings. However, we need a single summary clustering and number of clusters. We calculate the pairwise proportion of samples in which two genes are in different clusters. This is a measure of the distance between the posterior distributions of the expression ratio for the two genes. We use this distance value to cluster the genes with PAM. The method is implemented on the maize embryogenesis data described earlier. Results are described in Section 4.5.

4.3.3 Clustering the posterior distributions of the ρ_i ’s

Here, we make use of the full posterior distribution of each ρ_i , $p(\rho_i|X_i, Y_i, a, a_0, \nu, p)$ for $i = 1, \dots, n$. Using fixed values of the hyperparameters, (a, a_0, ν, p) , we can calculate the exact form of the posterior distribution in the two treatment case,

$p(\rho_i|X, Y, \alpha, \beta, \lambda, \psi) \equiv p_i(\rho_i|\text{all})$. When we place hyperprior distributions on the hyperparameters (or in the general m treatments case), the functional form of the posterior distributions of ρ_i are analytically intractable. Therefore, we also approximate them with the empirical distribution, $p_B(\rho_i|X, Y, \alpha, \beta, \lambda, \psi) \equiv p_{Bi}(\rho_i)$ based on B samples from the posterior distribution generated by MCMC.

The distance between two distributions, e.g. $p_i(\rho|\text{all})$ and $p_j(\rho|\text{all})$ or $p_{Bi}(\rho)$ and $p_{Bj}(\rho)$, can be measured via a distance metric. For example

$$L_1(p_{Bi}, p_{Bj}) = \int |p_{Bi}(\rho) - p_{Bj}(\rho)| d\rho$$

is the L_1 distance between $p_{Bi}(\rho)$ and $p_{Bj}(\rho)$. All probability densities are in L_1 space, so $L_1(f, g) = \int |f - g|$ is always well-defined. Another measure for the difference between distributions, the Kullback-Leibler (KL) divergence given by

$$d_{KL}(p_{Bi}, p_{Bj}) = \int \log \left(\frac{p_{Bi}(\rho)}{p_{Bj}(\rho)} \right) p_{Bi}(\rho) d\rho,$$

is not symmetric ($d_{KL}(f, g) \neq d_{KL}(g, f)$). However, Jeffreys (1946) proposed

$$d_J(p_{Bi}, p_{Bj}) = d_{KL}(p_{Bi}, p_{Bj}) + d_{KL}(p_{Bj}, p_{Bi})$$

which is the J-divergence between $p_{Bi}(\rho)$ and $p_{Bj}(\rho)$. We use the average KL divergence (half of the J-divergence) between the two distributions. Using the average of $d_{KL}(f, g)$ and $d_{KL}(g, f)$ creates a symmetric measure of divergence which can be used as a distance metric in clustering. The Kullback-Leibler divergence is attractive here because of its invariance to all non-singular transformations of the variable.

We obtain two different sets of pairwise distances between the posterior distributions, one measured on the set $\{p_i(\cdot|\text{all})\}_{i=1, \dots, n}$ and the other on $\{p_{Bi}(\cdot)\}_{i=1, \dots, n}$. These pairwise distances are used to cluster the genes into C clusters using PAM. We pick C from

the number of clusters found to be optimal by the standard methods described in Section 4.3.2. The methods are illustrated in the maize embryogenesis example described in Section 4.5.

4.4 Model and posterior clustering for m treatments

Consider now the more realistic case where m treatments are hybridized with R replications. With cDNA microarray data, it is generally desirable to include an even number of replications, R of each treatment in the design, to be able to accommodate a dye-swap to address potential dye-gene interactions. [In a dye-swap design, a pair of treatments A and B are hybridized on one slide by dyeing A red and B green, and on a different slide by now dyeing A green and B red.] This is not necessary in single hybridization technology. In either case, the total number of observations for each gene in this experiment will be $m \times R$.

For a given gene i , we use E_{ijk} to denote the gene expression measurement from the k^{th} replication of the j^{th} treatment after background correction and normalization. We do not include subscripts for slide or dye because we assume that the appropriate normalization has removed these effects.

The model that we introduce below assumes that we have obtained the same number of observations per treatment. Therefore, the formulation below is not suitable if the experiment is carried out using cDNA microarrays and a reference design in which each treatment is hybridized with a control or reference sample on a slide. Such data can be transformed and expressed as expression relative to reference expression in order to fit this model. The loop design, with or without a dye swap, however, fits directly into our framework, and we recommend it (Kerr and Churchill 2001). Any design with the same number of replications per treatment generated by a single hybridization technology

would also fit this model. While we do not discuss extensions here, the model below can be adapted to include the case where an uneven number of replications per treatment is available.

4.4.1 Bayesian Hierarchical Gamma Model

In order to estimate gene expression, we propose a Bayesian hierarchical model. As before, the sampling distribution is

$$E_{ijk} \sim \Gamma(a, \theta_{ij}) \quad (4.9)$$

for $i = 1, \dots, n$, $j = 1, \dots, m$, and $k = 1, \dots, R$ where a is constant for all genes and θ_{ij} depends on gene and treatment. Here, we also assume that the replications of a treatment will be exchangeable after normalization.

If gene i is not differentially expressed across the different treatments, then $\theta_{i1} = \dots = \theta_{im}$. This might be the case, for example, of a gene associated to photosynthesis if the m treatments consist in varying the moisture to which a plant tissue is exposed without varying the light conditions. If the expression of gene i differs across the treatments, then the θ_i 's are not all equal. Again, to simplify computations, we introduce the variable z , that indicates whether each gene is differentially expressed.

There are many ways to create this variable in the case of multiple treatments. In the two treatment case, a gene has only two options: equal expression at both treatments, or different expression. In the general $m > 2$ treatment case, there are more than two possible configurations of expression equality. For the case with three treatments, there can be equal expression under all treatments, three distinct types of equal expression under two treatments and not the third, and finally expression can be different under each treatment for a total of five configurations. The total number of configurations is

the Bell number of the number of treatments, B_m for m treatments (Bell 1934). We discuss three of the most simple choices for z .

First, we may wish to differentiate between all B_m possible configurations. One way to do this is to assign the B_m configurations to the set of integers $\{1, \dots, B_m\}$ and let $z_{it} = 1$ if the expression pattern observes the t^{th} configuration and $z_{it} = 0$ otherwise for $t = 1, \dots, B_m$. We can model each z_i as a multinomial(1, p) random variable where $p = (p_1, \dots, p_{B_m})$ are the probabilities of the different configurations and $\sum_{t=1}^{B_m} p_t = 1$. This gives us the following distribution for z_i : $p(z_i|p) = \prod_{t=1}^{B_m} p_t^{z_{it}}$.

A simpler alternative is to only concern ourselves with the sequential pairwise differences. There are $m-1$ of these differences and we can represent their presence or absence by letting $z_{it} = 1$ if expression is different under treatment t than under treatment $t+1$ and $z_{it} = 0$ otherwise for $t = 1, \dots, m-1$. Now, we can model z_i as $m-1$ independent Bernoulli random variables with different parameters. Let p_t be the probability that $z_{it} = 1$ for $t = 1, \dots, m-1$ and $p = (p_1, \dots, p_{m-1})$. Note that $\sum_{t=1}^{m-1} p_t$ is not necessarily one and these four indicators do not describe all possible expression configurations. This gives us $p(z_i|p) = \prod_{t=1}^{m-1} p_t^{z_{it}} (1 - p_t)^{1-z_{it}}$ as the distribution of z_i .

Finally, we could use the same indicator that we used in the two treatment case. Here we only distinguish between the null pattern (no differential expression) and anything else. There are only two alternatives, regardless of the number of treatments. For the i th gene on a slide, $z_i = 1$ if the gene is differentially expressed and $z_i = 0$ if the gene has the same expression under all treatments. Therefore, z_i is again a Bernoulli random variable with parameter p ; $p(z_i|p) = p^{z_i} (1 - p)^{1-z_i}$.

In choosing between these three indicator variables, the purpose of our model drives the choice. We wish to create posterior distributions for the expression ratios and to

calculate the probability of differential expression for each gene. These goals are met by the simplest indicator proposed. It also greatly simplifies the calculation of the posterior probability of differential expression and the posterior distribution of the expression ratios. For these reasons, we will use the third alternative for our indicator variable. If gene i is not differentially expressed across all m treatments ($z_i = 0$), $\theta_{i1} = \dots = \theta_{im}$; otherwise they are not all equal ($z_i = 1$). In the remainder of this section, we omit the subscript i to simplify notation.

The joint likelihood function for the observed intensities E_{jk} when there is differential expression is given by

$$p(E|\Theta, a, z = 1) = \prod_{j=1}^m \prod_{k=1}^R \frac{(\theta_j)^a}{\Gamma(a)} (E_{jk})^{a-1} \exp(-E_{jk}\theta_j). \quad (4.10)$$

Placing independent conjugate gamma priors on the θ 's, the joint distribution of Θ given differential expression and the hyperparameters is

$$p(\Theta|a_0, \nu, z = 1) = \prod_{j=1}^m \frac{\nu^{a_0}}{\Gamma(a_0)} (\theta_j)^{a_0-1} \exp(-\nu\theta_j).$$

The joint posterior distribution of θ when there is differential expression, given the observed intensities and the hyperparameters, is proportional to the product of the likelihood and the prior distributions:

$$p(\Theta|E, a, a_0, \nu, z = 1) \propto p(E|\Theta, a, z = 1)p(\Theta|a_0, \nu, z = 1). \quad (4.11)$$

Substituting the distributions from above and dropping all of the terms which are constant with respect to Θ , we get

$$\begin{aligned} p(\Theta|E, a, a_0, \nu, z = 1) &\propto \prod_{j=1}^m \prod_{k=1}^R (\theta_j^a \exp(-E_{jk}\theta_j)) \prod_{j=1}^m (\theta_j^{a_0-1} \exp(-\nu\theta_j)) \\ &= \left(\prod_{j=1}^m \theta_j \right)^{aR} \left(\prod_{j=1}^m \theta_j \right)^{a_0-1} \exp \left(- \sum_{j=1}^m \sum_{k=1}^R E_{jk}\theta_j \right) \exp \left(- \sum_{j=1}^m \nu\theta_j \right) \\ &= \left(\prod_{j=1}^m \theta_j \right)^{aR+a_0-1} \exp \left(- \sum_{j=1}^m \theta_j \left(\sum_{k=1}^R E_{jk} + \nu \right) \right) \end{aligned} \quad (4.12)$$

which is the product of independent $\Gamma(aR + a_0, \sum_{k=1}^R E_{jk} + \nu)$ distributions on each θ_j .

Let $\Psi_j = \theta_{j+1}/\theta_j$ for $j = 1, \dots, m-1$. Note that

$$\Psi_j = \frac{E(E_{jk})}{E(E_{(j+1)k})} = \frac{a}{\theta_j} \times \frac{\theta_{j+1}}{a} = \frac{\theta_{j+1}}{\theta_j}$$

for all k and for $j = 1, \dots, m-1$. We reparameterize $p(\theta|E, a, a_0, \nu, z = 1)$ in terms of Ψ as follows, where the last term is the determinant of the Jacobian of the inverse transformation and $\eta = (a, a_0, \nu)$.

$$\begin{aligned} p(\Psi, \theta_m | E, \eta, z = 1) &= p(\theta_1, \theta_2, \dots, \theta_m | E, \eta, z = 1) \mathbf{J} \\ &= p(\theta_1, \Psi_1 \theta_1, \dots, \prod_{j=1}^{m-1} \Psi_j \theta_1 | E, \eta, z = 1) \\ &\quad \times \begin{vmatrix} 1 & 0 & 0 & \dots & 0 \\ \Psi_1 & \theta_1 & 0 & \dots & 0 \\ \prod_{j=1}^2 \Psi_j & \Psi_2 \theta_1 & \Psi_1 \theta_1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \prod_{j=1}^{m-1} \Psi_j & \prod_{j=2}^{m-1} \Psi_j \theta_1 & \prod_{j \neq 2} \Psi_j \theta_1 & \dots & \prod_{j=1}^{m-2} \Psi_j \theta_1 \end{vmatrix} \\ &\propto \left((\theta_1)^m \prod_{j=1}^{m-1} (\Psi_j)^{m-j} \right)^{Ra+a_0-1} (\theta_1)^{m-1} \prod_{j=1}^{m-1} (\Psi_j)^{m-1-j} \\ &\quad \times \exp \left(- \left(\sum_{k=1}^R E_{1k} + \nu \right) \theta_1 - \sum_{j=2}^m \left(\sum_{k=1}^R E_{jk} + \nu \right) \prod_{j'=1}^{j-1} \Psi_{j'} \theta_1 \right) \\ &\propto (\theta_1)^{m(Ra+a_0-1)+m-1} \prod_{j=1}^{m-1} (\Psi_j)^{(m-j)(Ra+a_0-1)+m-j-1} \\ &\quad \times \exp \left(-\theta_1 \left(\left(\sum_{k=1}^R E_{1k} + \nu \right) - \sum_{j=2}^m \left(\sum_{k=1}^R E_{jk} + \nu \right) \prod_{j'=1}^{j-1} \Psi_{j'} \right) \right). \end{aligned}$$

The marginal posterior distribution of Ψ can then be obtained by integrating out θ_1 from the joint posterior distribution as follows

$$\begin{aligned} p(\Psi | E, \eta, z = 1) &\propto \prod_{j=1}^{m-1} (\Psi_j)^{(m-j)(Ra+a_0-1)+m-j-1} \int (\theta_1)^{m(Ra+a_0-1)+m-1} \\ &\quad \times \exp \left(-\theta_1 \left(\left(\sum_{k=1}^R E_{1k} + \nu \right) - \sum_{j=2}^m \left(\sum_{k=1}^R E_{jk} + \nu \right) \prod_{j'=1}^{j-1} \Psi_{j'} \right) \right) d\theta_1 \end{aligned}$$

$$\begin{aligned}
&= \prod_{j=1}^{m-1} (\Psi_j)^{(m-j)(Ra+a_0-1)+m-j-1} \Gamma(m(Ra+a_0-1)+m) \\
&\quad \left(\left(\left(\sum_{k=1}^R E_{1k} + \nu \right) - \sum_{j=2}^m \left(\sum_{k=1}^R E_{jk} + \nu \right) \prod_{j'=1}^{j-1} \Psi_{j'} \right) \right)^{-(m(Ra+a_0-1)+m)} \\
&\propto \prod_{j=1}^{m-1} (\Psi_j)^{(m-j)(Ra+a_0)-1} \\
&\quad \left(\left(\left(\sum_{k=1}^R E_{1k} + \nu \right) - \sum_{j=2}^m \left(\sum_{k=1}^R E_{jk} + \nu \right) \prod_{j'=1}^{j-1} \Psi_{j'} \right) \right)^{-m(Ra+a_0)}.
\end{aligned}$$

Thus, under the assumption of differential expression for this gene, the $m-1$ elements of Ψ do not have independent posterior distributions. However, each Ψ_j , conditional on all other Ψ_{-j} , has a posterior distribution that is proportional to a scaled F distribution. Note that the posterior distribution, when $m=2$, equals that previously derived for the two treatment case.

If the gene is not differentially expressed, $z=0$ and $\Psi_j \equiv 1$ for $j=1, \dots, m-1$. The unconditional posterior distribution of Ψ is then

$$\begin{aligned}
p(\Psi|E, a, a_0, \nu, p) &= p(\Psi|E, a, a_0, \nu, p, z=1)p(z=1|E, a, a_0, \nu, p) \\
&\quad + I_{\{1_{m-1}\}}(\Psi)p(z=0|E, a, a_0, \nu, p),
\end{aligned}$$

where $I_{\{1_{m-1}\}}(\cdot)$ is the indicator function for a vector of length $m-1$.

We now derive the posterior distribution of z . We use the fact that $p(z|E, a, a_0, \nu, p)$ can be obtained as the ratio of the joint conditional distribution $p(z, E|a, a_0, \nu, p)$ to the distribution of E conditional on (a, a_0, ν, p) (full derivation in the appendix):

$$\begin{aligned}
p(z|E, a, a_0, \nu, p) &= \frac{p(z, E|a, a_0, \nu, p)}{p(E|a, a_0, \nu, p)} \\
&\propto p(z, E|a, a_0, \nu, p) \\
&= \left(\frac{z\nu^{a_0m}\Gamma(aR+a_0)^m}{\Gamma(a_0)^m \prod_{j=1}^m \left(\sum_{k=1}^R E_{jk} + \nu \right)^{aR+a_0}} \right)
\end{aligned}$$

$$+ \frac{(1-z)\nu^{a_0}\Gamma(amR+a_0)}{\Gamma(a_0)\left(\sum_{j=1}^m\sum_{k=1}^R E_{jk}-\nu\right)^{amR+a_0}} \Bigg) p^z(1-p)^{1-z}.$$

Therefore,

$$p(z|E, a, a_0, \nu, p)$$

$$= K p^z(1-p)^{1-z} \left(\frac{z\nu^{a_0m}\Gamma(aR+a_0)^m}{\Gamma(a_0)^m \prod_{j=1}^m \left(\sum_{k=1}^R E_{jk} + \nu\right)^{aR+a_0}} + \frac{(1-z)\nu^{a_0}\Gamma(amR+a_0)}{\Gamma(a_0)\left(\sum_{j=1}^m\sum_{k=1}^R E_{jk}-\nu\right)^{amR+a_0}} \right),$$

where K is the normalizing constant and K^{-1} is given by

$$K^{-1} = \frac{p\nu^{a_0m}\Gamma(aR+a_0)^m}{\Gamma(a_0)^m \prod_{j=1}^m \left(\sum_{k=1}^R E_{jk} + \nu\right)^{aR+a_0}} + \frac{(1-p)\nu^{a_0}\Gamma(amR+a_0)}{\Gamma(a_0)\left(\sum_{j=1}^m\sum_{k=1}^R E_{jk}-\nu\right)^{amR+a_0}}.$$

We can now compute the probability that $z = 1$ (differential expression of the gene) as follows:

$$\begin{aligned} p(z=1|E, a, a_0, \nu, p) &= \left(\frac{p\nu^{a_0m}\Gamma(aR+a_0)^m}{\Gamma(a_0)^m \prod_{j=1}^m \left(\sum_{k=1}^R E_{jk} + \nu\right)^{aR+a_0}} + \frac{(1-p)\nu^{a_0}\Gamma(amR+a_0)}{\Gamma(a_0)\left(\sum_{j=1}^m\sum_{k=1}^R E_{jk}-\nu\right)^{amR+a_0}} \right)^{-1} \\ &\quad \times p \left(\frac{\nu^{a_0m} \prod_{j=1}^m \prod_{k=1}^R E_{jk}^{a-1} \Gamma(aR+a_0)^m}{\Gamma(a)^m \Gamma(a_0)^m \prod_{j=1}^m \left(\sum_{k=1}^R E_{jk} + \nu\right)^{aR+a_0}} \right) \\ &= \left(1 + \frac{\Gamma(a_0)^m \prod_{j=1}^m \left(\sum_{k=1}^R E_{jk} + \nu\right)^{aR+a_0}}{p\nu^{a_0m}\Gamma(aR+a_0)^m} \frac{(1-p)\nu^{a_0}\Gamma(amR+a_0)}{\Gamma(a_0)\left(\sum_{j=1}^m\sum_{k=1}^R E_{jk}-\nu\right)^{amR+a_0}} \right)^{-1} \\ &= \left(1 + \frac{(1-p)\Gamma(a_0)^{m-1}\Gamma(amR+a_0) \prod_{j=1}^m \left(\sum_{k=1}^R E_{jk} + \nu\right)^{aR+a_0}}{p\nu^{a_0(m-1)}\Gamma(aR+a_0)^m \left(\sum_{j=1}^m\sum_{k=1}^R E_{jk}-\nu\right)^{amR+a_0}} \right)^{-1}. \end{aligned}$$

This model has the same drawback encountered in the two treatment case, in that it does not account for the uncertainty in our estimate of the hyperparameters, (a, a_0, ν, p) .

This can be addressed by assigning a hyperprior to the hyperparameters and then including them in the estimation process. However, as before, these values are constant genome-wide and are estimated using all n genes. Since n is large, we get a good estimate of these values using marginal maximum likelihood methods.

4.4.2 Clustering the Ψ vectors

Here, we consider all treatments in an experiment. For each gene i we have created a vector $\Psi_i = (\Psi_{i1}, \dots, \Psi_{i(m-1)})$ which is the ratio of expected gene expression between the first and second, second and third, ... treatments of the i th gene in the experiment. To implement model-based clustering, we assume that the n Ψ_i 's come from a mixture of C distributions.

As in Section 4.3.2, we believe that it is reasonable to assume that the Ψ_i 's follow a multivariate log-normal distribution. Therefore, we can implement the standard clustering methods on the posterior means of Ψ and on the individual samples from the posterior in the same way as before, using multivariate normal model-based clustering. Again, this will give us the most likely number of clusters for each clustering.

Unlike the two treatment case, the functional form of the posterior distribution of Ψ is not tractable. However, we can calculate the empirical average Kullback-Leibler divergence using samples from the posterior as in Section 4.3.3. These distances can then be used to cluster the posterior distributions as before.

4.5 Maize embryogenesis

We now assess the relative performance of these clustering methods using cDNA microarray data from an embryogenesis experiment in maize. The experiment and data collection are described in Che et al. (2004). The data have been background cleaned,

scan integrated, and normalized as described in Love and Carriquiry (2004). The data for analysis consist of 12,160 genes whose expression were measured under five treatments with 12 replications per treatment. First, we will examine the comparison of embryogenic cells during early embryo development; at 7 and 14 days past separation from callus and placement in embryogenic medium. Second, we will examine expression trends over the whole course of embryo development; at 7,14, and 21 days past placement in an embryogenic medium.

4.5.1 Example using two treatments

In comparing gene expression at 7 and 14 days past placement in an embryogenic medium, we have reduced our problem to a two treatment experiment. We fitted the model and implemented the clustering techniques proposed in Section 4.3. We found that 713 genes of the 12,160 genes have posterior probability of differential expression greater than half. The analytical posterior distributions of the ρ_i are non-continuous in most cases (see Figure 4.1).

When we clustered the log posterior means of ρ , the optimal number of clusters based on differentially expressed genes was five. For the 1000 MCMC samples from the posterior of ρ , the optimal number of clusters was either four or six, with four being much more likely. In order to allow for a reasonable diversity of clusters, we used five groups in clustering when implementing the other methods.

The four different clustering methods (clustering log posterior means, clustering MCMC samples from the posterior distribution, clustering the posterior distributions using the analytical form, and clustering the posterior distributions using the empirical distributions from MCMC) each generated a set of five clusters of the differentially expressed genes. The shape of these clusters can be seen in Figures 4.2-4.5.

Figure 4.1 Posterior distributions of ρ for four genes. These have been calculated analytically. The top two panels show the posterior distributions of ρ for two genes whose expressions do not change significantly between day 7 and day 14 and the bottom two panels correspond to genes which are differentially expressed.

Figure 4.2 The distribution of samples from the posterior distribution of ρ (on log scale) for genes in the five clusters created by clustering the log posterior mean of ρ .

Figure 4.3 The distribution of samples from the posterior distribution of ρ (on log scale) for genes in the five clusters created by clustering each $\log(\rho_b)$, the 1000 MCMC samples from the posterior distribution of ρ .

Figure 4.4 The distribution of samples from the posterior distribution of ρ (on log scale) for genes in the five clusters created by clustering the posterior distributions of ρ , evaluated analytically.

Figure 4.5 The distribution of samples from the posterior distribution of ρ (on log scale) for genes in the five clusters created by clustering the posterior distributions of ρ , evaluated empirically.

Method	Group 1	Group 2	Group 3	Group 4	Group 5
log(means)	0.063 (30)	0.480 (121)	1.798 (300)	3.026 (237)	86.726 (25)
log(MCMC)	0.482 (120)	1.457 (25)	2.067 (411)	3.669 (101)	38.755 (56)
analytical KL	0.005 (14)	0.088 (12)	0.470 (125)	2.404 (544)	117.597 (18)
empirical KL	0.006 (11)	0.443 (135)	2.404 (544)	29.301 (18)	317.880 (5)

Table 4.1 Mean of the posterior distributions of the member of each cluster and membership (out of 713 total) in parentheses.

The sizes of the five groups varied by clustering methods. To compare groups, we ordered the groups by posterior mean of the MCMC samples taken from genes in that group. Table 4.1 shows the means and how many genes belong to each of the five groups obtained using each method. We see that there is a large group of about 500 of these genes with mean approximately equal to 2.5. These are genes for which expression levels were about 2.5 times higher on day 7 after transfer to embryogenic medium relative to levels at 14 days. The first two methods have separated this large group into 2 groups that are close. On the other hand, clustering using the distance between the full posterior distributions yields extreme groups (1,2, and 5 for analytical and 1, 4, and 5 for empirical) with posterior means far from the usual range. These small, more extreme genes will be easier to investigate further with sophisticated biological methods.

Examining the blastx-nt database entries for the functions of these genes, we can examine whether our clusters have a meaningful biological interpretation (<http://ncbi.nlm.nih.gov/BLAST>). Upon doing so, we find some other interesting features of the posterior distribution clustering. A little more than half of the genes clustered have a proposed function in the database, the rest have no known function. All histones, all ribosomal, and all but one of the known chloroplast/photosystem genes are placed in the same group when clustering with the full posterior methods (analytical and empirical KL), which does not occur when applying the other clustering methods

($\log(\text{means})$ and $\log(\text{MCMC})$). The small groups created by posterior distribution clustering have very few genes with known function, so it is difficult to judge their coherence. However, the plasma membrane genes are placed in the same cluster (this is true in all other clustering methods as well).

4.5.2 Example using three treatments

Now, comparing gene expression at 7, 14, and 21 days past placement in embryogenic medium, we have a three treatment case. We applied the model and clustering techniques from Section 4.4. We found that 866 genes of the 12160 genes have posterior probability of differential expression greater than half. Again, examining the blastx-nt database entries for the functions of these genes, we can examine the usefulness of our clusters. The posterior distributions of the Ψ_i are non-continuous in some cases (see Figure 4.6).

When we clustered the log posterior means of Ψ , the optimal number of differentially expressed clusters was 10. For the 1000 MCMC samples from the posterior of Ψ , the optimal number of clusters was ranged between 5 and 9 with 5 being the most likely and 7 next most likely. In order to allow for a reasonable diversity of clusters, we used 7 groups in clustering using our other methods.

The three different clustering methods (clustering log posterior means, clustering MCMC samples from the posterior distribution, and clustering the posterior distributions using the empirical distributions from MCMC) each generated a set of clusters (10 in the first method and 7 in each of the other methods) of the differentially expressed genes. The shape of some of these clusters can be seen in Figures 4.7-4.9.

The locations and sizes of the groups varied by clustering methods. To compare groups, we ordered the groups by the norm of the posterior mean of the MCMC samples taken from genes in that group. Table 4.2 shows the means and how many genes belong

Figure 4.6 Posterior distributions of Ψ for two genes. These have been approximated from the 1000 MCMC samples from the posterior distribution. The first panel shows the posterior distribution of Ψ for a gene which is relatively continuous and the second panel corresponds to a gene with a noncontinuous posterior distribution for Ψ .

Figure 4.7 The distribution of samples from the posterior distribution of Ψ (on log scale) for genes in the four smallest clusters created by clustering the log posterior mean of Ψ .

Figure 4.8 The distribution of samples from the posterior distribution of Ψ (on log scale) for genes in the four smallest clusters created by clustering the log of samples from the posterior of Ψ .

Figure 4.9 The distribution of samples from the posterior distribution of Ψ (on log scale) for genes in the four smallest clusters created by clustering the posterior distributions of Ψ , evaluated empirically.

to each of the clusters for each method. Clustering using the log of the posterior means yields more groups than the other methods. Of course, more clusters will create smaller groups.

We see that there is a large group of about 400 of these genes with mean approximately equal to (0.6,1.0). These are genes for which expression levels decreased by about 40% on day 14 after transfer to embryogenic medium relative to levels at 7 days and then stayed constant at 21 days after transfer. All three clustering methods have a cluster of this type. As in the two treatment case, clustering using the distance between the full posterior distributions yields extreme groups (1,2,5, and 7) with posterior means far from the usual range. These small, more extreme genes will be easier to investigate further with sophisticated biological methods.

Upon examining the blastx-nt database, we find some other interesting features of the posterior distribution clustering. Less than half of the genes clustered have a proposed function, the rest have no known function. The small groups created by posterior distribution clustering have very few genes with known function, so it is difficult to judge their biological significance. The three big functional categories, histones, ribosomal, and chloroplast/photosystem genes, are broken up into several groups when applying any of the clustering methods (3, 3, and 2 groups respectively). However, the replication factor and proliferating cell nuclear antigen genes are placed in the same cluster when using full posterior distributions to cluster which does not occur when applying the other clustering methods.

4.6 Discussion

There is much interest in clustering gene expression data. The possibilities are promising and many alternative methods have been proposed. This area in particu-

Clustering Method		Group									
		1	2	3	4	5	6	7	8	9	10
log(means)	14:7 days	0.034	0.286	0.885	0.550	0.287	1.233	2.129	0.963	13.233	663.577
	21:14 days	1.060	0.855	0.534	0.990	1.282	1.114	1.703	3.342	0.094	0.020
	size	16	62	58	404	38	44	199	16	10	19
log(MCMC)	14:7 days	0.284	0.507	0.983	0.703	1.183	2.010	289.562			
	21:14 days	0.882	0.999	0.563	1.009	1.066	1.823	0.394			
	size	95	354	42	46	69	216	44			
empirical KL	14:7 days	0.005	0.039	0.777	1.438	18.902	78.676	292.240			
	21:14 days	0.920	1.002	0.987	1.528	0.049	0.942	0.005			
	size	5	12	415	296	7	120	11			

Table 4.2 Marginal means of the posterior distributions of the member of each cluster and membership size (out of 866 total).

lar, with the so-called $p \gg n$ problem provides challenges to standard methods. As in identification of differential expression, information sharing between genes and moderation or shrinkage of estimates can bring us more realistic and useful results. Applying Bayesian methods to modeling gene expression and identifying differentially expressed genes has now gained in popularity because of this. Bayesian models provide full posterior inference about the parameters of interest and this large quantity of information can be useful in clustering the genes.

Statistical clustering has frequently been applied to gene expression data using the estimated expression levels without accounting for the uncertainty about the true levels of gene expression. Yeung et al. 2001 and Broët et al. (2002) address variability in expression by assuming that the expression estimates being used can be modeled as normal after log-transformation. In this manuscript, we propose methods for clustering that use the full nature of variability in the values, regardless of form.

Fitting a mixture model to data, when the number of components in the mixture is itself a parameter to be estimated can be challenging. We did not propose varying-dimensional algorithms in this manuscript but they could, in principle, be implemented to perform clustering as we propose here. Within a Bayesian framework, computation is typically carried out using Markov chain Monte Carlo (MCMC) methods, which in this type of applications must allow 'visiting' parameter spaces of varying dimensions. An extension of MCMC methods that is suitable for varying-dimensional problems is called Reversible Jump Markov chain Monte Carlo (RJMCMC) and was introduced by Green (1995). The application of RJMCMC to this problem is possible in certain circumstances. Broët et al. (2002) applied it to the two treatment case. They created a variable of the log ratios of average gene expression for the two treatments. Postulating that these values will come from a mixture of normal components of unknown number,

they are able to estimate the posterior distribution of the number of clusters and the allocation of genes to clusters. While it is reasonable to assume that these log-ratios will be close to normally distributed, the actual posterior distributions for these quantities can be more complicated under certain models. Also, RJMCMC methods are difficult to implement in real time and must be tuned to each particular use.

We have compared clustering methods using the posterior mean and the full posterior distributions of the variable of interest. The most standard method assumes that the posterior means are the observed data and uses model-based clustering to group them while considering the values as stochastic. The second method more fully accounts for the complete shape of the posterior distribution by repeatedly clustering samples from the posterior distribution. The final two methods use the divergence between pairs of posterior distributions, either analytical or empirical.

Here, we have developed a gamma-gamma-Bernoulli model for gene expression with an arbitrary number of treatments and a constant number of replications per treatment that includes our knowledge that some genes are differentially expressed. However, the clustering methods developed are not specific to this model. Any model of gene expression that generates posterior distributions of expression ratios can be employed in the same way as described here. In fact, any set of distributions can be clustered by the four methods described in this work. The cluster distributions used in the first two methods (model-based clustering of the means or individual draws from the posterior) will vary based on assumptions about the data.

In all methods described here, we restricted our attention to genes identified as significantly differentially expressed by grouping all genes with posterior probability of differential expression less than or equal to half into a cluster with the null (no change) expression pattern. This cut-off is reasonable in that all genes in the null expression

pattern cluster are most likely expression to the same degree in all treatments. However, any cut-off may be chosen depending on the focus of the experiment. A higher cut-off will narrow the genes being clustered to those with high probability of differential expression. A cut-off of zero will include all genes; we anticipate that genes with low probability of differential expression will again be clustered together because of the unique shape of their posterior distribution. However, increasing the number of genes greatly increases the computation time needed to implement the clustering methods.

To assess whether our methods produce clusterings with biological significance, we studied the characteristics of the clusters and the distribution of known gene functions within clusters. The clusters created by comparing the full posterior distributions included small clusters with extreme fluctuations in expression. These groups of genes experience dramatic expression changes of the same pattern and this gives evidence that they are commonly controlled or triggered. The blastx-nt descriptions of known genes in each cluster were more biologically consistent for the clusters created using the full posterior distributions. We expect genes with the same function to belong to the same cluster.

Some of the same patterns of gene expression observed in Chapter 2 were found in the clusters generated by the methods proposed here. Gene down-regulated at the beginning of embryogenesis included histones and ribosomal proteins. These mainly showed a pattern of down-regulation at day 14 with steady expression levels at day 21; however, some continued to down-regulate. Also identified as down-regulated with the proposed method in an α -zein (different from the one identified as down-regulated in Chapter 2) which is a storage gene. Included in those genes that up-regulated on both day 14 and day 21 are caleosin (another storage gene), a protease, and most of the glucosidases.

Certainly the added information in a posterior distribution above the posterior mean cannot hinder our attempts to find groups of objects that behave similarly. When the posterior mean is used for clustering, we are implicitly assuming that these values are observed without error. In fact, we know that there is substantial variation. When the posterior distributions are decidedly non-normal (even under transformation), summary statistics such as the mean may indeed impair our ability to distinguish between objects. We believe that our method creates useful clusterings using the full information of the posterior distribution. This should be helpful in any clustering scenario where we believe that the quantities of interest for clustering are not well-behaved or well represented by summary statistics. Though the analytical posterior distribution is the gold standard for measuring distances and clustering, the empirical distribution can be used in cases where an analytical representation of the posterior density is not feasible and has most of the advantages.

To fully compare the performance of the four clustering methods proposed here, and to assess whether the added analytical and computational burden results in more reliable gene groupings, it is necessary to conduct a simulation study. We plan on doing so in the future and will report on those results elsewhere.

4.7 References

- Banfield, J.D. and Raftery, A.E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics* 49, 803–821.
- Bell, E.T. (1934). Exponential numbers. *American Mathematical Monthly* 41, 411–419.

- Broët, P., Richardson, S., and Radvanyi, F. (2002). Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. *Journal of Computational Biology* 9(4) 671–683
- Bryan, J., Pollard, K.S., and van der Laan, M.J. (2002). Paired and unpaired comparison and clustering with gene expression data. *Statistica Sinica* 12(1) , 87–110
- Carlin, B.P. and Louis, T. (1997). *Bayes and Empirical Bayes* Chapman & Hall, London
- Che, P., Love, T.M., Frame, B.R., Wang, K., Carriquiry, A.L., and Howell, S.H. (2004). Gene expression program during maturation and germination of somatic embryos in maize cultures. submitted to *Plant Molecular Biology*.
- Devroye, L. (1987). *A Course In Density Estimation* Birkhäuser Verlag, Boston
- Fraley, C. and Raftery, A.E.(1999). MCLUST: Software for Model-Based Clustering and Discriminant Analysis. *Journal of Classification* 16, 297–306.
- Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis* Chapman & Hall, London
- Goldstein, D.R., Ghosh, D., and Conlon, E.M. (2002). Statistical issues in the clustering of gene expression data. *Statistica Sinica* 12(1) , 219–240
- Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82 711–732
- Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Journal of the Royal Statistical Society, Series A* 186, 453–461.
- Kaufman, L. and Rousseuw, P.J. (1990). *Finding Groups in Data- An Introduction to Cluster Analysis*. Wiley Series in Probability and Mathematical Sciences, New York

- Kerr, M.K., and Churchill, G.A. (2001). Experimental Design for Gene Expression Microarrays. *Biostatistics* 2, 183–201
- Kullback, S. and Leibler, R.A. (1951). On information and sufficiency. *Annals of Mathematical Statistics* 22(1), 79–86.
- Lazzeroni, L.C. and Owen, A.(2002). Plaid models for gene expression data. *Statistica Sinica* 12(1) , 61–86.
- Liang, J., and Kachalo, S. (2002). Computational analysis of microarray gene expression profiles: clustering, classification, and beyond. *Chemometrics and Intelligent Laboratory Systems* 62(2) , 199–216
- Love, T.M. and Carriquiry, A. (2004). Incorporating Multiple cDNA Microarray Slide Scans -Application to Somatic Embryogenesis in Maize. submitted to *The Journal of the American Statistical Association*.
- Newton, M.A., Kendzierski, C.M., Richmond, C.S., Blattner, F.R., and Tsui, K.W. (2001). On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data. *Journal of Computational Biology*, 8(1), 37–52.
- Richmond, C.S., Glasner, J.D., Mau, R., Jin, H., and Blattner, F.R. (1999) Geneome-wide expression profiling in Escherichia coli K-12 *Nucleic Acids Research*, 27(19), 3821–3835.
- Smyth, G.K., Yang, Y.H., and Speed, T. (2002). Statistical Issues in cDNA Microarray Data Analysis. *Functional Genomics:Methods and Protocols* Humana Press, Totowa, NJ.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E. and Ruzzo, W. L. (2001). Model-based clustering and data transformation for gene expression data. *Bioinformatics* 17,977-987.

4.8 Appendix

4.8.1 Posterior distribution of z in the two treatment case

Consider the posterior distribution of z , $p(z|X, Y, a, a_0, \nu, p)$. This is proportional to the joint distribution of z and the data as follows:

$$\begin{aligned}
p(z|X, Y, a, a_0, \nu, p) &= \frac{p(z, X, Y|a, a_0, \nu, p)}{p(X, Y|a, a_0, \nu, p)} \\
&\propto p(z, X, Y|a, a_0, \nu, p) \\
&= p(X, Y|a, a_0, \nu, p, z)p(z|a, a_0, \nu, p) \\
&= \left(z \int \int p(X, Y, \theta_X, \theta_Y|a, a_0, \nu, p, z=1) d\theta_X d\theta_Y \right. \\
&\quad \left. + (1-z) \int p(X, Y, \theta|a, a_0, \nu, p, z=0) d\theta \right) p(z|p) \\
&= \left(z \int \int p(X, Y|\theta_X, \theta_Y, a, z=1) p(\theta_X, \theta_Y|a_0, \nu, z=1) d\theta_X d\theta_Y \right. \\
&\quad \left. + (1-z) \int p(X, Y|\theta, a, z=0) p(\theta|a_0, \nu, z=0) d\theta \right) p^z(1-p)^{1-z} \\
&= \left(z \int \int \Gamma(a)^{-2} (\theta_X \theta_Y)^a (XY)^{a-1} e^{-X\theta_X - Y\theta_Y} \right. \\
&\quad \times \Gamma(a_0)^{-2} \nu^{2a_0} (\theta_X \theta_Y)^{a_0-1} e^{-\nu(\theta_X + \theta_Y)} d\theta_X d\theta_Y \\
&\quad \left. + (1-z) \int \Gamma(a)^{-2} \theta^{2a} (XY)^{a-1} e^{-\theta(X+Y)} \right. \\
&\quad \left. \times \Gamma(a_0)^{-1} \nu^{a_0} \theta^{a_0-1} e^{-\nu\theta} d\theta \right) p^z(1-p)^{1-z} \\
&= \left(\frac{z(XY)^{a-1} \nu^{2a_0}}{\Gamma(a)^2 \Gamma(a_0)^2} \int \int (\theta_X \theta_Y)^{a+a_0-1} e^{-\theta_X(X+\nu) - \theta_Y(Y+\nu)} d\theta_X d\theta_Y \right. \\
&\quad \left. + \frac{(1-z)(XY)^{a-1} \nu^{a_0}}{\Gamma(a)^2 \Gamma(a_0)} \int \theta^{2a+a_0-1} e^{-\theta(X+Y+\nu)} d\theta \right) p^z(1-p)^{1-z} \\
&= p^z(1-p)^{1-z} \left(\frac{z(XY)^{a-1} \nu^{2a_0}}{\Gamma(a)^2 \Gamma(a_0)^2} \times \frac{\Gamma(a+a_0)}{(X+\nu)^{a+a_0}} \times \frac{\Gamma(a+a_0)}{(Y+\nu)^{a+a_0}} \right. \\
&\quad \left. + \frac{(1-z)(XY)^{a-1} \nu^{a_0}}{\Gamma(a)^2 \Gamma(a_0)} \times \frac{\Gamma(2a+a_0)}{(X+Y+\nu)^{2a+a_0}} \right) \\
&= p^z(1-p)^{1-z} \left(\frac{z(XY)^{a-1} \nu^{2a_0}}{(X+\nu)^{a+a_0} (Y+\nu)^{a+a_0}} \frac{\Gamma(a+a_0)^2}{\Gamma(a)^2 \Gamma(a_0)^2} \right. \\
&\quad \left. + \frac{(1-z)(XY)^{a-1} \nu^{a_0}}{(X+Y+\nu)^{2a+a_0}} \frac{\Gamma(2a+a_0)}{\Gamma(a)^2 \Gamma(a_0)} \right).
\end{aligned}$$

4.8.2 Posterior distribution of z in the m treatment case

Consider the posterior distribution of z , $p(z|E, a, a_0, \nu, p)$. This is proportional to the joint distribution of z and the data as follows:

$$\begin{aligned}
p(z|E, a, a_0, \nu, p) &= \frac{p(z, E|a, a_0, \nu, p)}{p(E|a, a_0, \nu, p)} \\
&\propto p(z, E|a, a_0, \nu, p) \\
&= p(E|a, a_0, \nu, p, z)p(z|a, a_0, \nu, p) \\
&= \left(z \int p(E, \Theta|a, a_0, \nu, p, z=1)d\Theta \right. \\
&\quad \left. + (1-z) \int p(E, \theta|a, a_0, \nu, p, z=0)d\theta \right) p(z|p) \\
&= \left(z \int p(E|\Theta, a, z=1)p(\Theta|a_0, \nu, z=1)d\Theta \right. \\
&\quad \left. + (1-z) \int p(E|\theta, a, z=0)p(\theta|a_0, \nu, z=0)d\theta \right) p^z(1-p)^{1-z} \\
&= \left(z \int \prod_{j=1}^m \prod_{k=1}^R \frac{\theta_j^a}{\Gamma(a)} E_{jk}^{a-1} e^{-E_{jk}\theta_j} \prod_{j=1}^m \frac{\nu^{a_0}}{\Gamma(a_0)} \theta_j^{a_0-1} e^{-\nu\theta_j} d\Theta \right. \\
&\quad \left. + (1-z) \int \prod_{j=1}^m \prod_{k=1}^R \frac{\theta^a}{\Gamma(a)} E_{jk}^{a-1} e^{-E_{jk}\theta} \frac{\nu^{a_0}}{\Gamma(a_0)} \theta^{a_0-1} e^{-\nu\theta} d\theta \right) p^z(1-p)^{1-z} \\
&= \left(\frac{z \nu^{a_0 m} \prod_{j=1}^m \prod_{k=1}^R E_{jk}^{a-1}}{\Gamma(a)^{mR} \Gamma(a_0)^m} \right. \\
&\quad \int \left(\prod_{j=1}^m \theta_j \right)^{aR+a_0-1} \exp \left(- \sum_{j=1}^m \theta_j \left(\sum_{k=1}^R E_{jk} + \nu \right) \right) d\Theta \\
&\quad + \frac{(1-z) \nu^{a_0} \prod_{j=1}^m \prod_{k=1}^R E_{jk}^{a-1}}{\Gamma(a)^{mR} \Gamma(a_0)} \\
&\quad \left. \int \theta^{aR+a_0-1} \exp \left(- \theta \left(\sum_{j=1}^m \sum_{k=1}^R E_{jk} - \nu \right) \right) d\theta \right) p^z(1-p)^{1-z} \\
&= \left(\frac{z \nu^{a_0 m} \Gamma(aR+a_0)^m}{\Gamma(a_0)^m \prod_{j=1}^m \left(\sum_{k=1}^R E_{jk} + \nu \right)^{aR+a_0}} \right. \\
&\quad \left. + \frac{(1-z) \nu^{a_0} \Gamma(amR+a_0)}{\Gamma(a_0) \left(\sum_{j=1}^m \sum_{k=1}^R E_{jk} - \nu \right)^{amR+a_0}} \right) p^z(1-p)^{1-z}.
\end{aligned}$$

CHAPTER 5. GENERAL CONCLUSIONS

5.1 General Discussion

We have developed tools to analyze gene expression data from microarray data. The data used to motivate these methods was generated by a cDNA microarray experiment on maize embryogenesis. Five time points during the development of somatic embryos in sprouting plants were measured with 12 replications for each time point. The goal of this experiment was to identify genes with patterns of changing expression during embryo development.

Using standard methodology, we originally identified 1,026 differentially expressed genes. These could be classified into 12 types of expression change over time. The standard methodology used consisted of selecting the 'best' reading from available scans of each slide and calculating the background-corrected, normalized values of gene expression in each case. Independent two-way analysis of variance (ANOVA) tests were performed on each gene. The p-values obtained from these tests were corrected for multiple comparisons and used to identify the set of significantly differentially expressed genes. Model-based clustering of the time-sequential ratios of mean expression were used to identify 12 clusters within the differentially expressed genes.

One source of potentially significant measurement error is the settings of the instruments (laser and sensor) that are used to obtain the data. Because the 'optimal' settings may vary from slide to slide, operators often obtain multiple readings of each slide and

then choose the 'best', meaning the reading that includes the fewest saturated spots and the fewest under-exposed spots. In Chapter 3, we propose a general hierarchical modeling approach that allows incorporation of as many readings as may be available for each slide into the model, even if the number of readings per slide vary across slides. The basic premise is that each reading of a spot contains some information about the true expression of the gene and that if an appropriate scaling factor for each spot can be estimated, then all readings for a spot estimate the same quantity and can be combined.

We ran a simulation study and assessed bias and root mean squared error of the gene expression estimators derived in Chapter 3 over repeated sampling. Using simulated gene expression data, we applied several approaches (including the approach proposed here) to estimate gene expression for 1,000 genes and compared the various methods on the basis of bias, RMSE and effective range of the estimates that were obtained. The hierarchical modeling approach we propose had smaller bias and smaller RMSE than all other estimators, suggesting that our method is effective in combining the information from different scans. Also, the effective range of expression values is greatly increased using our proposed method. This is an improvement because we assume that there is a larger range of expression values than that observable from the scanner. These simulation results may be overly optimistic and should be interpreted cautiously because the simulated data were generated from the model proposed here.

We implemented the proposed hierarchical modeling approach on the maize embryo-genesis data. For each of the 30 slides in this experiment, three readings with various settings were originally scanned. We repeated the original analysis of these data using posterior estimates of the expression values generated using all three readings instead of observed values from an individual reading. We note that the variance of expression estimates is lower when based on three readings, as would be expected. Because of the

smaller bias and RMSE in gene expression estimates, inferences about the set of genes involved in somatic embryogenesis in maize change drastically when statistical analyses are based on one or on three readings of each slide. In this experiment, 3,840 genes were identified as differentially expressed when using all of the available readings; a large improvement over the 1,026 genes originally identified. This stronger conclusion is possible because of the increased information used when all readings are incorporated.

Clustering genes using expression data has the potential to help identify biological relationships and systems. We desire groups of genes with similar changes in expression over different treatments. Bayesian models provide full posterior inference about the parameters of interest and this information can be useful in clustering the genes. In Chapter 4, we derived a hierarchical mixture model of gene expression values that includes our knowledge that some genes are differentially expressed. This model is a gamma-gamma-Bernoulli hierarchical model that assumes a design with m treatments each measured with R replications. This model is used to generate posterior distributions of the variables of interest in our clustering methods. It is also used to identify genes with low posterior probability of differential expression.

Finally, we have compared clustering methods using either the posterior mean or the full posterior distributions of the variable of interest. In all methods, we first create a cluster of genes with the null (no change) expression pattern from those genes with small posterior probability of differential expression. The most standard clustering method assumes that the posterior means summarize the distributions and uses model-based clustering to group the means while considering the values as stochastic. The second method more fully accounts for the complete shape of the posterior distribution by repeatedly clustering samples from the posterior distribution. The final two methods use a variant of Kullback-Leibler divergence between pairs of posterior distributions, either

analytically or empirically. These measures of difference between the distributions are used to cluster distributions that are similar in shape and separate distributions with different shapes.

We implemented all four proposed clustering methods on the maize embryogenesis data. First, we clustered genes based on their pattern of behavior between days 7 and 14 after immersion in embryogenic medium. 11,447 genes were clustered in the “null” cluster as genes with no change in gene expression between the time points. The remaining 713 genes were clustered into five groups by each of the four clustering methods. Second, we clustered genes based on their expression pattern over days 7, 14, and 21 after immersion in embryogenesis medium. 11,294 genes were clustered as genes with no change in expression over the interval and placed in the “null” cluster. The other 866 genes were clustered into seven groups by the proposed clustering methods. We studied the characteristics of the clusters and the distribution of known gene functions within clusters. The clusters created by comparing the full posterior distributions included small clusters of genes with extreme fluctuations in expression. These groups of genes experience dramatic expression changes of the same pattern and this gives evidence that they are commonly controlled or triggered. The blastx-nt descriptions of known genes in each cluster were more biologically consistent for the clusters created using the full posterior distributions. We expect genes with the same function to belong to the same cluster. The creation of more biologically useful and consistent clusters is possible because of the added posterior information at work in clustering with full posterior distributions.

The application of sophisticated statistical techniques to the specific problems of microarray data have improved our estimation of expression intensities, our discovery of differentially expressed genes, and our clustering of genes into groups with similar expression patterns. In doing so, we hope to improve future work in micrarrays and molecular

biology as well as to have introduced novel statistical applications and methods.

5.2 Recommendations for future research

There are several directions which future research could take out of this work. These include expanding applications to other technologies, improving preprocessing of microarray data, expanding the model for multiple scans, expanding the model of expression to include blocked effects or covariates, creating a measure for choosing the most appropriate number of clusters, and integrating clustering with hierarchical modeling of expression values.

5.2.1 Related technologies

The first potential extensions of this work are to related technologies. Much of this work could be applied to other related and diverse problems. Various molecular biology technologies such as gel electrophoresis, use fluorescent dyes to measure quantities of molecules and these can be repeatedly scanned under changing settings. Measuring unknown concentrations of a group of chemicals in a liquid can be done by measuring the amounts in different dilutions of the liquid. The same technique of titrations is used to measure the viral load or amount of neutralizing antibodies in plasma. In these cases, the model for incorporating scaled replications presented in Chapter 3 can be applied. The clustering methods in Chapter 4 are more widely applicable to any situation where a posterior distribution of parameters of interest can be developed. It will be more useful than standard methods when these posterior distributions are discontinuous or multi-modal.

5.2.2 Improved and consistent preprocessing of expression values

The application of Bayesian hierarchical modeling to incorporating multiple scans points to the possibility of implementing such models in other aspects of preprocessing and normalization. Also needing development is a method for assessing the relative quality of different preprocessing methods. An intuitive approach is to measure the consistency between replicates within the experiment. If replicates are truly replicating the same measurement, then a method which creates clean data with the most consistency across replicates is desirable.

5.2.3 Intensity-dependent scaling of multiple scans

We have assumed here that a single multiplicative factor δ_j is applicable to expression levels of all spots on a slide. That is, if a specific laser and sensor setting tends to increase expression levels multiplicatively, we assume that the multiplicative factor is uniform across all spots on a slide. This assumption may not hold in all situations, but modeling each spot within a slide individually makes the problem intractable from an analytical point of view. Simulation results show that the bias with which we can estimate gene expression is associated to expression levels, indicating that different spots on the slide might require different scalings to correct for the effect of the same laser and sensor settings.

5.2.4 Blocked or covariate effects in the hierarchical model

A useful extension of the hierarchical model for gene expression under multiple treatments would be one which includes blocked or covariate effects. In many experiments, such as that presented in Chapter 2, we have a covariate within the treatment replicates. Estimating and controlling for the variation due to the covariate will improve our conclusions about differential expression. In some cases, both treatment variables may

be of biological interest and should be estimated together.

5.2.5 Choosing the appropriate number of clusters using distribution divergences

Currently we rely on the number of clusters identified via standard methods in order to identify the most likely number of groups in the dataset. A statistic like BIC for model-based clustering, to identify the most likely number of clusters would be useful in the development of posterior distribution clustering. Such a statistic would measure how well all genes fit into the clusters they were assigned for each possible number of clusters. We have not begun to explore the possibility of developing alternative model selection criteria when the items being clustered are distributions, but this may be a promising area for future research.

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this thesis. First and foremost, Dr. Alicia Carriquiry for her guidance, patience and support throughout the research and the writing of this thesis.

For including me in discussions and much help answering questions and explaining the biological contexts of the problem, I thank Stephen Howell. I am also grateful to Ping Che, Kan Wang, Browyn Frame for their work on the maize embryogenesis experiment and Dan Nettleton and Karin Dorman for suggestions and aid in writing and editing this thesis.

I would additionally like to thank Dr. Isaacson for his guidance throughout the initial stages of my graduate career and Dr. Kaiser for his inspirational teaching style and continued support.