BAYESIAN LEARNING IN BIOINFORMATICS

A Dissertation

by

DAVID L. GOLD

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2007

Major Subject: Statistics

BAYESIAN LEARNING IN BIOINFORMATICS

A Dissertation

by

DAVID L. GOLD

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Bani Mallick
Committee Members,	Kevin R. Coombes
	David Dahl
	Faming Liang
	Edward Dougherty
Head of Department,	Simon Sheather

August 2007

Major Subject: Statistics

ABSTRACT

Bayesian Learning in Bioinformatics. (August 2007)
David L. Gold, B.A., The University of Texas;
M.S., Texas A&M University
Chair of Advisory Committee: Dr. Bani Mallick

Life sciences research is advancing in breadth and scope, affecting many areas of life including medical care and government policy. The field of Bioinformatics, in particular, is growing very rapidly with the help of computer science, statistics, applied mathematics, and engineering. New high-throughput technologies are making it possible to measure genomic variation across phenotypes in organisms at costs that were once inconceivable. In conjunction, and partly as a consequence, massive amounts of information about the genomes of many organisms are becoming accessible in the public domain. Some of the important and exciting questions in the post-genomics era are how to integrate all of the information available from diverse sources.

Learning in complex systems biology requires that information be shared in a natural and interpretable way, to integrate knowledge and data. The statistical sciences can support the advancement of learning in Bioinformatics in many ways, not the least of which is by developing methodologies that can support the synchronization of efforts across sciences, offering real-time learning tools that can be shared across many fields from basic science to the clinical applications. This research is an introduction to several current research problems in Bioinformatics that addresses integration of information, and discusses statistical methodologies from the Bayesian school of thought that may be applied.

Bayesian statistical methodologies are proposed to integrate biological knowledge and improve statistical inference for three relevant Bioinformatics applications: gene expression arrays, BAC and aCGH arrays, and real-time gene expression experiments. A unified Bayesian model is proposed to perform detection of genes and gene classes, defined from historical pathways, with gene expression arrays. A novel Bayesian statistical method is proposed to infer chromosomal copy number aberrations in clinical populations with BAC or aCGH experiments. A theoretical model is proposed, motivated from historical work in mathematical biology, for inference with real-time gene expression experiments, and fit with Bayesian methods. Simulation and case studies show that Bayesian methodologies show great promise to improve the way we learn with high-throughput Bioinformatics experiments.

To Marlene S. Gold and Daisy R. Gold

ACKNOWLEDGMENTS

The work embodied here advances research goals I have shared with my mentors and colleagues for the past seven years. Several of my early mentors, by their fine example, encouraged me to progress far beyond my expectations. Thank you to the many fine faculty at The University of Texas M.D. Anderson Cancer Center, above all Kevin Coombes, Keith Baggerly and Don Berry for building such an incredible Bioinformatics research community in Houston, TX.

The fine education that I received at Texas A&M University will no doubt help to guide me through the many trials I expect to face. It is with joy and sadness that I go to face greater challenges. Bani Mallick taught me so much more than I expected. I will always value his truly free mind and elegant calm. To the many fine faculty in the Department of Statistics at Texas A&M University, I applaud and thank you for your dedication. Gig'em.

To my father, loving deceased mother, and wife Daisy, thank you for always believing in me, and more.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

TABLE Page

LIST OF FIGURES

FIGURE Page

FIGURE Page

CHAPTER I

INTRODUCTION

Life sciences research is advancing in breadth and scope, affecting many areas of life including medical care and government policy. The field of Bioinformatics in particular is growing very rapidly with the help of computer science, statistics, applied mathematics, and medical engineering. New high-throughput technologies are making it possible to measure variation in and across genomes as previously unheard of cost. In conjunction, and partly as a consequence, massive amounts of information about the genomes of many organisms are becoming accessible in the public domain. One of a host of very important and exciting questions in the post-genomic era is how to integrate all of the information available from so many diverse sources.

Learning in complex systems biology requires that information be shared in a natural and interpretable way, to integrate knowledge and data. The statistical sciences can support the advancement of learning in Bioinformatics in many ways, not the least of which is by developing methodologies to support synchronization of efforts across sciences, offering real-time learning tools that can be shared across many fields from basic science to clinical applications. This research offers an introduction to several current research problems in Bioinformatics dealing with integration of information with discussion of Bayesian statistical methodologies that show promising solutions.

In the early days of Bioinformatics, much of the work concentrated on string process-

---

This dissertation follows the style of the *Journal of the Royal Statistical Society*.

ing, sequencing the genomes of many organisms. With the completion of sequencing of many genomes, including the human genome, and vast advances in the breadth and quality of high-throughput experiments, attention has turned to phenotyping and genotyping studies, largely in support of the clinical science. One such class of experiments, called microarray mRNA expression experiments, includes a very general class of high-throughput experiments designed to measure variation in mRNA transcripts. More generally, microarray experiments may be used to detect variation in DNA and proteins as well. High-throughput genomics technologies may be characterized as large and relatively inexpensive experiments designed for exploring genomic variation; large in the sense that thousands of genes, perhaps entire genomes, may be measured simultaneously and inexpensive in the sense that the relative cost on a per-gene basis is vastly reduced to a fraction of the cost of measuring the individual genes directly. One disadvantage of high-throughput experiments is that the experimental conditions are suboptimal for all genes. Nevertheless, the efficiency and cost of the experiments offer advantageous opportunities to learn about and explore new possibilities in genomics.

A growing vision in the medical community is customized medicine, or personalized medicine, making direct use of a patient's unique genomic signature. The concept is not new, as clinicians use a great amount of information about a patient to make a diagnoses and prognoses. What is new, is the benefit of using one's entire genomic disease profile, possibly with high-throughput experimental results. In order to make personalized medicine a reality, with the use of high-throughput technologies, the quality of both the experiments, and the data analysis must improve. Analyzing high-throughput genomics data is complicated by the overwhelming number of variables (gene, transcripts, or proteins) that need to be understood, along with the potential

relationships between them. It is typically nontrivial to learn about gene interactions in array studies, as these are often underpowered for such purposes. Typically, the designs include a few factors, such as tissues or treatments, for which expression is observed. A basic understanding of the molecular pathology of disease is essential for Bioinformatics analysis, but even more prior biological knowledge is required for an understanding of genes involved.

Bayesian statistics is a growing field, recognized in the applied life science and clinical research literature for its very flexible learning approach, which allows for integration of historical information with data. The makes Bayesian statistics attractive to biologists, who depend on historical information to make conclusions about complex organisms. The Bayesian paradigm is also more intuitive than frequentist statistics, relying on basic probability concepts to make conclusions, rather than p-values. This makes it attractive in large Bioinformatics studies, often requiring expertise across disciplines.

In this three part dissertation, each chapter focuses on an area of specific research aimed at improving learning in high-throughput Bioinformatics. Chapter II is devoted to gene detection with microarray expression experiments. Microarray experiments generate an exhaustive quantity of data for thousands of genes. Identifying effective targets for treatment in high-throughput experiments is typically complicated by the uncertainty in the collections of dependent genes that are responsible for events like cancer. Higher-level information about the genes, concerning gene classes, may be defined from historical pathways. The utility of historical pathways for microarray analysis is investigated in a Bayesian paradigm, called Bayesian Learning for Microarrays (BLM). The Bayesian approach is ideal for investigating the utility of historical

knowledge, as the strength of the prior information on the results may be gauged and controlled. In Chapter III, a novel method is proposed for modeling cytogentics with BAC arrays and aCGH. One difficulty in modeling high-throughput human chromosomal data is that genetic instability can be subject specific and therefore disease populations are heterogeneous. A Bayesian Change Point Analysis (BCPA) model is developed for high-throughput aCGH experiments. The method very flexibly allows one to model dependencies between inhomogeneous samples. Chapter IV delves into a new area of research, experiments that offer real-time profiles of gene expression in living cells. Experimental designs of microarrays are limited, lacking to the temporal resolution necessary in order to make detailed inferences about interactions between genes and extracellular events. New opportunities for discovery are possible. Statistical inference is challenging at many levels. There is uncertainty in the historical gene pathways, and the theoretical models proposed in the literature are not tailored for these experiments. A class of theoretical models are offered for inference. Experimental design issues and modeling assumptions are thoroughly discussed. Posterior inference is conducted with a case study of the publicly available S.O.S. data.

CHAPTER II

BAYESIAN LEARNING FOR MICROARRAYS

II.1. Introduction

Clinical researchers are greatly interested in discovering gene classes, collections of interacting genes, that are associated with disease. Many investigators believe that events such as disease onset are manifestations of highly evolved and complex chains of molecular interactions. Gene expression is linked in multifaceted biochemical pathways varying according to temporal sequences, prearranged by a genetic program. The molecular precursors to human disease are especially difficult to determine after the necessary changes for onset have been made, i.e. after transformation. It is advantageous to be able to interpret gene-wise events in a higher-level sense concerning gene pathways. Collective inference on genes and classes of dependent genes, defined from say historical pathways, provides investigators with an additional level of information about the underlying biology driving morphological changes, while accounting for the uncertainty in both.

The uncertainty in gene pathways responsible for disease typically complicates detection of effective targets for medical treatment. Rather than pathways, microarray technologies measure gene specific events. The enormity of the data, as measured by the number of genes on a microarray (typically in the thousands) relative to the number of arrays (typically less than 100), requires special methodologies for statistical analysis in order to draw logical conclusions about the genes. It is unrealistic to expect to determine the dependencies between all of the genes in a microarray

experiment, given the typical sample sizes. Use of historical pathways may serve as a vantage point to initiate the learning process.

We investigate a fully Bayesian method called Bayesian Learning for Microarrays, to detect changes in genes and gene classes defined from historical pathways. BLM is a unified approach that may serve to account for the uncertainty in genes and gene classes collectively. BLM allows flexibility to incorporate historical pathway knowledge to the extent of one's beliefs, and explicitly define gene-wise dependencies. The Bayesian approach is also ideal for investigating the utility of the historical pathway knowledge with prior sensitivity analysis. Examples with simulated data and case studies with public array data sets are used to demonstrate the utility of BLM for microarray analysis.

II.2. High-throughput Gene Expression Experiments

While DNA contains the blue print for life, differences between tissues are manifestations of the ways in which genes are expressed. DNA is related to proteins, the cellular machinery responsible for cellular events, through the processes of transcription and translation. DNA sequences along chromosomes contain the information needed to synthesize peptide chains, i.e. proteins, but that information is not utilized until it is communicated. Segments of DNA are initially transcribed to messenger RNA (mRNA), shorter transcripts containing copies of the genetic code that can leave the nucleus. The result of transcription is also known as gene expression. Once in the cytoplasm, mRNA transcripts can be translated to proteins with the help of ribosomes. Ribosomes attach to the mRNA and following the information contained therein, order to assemble peptide chains (Figure 1).

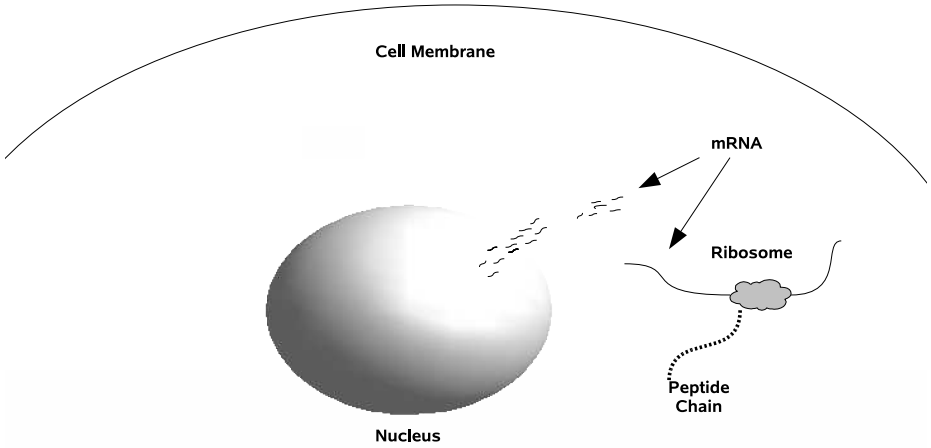

Fig. 1 Transcription and translation - mRNA leaves the nucleus and is translated to a peptide chain by a ribosome.

Figure 2 shows a diagram of a gene expression microarray experiment. Samples of mRNA are collected from different specimens. The mRNA is reverse transcribed to more stable clone DNA (cDNA), and dye labeled. For example, in a cancer study, normal tissue may be labeled with Cy5 and cancer tissue with Cy3. The dyes are actually small molecules attached to the cytosine bases along the cDNA during the sample processing. On the array surface, spots are positioned, each corresponding to a single gene. At each spot, homogeneous DNA sequences are demobilized at one end to the array surface by a ligand. The labeled cDNA from each sample is allowed time to permeate the array. The labeled cDNA sequences have a natural affinity to bind to their complementary sequences on the array, a process called hybridization. A laser is shined over each spot and the excitation causes the respective dye materials to emit different wavelength. Cy3 emits green light and Cy5 red light, as interpreted by the human retina. Spots with more (or less) hybridization should emit more (or less) relative fluorescent light between the channels. The slides are scanned for each dye

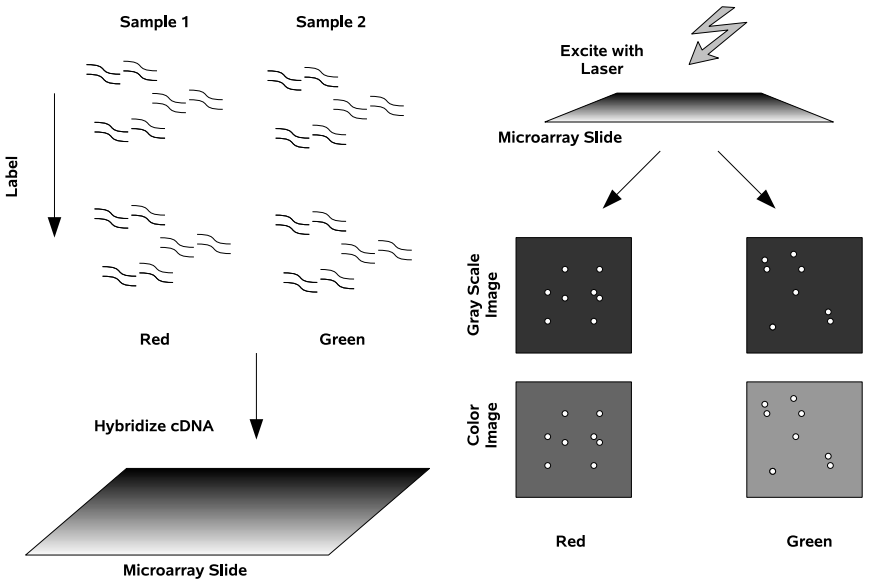

Fig. 2 Two-channel microarray experiment

and gray scale images are produced. The images are quantified, i.e. the relative pixel strengths with in each spot between the dyes are summarized numerically. There are many more steps involved in conducting a microarray experiment, from sample processing to image acquisition, contributing to experimental confounding variation (see: Yang and Speed, 2002).

II.2.1. Preprocessing Microarray Data

Microarrays are known to exhibit sources of variation attributable to the many technological steps involved in manufacturing an experiment (Bolstad *et al.* , 2003). A complete discussion of microarray data cleaning and preprocessing is beyond the scope of this text, taking us away from the current topics of interest. A brief sum-

mary includes methods to account and adjust for variation accumulated during the experimental procedures, unrelated to biology. For example, spatial variation on the chip, or variation between chips, if unaccounted for, may lead to faulty conclusions. Moreover, variation in experiments within labs on different days, and between labs has also been noted in the literature as distorting conclusions (Conlon *et al.*, 2006). The methods and procedures for preprocessing are platform and technology dependent. In many cases, special steps are needed to adjust for lab-specific effects. For more discussion see work by Geller *et al.* (2003), Bolstad *et al.* (2003), Gold *et al.* (2005), and Lewin *et al.* (2005).

II.2.2. Historical Pathways

Prior information may serve as a useful vantage point to initiate learning about disregulation in gene pathways in microarray experiments. The biological motivation is that genes are known to be dependent, and it makes sense to start with historically documented relations. Public domain access to information about genes and the genomes of many organisms is rapidly advancing. This is partly a consequence of high-throughput technologies. Pathguide, www.pathguide.org, maintains an extensive list of online gene database resource tools. Some of the more popular resources for the human genome are Gene Ontology (GO), the Kyoto Encyclopedia of Genes and Genomes (KEGG), and BioCarta. There is currently little uniformity in the way information is reported in each database. Integrating information between the databases requires good understanding of the database formats. For example, GO maintains and updates a public database with a mapping from the known genes to a highly structured vocabulary describing the biological processes, molecular functions, and cellular locations of gene products. Each gene annotation in GO is accompanied

by an evidence code, indicating how the information was curated. Pathway databases such as BioCarta provide symbolic reproductions of known protein interactions in the form of metabolic pathway maps. Efforts are underway to design semantic solutions that will integrate all of this information.

Despite the massive amount of prior information available, there is still much uncertainty attached to gene pathways. The rich class of known protein interactions represents only a fraction of the plausible relationships operating in living cells. A practical concern is that no one technology measures all of the different types of interactions needed to learn a complete pathway. The key issue in gene expression studies is that detection of biology pathways is typically hindered by underpowered study designs. Historical pathways may help to complete the picture.

Incorporating historical pathways into an array analysis explicitly is a goal that many share in Bioinformatics, although there are drawbacks. The prior information of many genomes is incomplete with poor coverage of the genes. In some cases the prior information is of dubious quality, or the pathway dependencies only are known to exist under a specific set of circumstances. Some sources of prior information may be more useful than others for array analysis. One form of prior information that has obvious benefits includes information about operon classes of genes. Genes in the same operon class have the same consensus upstream binding motif, and therefore are regulated by the same transcription factor (TF). When a TF binds to the promoter region of a gene, expression is initialized. In expression array experiments, this is an obvious source of historical information that would seem useful, as expression is what is being measured. TF databases exist for some organisms, such as yeast and e-coli. A goal is to make collective inference on genes and gene classes, defined from historical path-

Table 1. Counts by class and differential expression

	Genes in Class b	*Not in Class b*	*Total*
Genes Changed	n_{11}	n_{11}	$n_{1\cdot}$
Genes Not Changed	n_{21}	n_{22}	$n_{2\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

ways, while accounting for the uncertainty in both, by integrating historical pathways explicitly into the analysis.

II.2.3. Gene Class Detection

There is a growing body of evidence to suggest that known gene regulatory factors can explain variation on arrays (Tamada *et al.*, 2003; Allocco *et al.*, 2004; Bharjwaj *et al.*, 2005; Nagaraj *et al.*, 2004), although these results are premature. Nevertheless, this is important empirical evidence, as it demonstrates that some known sources of biological dependency between genes can be detected in array studies, overcoming the high levels of noise attributable to other sources (Bolstad *et al.*, 2003).

Enrichment analysis (EA) is a conventional approach for gene class inference. EA involves determining if a gene class is 'enriched' for changed genes, disproportionately to the overall fraction of genes showing change. Given a set of k statistically interesting genes, and biological class annotations, 2×2 tables are constructed for each class, as in Table 1. The count n_{11} is the number of genes changing in class b. One-sided Fisher Exact tests for each class are performed to determine if n_{11} is statistically larger than expected, i.e. than in relative proportion to the genes changing overall. If so, the class is deemed interesting. Curtis *et al.* (2005) reviewed heuristic approaches for inference on GO categories, such as Enrichment Analysis and Gene Set Enrichment Analysis, given the results of univariate gene detection analysis (Gold

et al., 2007).

Barry *et al.* (2005) detailed a method to make inferences on a priori gene category assignments by permutation testing, to produce false discovery rates of expression change for each category. Battacharjee *et al.* (2004) developed an hierarchical Bayesian method to study gene expression across organs, allowing for uncertainty in gene detection and enrichment of GO categories. Van Der Laan *et al.* (2001) used a parametric bootstrap to perform gene subset selection accounting for covariance between the genes. Dobra *et al.* (2004) developed a method to learn about gene pathways graphs which can include prior information. Lu *et al.* (2005) lists a systematic approach to multivariate gene detection with Hotelling's T^2 statistic. Pan (2006) performed stratified gene detection given GO biological process annotations. Zou and Hastie (2005) developed Elastic Nets, similar to LASSO, for variable selection and inference of gene pathways. Liao *et al.* (2007) developed Network Component Analysis for learning about gene pathway structure in multivariate data. A very specific application is offered in Sun *et al.* (2006), demonstrating a novel Bayesian approach to integrate data on transcriptional binding affinity with expression array data. Moloshok *et al.* (2002) performed Bayesian Decomposition, a pattern recognition algorithm that allows genes to cluster, and borrow strength, in multiple expression patterns associated with historical pathways. Historical pathway information is integrated into the analysis by defining *a priori* expression patterns that the genes that are likely to show across the samples. The prior distribution specifies the probability that each gene will show one or more of the patterns. This method can be awkward to employ, as one must translate the historical pathway information into expression patterns. Parmigiani *et al.* (2002) employed gene-wise borrowing to learn about tumor subclasses in cancer. They define a mixture model to classify each gene

in each tumor sample as normally expressed, up, or down regulated. The gene specific parameters are allowed to borrow strength across the whole genome. They did not use prior information, although it would not be difficult to extend their model to use prior information.

The problems with many of the aforementioned methods is that inference is limited to the genes or the pathways, but not to both. Historical pathways are typically not integrated explicitly into the analysis, and conclusions concerning pathways are made apart from the individual genes. For example, in EA, a heuristic approach is taken by first selecting a list genes for a contrast of interest and then given the list, then choosing a set of pathways. Historical pathways are ignored during gene detection. Integrating historical pathways with array data is a part of learning, something investigators do naturally. A Bayesian hierarchical linear model is introduced, Bayesian Learning for Microarrays (BLM), to flexibly allow for many sources of uncertainty and integrate prior information of gene biology, to the extent that it fits with one's beliefs. The advantages can be applied in many settings, to integrate information from many sources, while accounting for the fundamental limitations of the data, namely high noise. The Bayesian approach offers: (1) results that are interpretable in a higher-level sense concerning pathways or regulatory processes, (2) pooling information between genes to improve detection and (3) computational methods that, despite practical concerns, are quite achievable for routine analysis. We show that a fully Bayesian construction with prior information, is feasible and practical for regular high-throughput genomics analysis.

II.3. Bayesian Learning for Microarrays

The Bayesian Learning for Microarray (BLM) class of models are a series of extensions of the multistage-linear model, designed to model dependency in gene expression from historical pathways. Dependence is induced through a hierarchical structure, allowing class-level, or in terms of gene pathways, higher-level inference. The basic model is outlined for the case that the historical pathway information is assumed to be complete. Extensions of the basic model are presented, allowing for more uncertainty in the historical information.

II.3.1. BLM1

In this section, we develop the model assuming that the gene classes are defined from complete historical pathway information. Suppose we have microarray measurements of normalized log transformed gene expression for $i = 1, ..., N$ genes, across $j = 1, ..., J$ treatment groups or experimental factors, and $k = 1, ..., K$; replicate arrays for each treatment. Also, assume that we have available information about p historical gene pathways. Denoted the normalized log transformed fluorescent intensity for gene i in group j and replicate k as Y_{ijk} and let X be an experimental design matrix for one gene. The Bayesian multistage linear model, called Bayesian Learning for Microarrays 1 (BLM1), takes the $J \cdot K$ dimensional vector of gene expression measurements for gene i, $Y_i = (Y_{i11}, Y_{i12}, ..., Y_{iJK})^T$ as iid normal

$$
\begin{aligned}
Y_i &\sim N\left(X\beta_i, \sigma_i^2 I_{J\cdot K}\right) \\
\beta_{ij} &\sim N\left(Z_i\theta_j, \omega\sigma_i^2\right) \\
\sigma_i^2 &\sim \text{IG}\left(\frac{\gamma_{1i}}{2}, \frac{\gamma_{2i}}{2}\right) \\
\theta_{cj} &\sim N\left(\theta_0, \Omega_0\right) \qquad (2.1)
\end{aligned}
$$

with mean vector $X\beta_i$, and diagonal covariance matrix $\sigma_i^2 I_{J\cdot K}$, for the $J \cdot K$ identity matrix $I_{J\cdot K}$. The $(J\cdot K)\times J$ known full rank design matrix X includes as columns the J experimental factor or treatment covariates. The coefficients β_{ij}'s are assumed to be unknown and modeled as *a priori* iid normal with mean depending linearly on the p dimensional hyperparameter vector $\theta_j = (\theta_{j1}, ..., \theta_{jp})^T$ through the inner product with the i^{th} row of the $N \times p$ dimensional connectivity matrix Z, denoted Z_i. We explicitly take as given, that for a collection of p historical pathways, p gene 'classes' may be defined. The connectivity matrix Z incorporates this information into our model. Z is defined such that $Z_{ic} = 0$ if there is no historical evidence that gene i is a member of pathway c, and $Z_{ic} \neq 0$ otherwise. For example, one may define $Z_{ic} = 1$ if gene i is known to be up-regulated in pathway c, $Z_{ic} = -1$ if gene i is down-regulated and $Z_{ic} = 0$ otherwise. The p dimensional hyperparamter vector $\theta_j = (\theta_{j1}, ..., \theta_{jp})^T$ is included as a latent variable, to impose hierarchical dependence between the β_{ij}'s across the genes. The unknown latent hyperparamters θ_{cj}'s are assumed to be iid normal with mean θ_0 and variance Ω_0. In this framework, the hyperprior parameter θ_o is a prior guess at the level of change expected on average across the genes. Unlike covariance estimation, there is no sample-size limitation to the number of dependencies that may be defined between the genes within a class, and a gene may be a member of more than one class. All that we require is that Z be full rank. For now

we assume that Z is fully known. This assumption is relaxed in the following sections.

One can vary the level of borrowing between the genes by the global parameter ω. The prior weight ω in (2.1) is essentially the desired prior inverse sample size. Increasing ω places more weight on the data. Decreasing ω will increase borrowing between genes, reducing the posterior variance in β while increasing the bias. Striking an effective balance between bias and variance can facilitate improvements in sensitivity and specificity for detection. An intuitive way of thinking about the model, is as a multi-stage linear model, where we are regressing Y on the the design matrix X at the first stage, and at the second stage we are regressing the unknown coefficients β on the annotations.

Box 1.1. Summary of $N \times p$ connectivity matrix Z

- i^{th} row of Z links the treatment effects, β_i, in gene i to the hyperparameter vector θ_j
- c^{th} column of Z links θ_j to the $j = 1, ..., J$ treatment effects across all genes (weighted by Z)
- The relative direction of change in log gene expression gene depends on sign of Z
- Magnitude of Z accounts for strength of association
- Averaging of probes mapping to same gene is natural through Z

An extension of the above model, leading more conveniently to inference on the gene

classes, treats θ_{jc} as arising from a mixture,

$$\begin{aligned} Y_i &\sim N\left(X\beta_i, \sigma_i^2\right) \\ \beta_{ij} &\sim N\left(Z_i\theta_j, \omega\sigma_i^2\right) \\ \sigma_i^2 &\sim \text{IG}\left(\frac{\gamma_{1i}}{2}, \frac{\gamma_{2i}}{2}\right) \\ \theta_{jc} &\sim \pi_{\theta_{jc}} N\left(\theta_0, \Omega_0\right) \cdot I(\theta_{jc} > 0) + (1 - \pi_{\theta_{jc}}) \cdot 1_{\{0\}} \\ \pi_{\theta_{jc}} &\sim \text{Beta}(\upsilon_{1jc}, \upsilon_{2jc}), \end{aligned} \tag{2.2}$$

where $\pi_{\theta_{jc}}$ is the probability that H_o: pathway c given treatment j is unactivated. Bayesian false discovery rate estimation (bFDR) is one approach to make inference on genes or gene classes in (2.2). More on this is discussed in Section 3.5.

II.3.2. BLM2

In models (2.1) and (2.2) Z is assumed fully known. However, with some sources of prior information there remains uncertainty as to the relative direction of change in gene expression between experimental treatment conditions, and consequently uncertainty in the sign of Z_{ic}. This is the case for example with GO annotations, as there is little information to suggest up or down regulation of gene expression given an experimental stimulus, and consequently a positive or negative sign on Z_{ic}. In this situation, gene-pathway associations may be represented in Z in dichotomous form, $Z_{ic} \in \{0, 1\}$, or as weights representing the degree of belief that a particular gene is involved in a given pathway or process, e.g. based on evidence codes provided by GO. The extended model with a truncated normal prior for θ_j, added as an identifiability constraint, is

$$
\begin{aligned}
Y_i &\sim N\left(X\beta_i, \sigma_i^2\right) \\
\beta_{ij} &\sim \pi_{ij} N\left(Z_i\theta_j, \omega\sigma_i^2\right) + (1-\pi_{ij}) N\left(-Z_i\theta_j, \omega\sigma_i^2\right) \\
\sigma_i^2 &\sim \text{IG}\left(\frac{\gamma_{1i}}{2}, \frac{\gamma_{2i}}{2}\right) \\
\pi_{ij} &\sim \text{Beta}(\nu_{1ij}, \nu_{2ij}) \\
\theta_{jc} &\sim N\left(\theta_0, \Omega_0\right) \cdot I(\theta_{jc} > 0).
\end{aligned} \tag{2.3}
$$

The treatment effects, β_{ij}'s, are modeled as arising from a mixture with mean alternating in sign between the components. This model is useful for learning about changes in genes within gene classes defined from historical pathways for which little information is available concerning the direction of fold change in gene expression given an experimental stimulus. The corresponding mixture model in θ is

$$
\begin{aligned}
Y_i &\sim N\left(X\beta_i, \sigma_i^2\right) \\
\beta_{ij} &\sim \pi_{ij} N\left(Z_i\theta_j, \omega\sigma_i^2\right) + (1-\pi_{ij}) N\left(-Z_i\theta_j, \omega\sigma_i^2\right) \\
\sigma_i^2 &\sim \text{IG}\left(\frac{\gamma_{1i}}{2}, \frac{\gamma_{2i}}{2}\right) \\
\pi_{ij} &\sim \text{Beta}(\nu_{1ij}, \nu_{2ij}) \\
\theta_{jc} &\sim \pi_{\theta_{jc}} N\left(\theta_0, \Omega_0\right) \cdot I(\theta_{jc} > 0) + (1-\pi_{\theta_{jc}}) \cdot 1_{\{0\}} \\
\pi_{\theta_{jc}} &\sim \text{Beta}(\upsilon_{1jc}, \upsilon_{2jc}).
\end{aligned} \tag{2.4}
$$

II.3.3. BLM3

BLM1 and BLM2 treat the genes within a class as dependent, i.e. the whole class of genes is either activated or deactivated in tandem. This assumption is unrealistic in the following sense. In complex organisms, with incomplete pathway information, some genes might be observed to change in an important pathway, although it is unrealistic to expect all the genes to show the same activity. The changes observed do not verify or nullify the *a priori* gene dependencies. The observed changes must, very rationally argued, be accepted as uniquely observed given the specific treatment conditions studied, and not expected to generalize without further evidence. Models BLM1 and BLM2 enforce dependency between all the genes in a respective class, regardless of whether or not every gene actually exhibits treatment effect(s). This is fine if there is very strong prior information to suggest that the genes are linked, i.e. strong dependency. This may be the case, for example in lower organisms when studying transcriptional regulation. In the broader class of experiments, with weaker forms of evidence about gene dependence, the net effect of strictly imposing hierarchical dependence is to shrink genes effects that are essentially zero away from zero, and gene effects that are non-zero toward zero. It seems counterintuitive to borrow strength between genes, some of which are changing and others that are not. BLM1 and BLM2 offer advantages, but nevertheless drawbacks as well, bringing us to the next extension. A new random variable is introduced, $\Psi_{ij} \in \{1, 2, 3\}$, to account for the states of genes expression. In the following model,

$$
\begin{aligned}
Y_i|X\beta_i,\sigma_i^2 &\sim N\left(X\beta_i,\sigma_i^2\right) \\
\beta_{ij}|\Psi_{ij},Z\theta_j,\omega\sigma_i^2 &\sim \begin{cases} N\left(-Z_i\theta_j,\omega\sigma_i^2\right) & \Psi_{ij}=1 \\ 1_{\{\beta_{ij}=0\}} & \Psi_{ij}=2 \\ N\left(Z_i\theta_j,\omega\sigma_i^2\right) & \Psi_{ij}=3 \end{cases} \\
\pi(\Psi_{ij}) &\sim \text{Multinom}(3;p_1,p_2,p_3) \\
\underline{p} &\sim \text{Dir}(c,A) \\
\sigma_i^2 &\sim \text{IG}\left(\frac{\gamma_{1i}}{2},\frac{\gamma_{2i}}{2}\right) \\
\theta_{jc} &\sim N\left(\theta_0,\Omega_{0cj}\right)\cdot I(\theta_{jc}>0). \qquad (2.5)
\end{aligned}
$$

$\Psi_{ij}=1$ if gene i is down-regulated given treatment j, $\Psi_{ij}=2$ if gene i is unchanged and $\Psi_{ij}=3$ is expression in gene i is increased. Borrowing strength is conditional upon detection. The treatment effects β_{ij} depend on Z only if differential expression is detected. If a gene has no treatment effect, then the model does not borrow over that gene. In that sense, the elements of the *posterior* connectivity matrix are implicitly updated as zeros in the case that a gene is not found to change. We are in essence defining a new random variable Z' and allowing

$$
Z'_{ic}=\begin{cases} -1 & p_1 \\ 0 & p_2 \\ 1 & p_3 \end{cases},
$$

for all classes c, updating the respective probabilities $\underline{p}$ given the data. The motivation for this innovation is that in factorial designs it is possible to learn under which conditions genes in a historical pathway interact, although the entire pathway may not be activated/deactivated. The pathway may be disrupted. This level of uncer-

tainty has not been exploited in gene expression inference. We take caution, as this feature of the model does not imply unchanged genes are un-associated with members of their class. High-throughput study designs are incomplete. It is unrealistic to expect to obtain a complete and general picture of the connectivity matrix.

Here is makes sense to model θ_j with an informative prior, as it accounts for the event that a gene has actually changed. In this regard, the posterior will of course depend on the number of genes changed. In order to maintain prior 'strength' for sharing, the prior variance of θ may be inversely weighted by the number of β's $\neq 0$. For example, one may set $\theta_0 = 1$ corresponding to prior belief that a 2-fold change is *a priori* real, and set

$$\Omega_{ocj} = \sigma_\theta^2 \times \frac{1}{\sum_{i \in A_c}^{n} 1_{\beta_{ij} \neq 0}} \tag{2.6}$$

where σ_θ^2 is based on a prior belief of the support of θ when just one gene has changed, and A_c is the index set of genes annotated with class $c = 1, ..., p$. It may be advantageous to select a region of high support where it is strongly believed that effects are real. For example, setting $\sigma_\theta^2 = .25^2$ produces a prior such that $\Phi(\log_2(1.5); 1, .25^2) = 0.0465$, putting little prior mass on fold changes less than 1.5.

Information within classes is shared in a natural way between genes. This is the model chosen for the general analysis, letting us borrow information between the genes that are changing while discriminating from those that are not changing. Only genes in the same class borrow information. This may be useful when a small group of genes within the same biological class are changing but at levels too low to be detected individually.

The concept of using a connectivity matrix Z to model gene dependency is not new, Liao *et al.* (2007). BLM extends the idea, in a Bayesian framework, by allowing for uncertainty in Z. BLM shares similarities with Bayesian Decomposition of Moloshok *et al.* (2002). Like Bayesian Decomposition, the genes are allowed to borrow with prior information, but only in fixed experimental groups. Bayesian Decomposition can discover expression patterns. This is very powerful for grouping genes in, for instance, time course array studies. BLM is not designed to detect patterns, but rather gene classes given fixed experimental covariates. Another major difference in BLM is the way in which is accounts for the uncertainty in historical pathways, either in the direction of fold change, or the event that a gene is differentially expressed. BLM allows genes to flexibly share information in experimental groups when differentially expressed, rather than limited to distinct patterns. Both methods can powerfully detect changes in gene classes through borrowing, although they have different strengths and weaknesses. BLM is more flexible in the way it allows genes in the same class to borrow information across experimental groups, whereas Bayesian Decomposition is more flexible in the patterns that it may detect. BLM3 also shares similarities with the mixture model of Parmigiani *et al.* (2002). They let genes borrow information at the whole genenomic level to detect tumor subclasses, but did not make use of prior information. It would not be difficult to extend their model to include prior information. If one is willing to define each tumor sample as a unique experimental factor, then their model can be viewed as an adaptation of BLM3 without prior information.

II.3.4. Posterior Computation

Gibbs sampling (Geman and Geman, 1984) for hierarchical mixture models is described in Gelman *et al.* (2004). Block sampling, Wilkenson *et al.* (2002), is helpful for posterior sampling when the number of genes is large. The data augmentation approach of Hodges (1998) is convenient if the number of genes is moderate. Lindley and Smith (1972) derived the marginal posterior distributions for the 3rd stage parameters, θ's, integrating out the β's and Bayes Factor can be applied as discussed in Raftery *et al.* (1997). George and McCulloch (1993) discuss variable selection. Simulations and model fitting were performed with R scripts. The full conditionals are listed in the Appendix.

II.3.5. False Discovery Analysis

Consider the model in BLM1 (2.1) and let the null hypothesis be for change in gene i and treatment effect j be $H_0^{(ij)} : \beta_{ij} = 0$ and the alternative $H_a^{(ij)} : \beta_{ij} \neq 0$. Define the posterior rejection region as

$$P(|\beta_{ij}| > \gamma \ \mid \ y) > (1 - \alpha),$$

for some user specified γ and α. Let the variable $r_\gamma^{(ij)}(y) = 1$ if $H_a^{(ij)}$ is chosen, i.e. β_{ij} selected for change, and $r_\gamma^{(ij)}(y) = 0$ otherwise. In order to estimate the $FDR(\gamma)$, a measure is needed of the posterior probability $P(\beta_{ij} > \gamma|y)$, in the case that $H_0^{(ij)}$ is true, or in the Bayesian framework, a measure of $P(H_0^{(ij)}|r_\gamma^{(ij)}(y) = 1, y)$.

In the typical setting, inference on the gene 'classes' involves making probability statements on the θ_{jc}'s, i.e. $P(\delta_1 < \theta_{jc} < \delta_2)$. Posterior inference on θ in the case of

BLM2 (2.3) is complicated by the modeling constraint that $\theta > 0$. An approach to determine if a process or pathway is interesting, is to compare the posterior distribution of θ actually observed, with what would have been observed if none of the genes in the class had changed. There are a number of ways discussed in the literature to accomplish this, such as bootstrap resampling or model simulation (Tadesse *et al.*, 2004). For example, one may fit model (2.1) to derive posterior samples of the gene-wise parameter effects, β_{i2}'s and θ_{jc}'s, by Gibbs sampling. In a repeat analysis, the model is fit to a simulated data set of a transformed response, e.g. subtracting the fitted treatment effects $\beta_{i2}^{(t)}$ from y_i, $\tilde{y}_i = Y_i - X^T\beta^{(t)}$, for each Gibbs replicate t, to generate new posteriors for the $\tilde{\theta}_{jc}$'s with the estimated cancer effects removed, $\pi_o(\tilde{\theta}_{jc}|\text{Data})$. Simulations reveal that this approach can approximate the null case quite reasonably, see Figure 3.

Consider the gene classes $c = 1, ..., M$ and suppose that a class c is selected for change if

$$P(|\theta_c| > \gamma \ \mid \ y) > (1 - \alpha),$$

for different user specified γ's, and α presumed fixed. Let the variable $r_\gamma^c(y) = 1$ if class c is selected and $r_\gamma^c(y) = 0$ otherwise. Under the null hypothesis $H_0^{(c)}$ in BLM1, none of the genes in pathway b are differentially expressed. The alternative hypothesis is that all genes show at least some differential expression. In the case that the $H_0^{(c)}$ is true, the posterior mode of $(\theta_c|H_0^{(c)}, y)$ will be at (or in practice very near) 0. For those classes for which the $H_0^{(c)}$ is rejected, one can determine

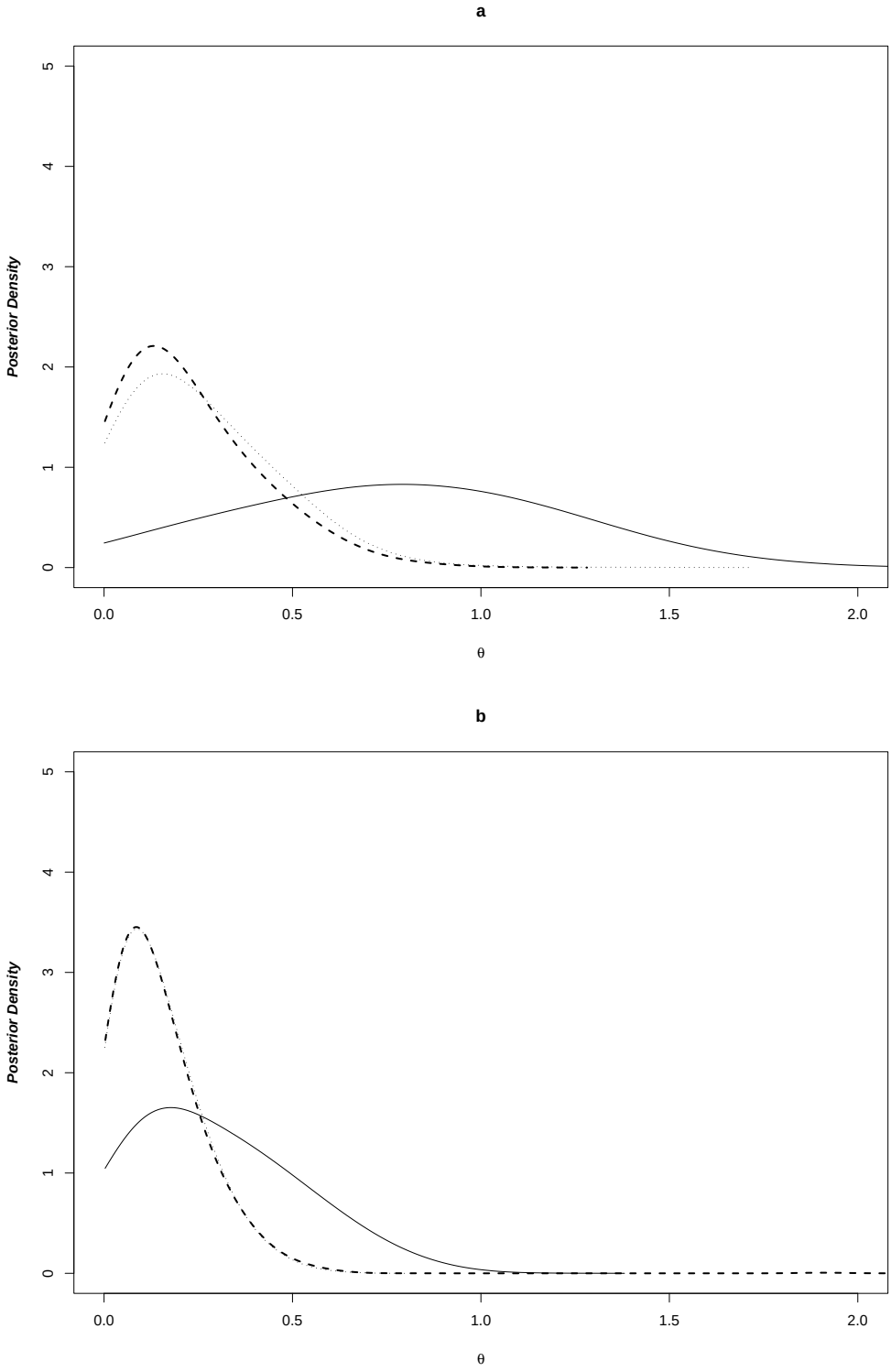

Fig. 3 Simulated posterior results, kernel fits to 1000 Gibbs samples, $(- - -)$ null, $(\cdots)$ simulated null and (—) true differences, for (a) $n = 10$ and (b) $n = 25$ genes. Expression values were simulated with 8 normals as *iid* $N(9, 1)$ and 9 cancers as *iid* $N(9 + \delta, 1)$ for δ *iid* $N(0, 1)$.

$$P(|\theta_c - \tilde{m}_c| > \gamma \ \mid \ y),$$

for $\tilde{m}$, the mode of $\theta|y$. The False Discovery Rate (FDR) is defined as

$$FDR(\gamma) = \frac{\sum_c P(|\theta_c - \tilde{m}_c| > \gamma \ \mid \ r_\gamma^c(y) = 1, y) \cdot r_\gamma^c(y)}{\sum_c r_\gamma^c(y)}, \tag{2.7}$$

the posterior probability that $|\theta_c|$ is greater than γ, given $H_0^{(c)}$. The True Negative Rate (TNR) is defined as

$$TNR(\gamma) = \frac{\sum_c P(|\theta_c - \tilde{m}_c| < \gamma \ \mid \ r_\gamma^c(y) = 0, y) \cdot (1 - r_\gamma^c(y))}{\sum_c (1 - r_\gamma^c(y))}. \tag{2.8}$$

Both the FDR and the TNR yield information about the respective decision rules. In order to achieve a given $FDR(\gamma)$, γ may be selected, and a respective list of pathways chosen. Choices of γ that are very conservative, i.e. very large in this case, tend to yield low FDR's and high TNR's. For practical purposes, a particular value of γ may be choose to yield an acceptable combination of the FDR and TNR pair. Consideration of both rates is feasible, and straightforward to estimate in a Bayesian paradigm.

Notice though, that the FDR is a mean, a rate over different gene classes. The rate can vary, although this variability is often ignored or not mentioned in the literature. A more amenable solution for gene class detection may be to consider the entire distribution of

$$P(|\theta_c - \tilde{m}_c| > \gamma \ \mid \ y),$$

for example the standard deviation and the percentiles, over all classes $c = 1, ..., M$. Dividing the standard deviation by the appropriate number of classes, for each γ, yields an approximate confidence region. The local FDR for gene class c, $lFDR_c$, is defined as

$$lFDR_c = FDR(\gamma*) \quad \text{for} \quad \gamma* = \text{argmax}_\gamma r^c_\gamma(y) = 1. \tag{2.9}$$

which may be reported for each class, with error bounds.

Suppose that we model θ by (2.2) as arising from a mixture of a truncated normal and a point mass at 0. An approach for FDR estimation attributable to Whittemore (2006), described as the Bayesian FDR ($bFDR$), is defined as

$$bFDR(\gamma) = \frac{\sum_c P(H^c_0 | r^c_\gamma(y) = 1, y) \cdot r^c_\gamma(y)}{\sum_c r^c_\gamma(y)}. \tag{2.10}$$

This definition of FDR integrates the posterior probability of H^c_0 over all gene classes that test positive for change.

In the case of BLM3, there is uncertainty as to which genes, if any, changed within a class, as there is no strict requirement that all of the genes changed in a class, even if a pathway is activated/deactivated. Consider the gene-wise rejection rule $r^\gamma_{ij}(y) = 1$ if $\max\{Pr(\beta_{ij} > 0|\text{y}), Pr(\beta_{ij} < 0|y)\} > \gamma$ for $\gamma \in [0, 1]$. An estimate of the $FDR_{ij}(\gamma)$ for gene i given treatment j, according to the rejection region specified by γ, is

$$FDR_{ij}(\gamma) = \frac{\sum_i P(\beta_{ij} = 0 \ \mid \ r^\gamma_i(y) = 1) \cdot r^\gamma_i(y)}{\sum_i r^\gamma_i(y)}. \tag{2.11}$$

This definition of FDR integrates the false positive rate over the posterior parameter

space of genes that test positive for change. An estimate of the local FDR, $lFDR_{ij}$ for gene i and treatment is

$$lFDR_{ij} = FDR(\max\{Pr(\beta_{ij} > 0 \mid y), Pr(\beta_{ij} < 0 \mid y)\}). \tag{2.12}$$

Also of interest may be some percentile p of $P(\beta_i = 0 | r_i^\gamma(y) = 1)$ over $\{i : r_i^\gamma(y) = 1\}$ or the standard deviation. These quantities are straight-forward and not cumbersome to compute in the Bayesian paradigm. The TNR is computed similarly. The choice of γ may be selected striking a desired balance between FDR and TNR.

Suppose one wishes to make inference on gene class c, under the modeling assumptions of BLM3. There are many considerations when selecting a class in this context, and the decision of how to proceed will guided by investigators' needs. One recommendation is to consider classes of genes such that the % of genes discovered for change is high, at some pre-specified FDR. According to the null hypothesis, $H_0^{(c)}$: none of the genes in class c are differentially expression, while under the alternative hypothesis, $H_1^{(c)}$: at least some genes are differentially expressed. In each class c the posterior probability of the number of genes, n_c, showing differential expression, may be used to define the rejection rule, $r_\gamma^c(Y) = 1$ for

$$\pi(n^{(c)} \geq n_\gamma \mid Y) \geq (1 - \alpha)$$

and fail to reject otherwise. This approach has been considered for gene subset selection (Bhattacharjee *et al.*, 2004). For the rejected pathways, the probability of a false positive may be estimated by

$$FDR_c(\gamma) = \pi(n_0^{(c)} \geq n_\gamma \mid Y) \tag{2.13}$$

where the posterior of $n_0^{(c)}$ is derived by repeating the modeling on a transformed data set, with the gene-wise effects for treatment removed. There are a number of different approaches proposed in the literature for simulating from the null distribution, by simulation modeling (Tadesse, 2004) or by bootstrap (Van der Laan M. and Bryan, 2001) versions of the data. These methods make certain assumption about the distribution of the data which may not always be realistic. An alternative, which in some situations may be more robust is the plug in estimator, described below, where gene wise effects, estimated by least squares ANOVA or by simple averages, are subtracted from the treatment groups in order to remove treatment differences from the data. In small data sets, the plug in approach can bias the FDR up, due to over fitting, which will induce small changes between the groups, although in practice the residual variation from the estimated treatment effects will often be overwhelmed by the effects of noise and outliers in the data.

In the context of BLM1 and BLM2, the hierarchical parameter θ embodies information which serves to inform us about changes within the respective gene classes. Collective inference on the genes and pathways with BLM3 is possible by considering first inference on θ followed by gene selection. The hierarchical approach to learning is a reasonable course. Alternatively, the joint posterior of $(n_c, \beta_{1j}, \beta_{2j}, ...|y)$ may be considered. Gene classes and genes may be selected collectively according to the joint posterior, defined by a threshold. A disadvantage to this approach is that it selects genes in classes with a high proportion of genes changing. Many investigators will find this disturbing. It is not considered further here, although such an approach may

supplement to one-at-a-time gene analysis.

II.4. Simulation Studies

In order to compare the effectiveness of the model to borrow strength, simulation studies were conducted under different conditions. Several simulation studies were conducted as part of a broader effort to understand the operating characteristics of models BLM1, BLM2 and BLM3. Intuitively, borrowing strength should increase the power to detect real differences in the data, reducing the threshold for detecting small but consistent changes. In practice, the benefits can be minimal depending on the sample and gene class sizes.

In the first simulation, data was generated for $n = 25$, 50 and 100 genes and $m = 17$ samples, with $m_0 = 8$ simulated normal samples and $m_1 = 9$ simulated disease samples. The data was simulated from a $Y_{ij} \sim N(\mu_{ij}, 1)$ iid distribution, with $\mu_{i1} = 0$ for $i = 1, ..., n$ and $\mu_{i2} = \delta$ for $\delta = .58$, .75 and 1. The results were compared for different levels of borrowing, $\omega = 1, 10$. At each iteration, the posterior probability of $\beta > 0$, the coefficient responsible for change, was averaged across the genes. This was repeated 100 times. The simulations were run with R scripts in parallel on a Dell 8200 server with dual core Intel Xeon processors, in under 24 hours.

Table 2 shows the results for BLM1. Notice that the rates of detection increase as delta increases and ω falls and the level of borrowing increases. The effect of increasing the sample size is largely to increase the precision.

In Table 3 the FDR's and FNR's are displayed, estimated as the average rates

Table 2. Simulation BLM1 (2.1), $P(\beta > 0|y)$ mean and sd

		$\omega = 10$	
n	$\delta = 0.58$	$\delta = 0.75$	$\delta = 1.00$
25	0.7993 (0.0420)	0.8677 (0.0349)	0.9313 (0.0212)
50	0.8069 (0.0318)	0.8682 (0.0252)	0.9306 (0.0154)
100	0.8051 (0.0208)	0.8697 (0.0172)	0.9288 (0.0116)
		$\omega = 1$	
n	$\delta = 0.58$	$\delta = 0.75$	$\delta = 1.00$
25	0.8460 (0.0411)	0.8991 (0.0279)	0.9593 (0.0160)
50	0.8420 (0.0297)	0.9044 (0.0223)	0.9578 (0.0128)
100	0.8454 (0.0199)	0.9041 (0.0147)	0.9581 (0.0083)
n		p-value	
50	0.1470 (0.2603)	0.0847 (0.1948)	0.0337 (0.1101)

Table 3. Simulation BLM1 (2.1), FDR/FNR

		$\omega = 10$	
n	$\delta = 0.58$	$\delta = 0.75$	$\delta = 1.00$
25	0.0355/0.6787	0.0321/0.7052	0.0214/0.7465
50	0.0348/0.6601	0.0296/0.7094	0.0224/0.7618
100	0.0358/0.6666	0.0299/0.7055	0.0218/0.7589
		$\omega = 1$	
n	$\delta = 0.58$	$\delta = 0.75$	$\delta = 1.00$
25	0.0377/0.7060	0.0314/0.7517	0.0184/0.7983
50	0.0362/0.7143	0.0290/0.7408	0.0189/0.7909
100	0.0368/0.7126	0.0293/0.7500	0.0189/0.7965

across 100 simulated data sets. Of course in this simulation, all of the genes change, so FDR's closer to zero and FNR's near 1 are better. As one would hope, the effect of increasing the level of borrowing is to reduce the FDR, while increasing the FNR.

The simulation was repeated for BLM2 (2.3), with δ alternating in sign independently at random with probability 1/2 for each gene. The results are summarized in Tables 4 and 5. Detection rates do increase with borrowing, although there is not a clear advantage in producing improved FDR's and FNR's with borrowing, until ω is reduced to 0.25. The additional level of uncertainty, i.e. in the sign of Z, does lead to additional variability in the results, which are manifested in the FDR's and FNR's.

Table 4. Simulation BLM2 (2.3), $P(\beta > 0|y)$ mean and sd

		$\omega = 10$	
n	$\delta = 0.58$	$\delta = 0.75$	$\delta = 1.00$
25	0.8043 (0.0425)	0.8577 (0.0386)	0.9273 (0.0226)
50	0.7951 (0.0283)	0.8652 (0.0244)	0.9234 (0.0193)
100	0.8005 (0.0241)	0.8597 (0.0186)	0.9245 (0.0136)
		$\omega = 1$	
n	$\delta = 0.58$	$\delta = 0.75$	$\delta = 1.00$
25	0.7964 (0.0529)	0.8571 (0.0405)	0.9228 (0.0271)
50	0.8055 (0.0248)	0.8656 (0.0255)	0.9268 (0.0147)
100	0.7987 (0.0208)	0.8630 (0.0177)	0.9265 (0.0118)
		$\omega = .25$	
n	$\delta = 0.58$	$\delta = 0.75$	$\delta = 1.00$
25	0.7983 (0.0448)	0.8833 (0.0373)	0.9592 (0.0173)
50	0.8030 (0.0320)	0.8816 (0.0263)	0.9587 (0.0140)
100	0.8027 (0.0223)	0.8875 (0.0188)	0.9599 (0.0106)
n		p-value	
50	0.2873 (0.5130)	0.1690 (0.3879)	0.0672 (0.2208)

Table 5. Simulation BLM2 (2.3), FDR/FNR

		$\omega = 10$	
n	$\delta = 0.58$	$\delta = 0.75$	$\delta = 1.00$
25	0.0353/0.7292	0.0305/0.7459	0.0233/0.7720
50	0.0352/0.7278	0.0300/0.7418	0.0237/0.7646
100	0.0365/0.7312	0.0305/0.7485	0.0225/0.7682
		$\omega = 1$	
n	$\delta = 0.58$	$\delta = 0.75$	$\delta = 1.00$
25	0.0348/0.7321	0.0301/0.7471	0.0225/0.7701
50	0.0362/0.7347	0.0301/0.7496	0.0234/0.76954
100	0.0368/0.7304	0.0306/0.7457	0.0229/0.7755
		$\omega = .25$	
n	$\delta = 0.58$	$\delta = 0.75$	$\delta = 1.00$
25	0.0474/0.7354	0.0360/0.7610	0.0175/0.7899
50	0.0499/0.7409	0.0348/0.7623	0.0173/0.7835
100	0.0490/0.7438	0.0345/0.7644	0.0169/0.7880

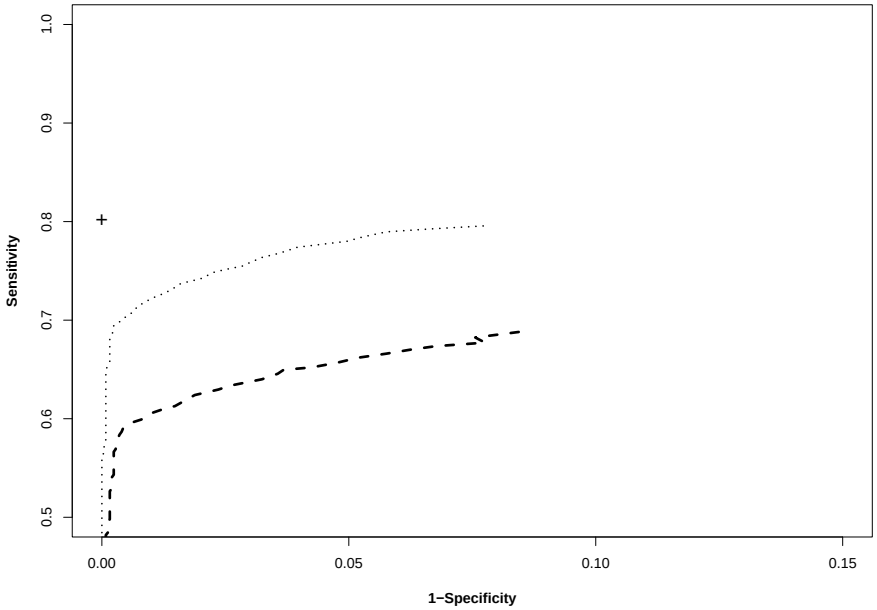

Fig. 4 ROC curve BLM3: $+$ $\omega = 1$, $\cdots$ $\omega = 10$, $---$ $\omega = 100$.

In order to assess BLM3, data was generated for $n = 50$ genes and $m = 17$ samples, and two types, with $m_0 = 8$ simulated normal samples and $m_1 = 9$ simulated disease samples. The data was simulated as iid normal $Y_{ij} \sim N(\mu_{ij}, .25^2)$ with $\mu_{i1} = 0$ for $i = 1, ..., 50$ and $\mu_{i2} = 0$ for $i = 1, ..., 25$ and $\mu_{i2} = i/50$ for $i = 26, ..., 50$. The results for model (2.5) were compared for different levels of borrowing strength through the parameter $\omega = 1$, 10 and 100. The results are summarized in Figure 4, by ROC curves, averaged over 100 iterations.

The results show that a certain level of borrowing does lead to better estimation. For $\omega = 1$ perfect specificity was achieved in all 100 simulations, although the sensitivity varied. For $\omega = 10$ and 100, the ROC curve does rise as the Specificity improves. The Sensitivity appears to taper out to 80% for $\omega = 10$. Notice that at lower ω, the

same level of sensitivity can be achieved at an improved specificity. Each ω ultimately can achieve the same sensitivity at near 80%, although the effect of borrowing is to improve the specificity.

II.5. Yeast Time Course Data

This microarray study was conducted by Klevecz *et al.* (2004) with Affymetrix S98 Arrays gene expression arrays. Oscillations in transcription were monitored through 3 complete cycles of respiration and reduction by dissolved oxygen (DO). Every 4 minutes RNA samples were collected and hybridized to arrays. Two striking global patterns in gene expression were observed, a shift in expression between time points 1-10 and 11-32, and a periodic trend resembling a sine wave. The data was originally preprocessed by Affymetrix Microarray Analysis Suite 5.0 (MAS 5.0). The MAS 5.0 expression values were obtained, and transformed by the base 2 logarithm. The data may be found at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2583. Gene classes were obtained from the Yeastract public database, according to transcription factor (TF), http://www.yeastract.com. Seventeen gene classes were obtained, according to the consensus motifs of 17 TF's, the smallest class including 5 genes and the largest 222 genes.

A post-hoc analysis was carried out. BLM2 (2.3) and (2.4) were fit to each TF gene class independently. A design matrix was constructed including two terms, accounting for a shift in expression at time point 10, and the sine wave. The FDR was estimated for each class variable θ_c, for $c = 1, ..., 17$, from the results. The rejection rule $r_\gamma^{cj}(y) = 1$ for each TF c and factor j was defined as $P(\theta_{cj} > \gamma \mid y) > 0.90$. The $lFDR$ is reported for each TF class, along with the $lTNR$. The results are

Table 6. BLM2 (2.3) $\omega = 10$, FDR/TNR

	TF	*Shift*	*Sine Wave*
1.	Dal82	1/0.1976	1/0.1866
2.	FLO8	0.922/0.3524	0.9205/0.3015
3.	GAT1	1/0.1976	1/0.1866
4.	HAA1	0.922/0.3524	0.9205/0.3015
5.	HAP2	1/0.1976	1/0.1866
6.	HPC2	1/0.1976	1/0.1866
7.	MOT3	1/0.1976	1/0.1866
8.	NDT80	1/0.1976	1/0.1866
9.	NRG1	1/0.1976	1/0.1866
10.	PHO4	1/0.1976	0.9205/0.3015
11.	RDR1	1/0.1976	1/0.1866
12.	SMP1	1/0.1976	0.9205/0.3015
13.	THI2	1/0.1976	1/0.1866
14.	CBF1	1/0.1976	1/0.1866
15.	GCR2	1/0.1976	1/0.1866
16.	HAP5	1/0.1976	1/0.1866
17.	MET31	1/0.1976	0.9205/0.3015

Table 7. BLM2 (2.3) $\omega = 1$, FDR/TNR

	TF	*Shift*	*Sine Wave*
1.	Dal82	0.8642/0.1354	0.9359/0.1465
2.	FLO8	0.7051/0.2032	0.9011/0.2532
3.	GAT1	0.5385/0.2608	0.9359/0.1465
4.	HAA1	0.2871/0.3843	0.9151/0.2198
5.	HAP2	0.0206/0.7952	0.9574/0.1177
6.	HPC2	0.8642/0.1354	0.9359/0.1465
7.	MOT3	0.7051/0.2032	0.9359/0.1465
8.	NDT80	0.0553/0.6488	0.9574/0.1177
9.	NRG1	0.1758/0.4496	0.9574/0.1177
10.	PHO4	0.9320/0.1199	0.9151/0.2198
11.	RDR1	0.7793/0.1905	0.9574/0.1177
12.	SMP1	0.0170/0.8874	0.9151/0.2198
13.	THI2	0.0013/0.9030	0.9359/0.1465
14.	CBF1	0.0188/0.8566	0.9574/0.1177
15.	GCR2	0.0206/0.7952	0.9574/0.1177
16.	HAP5	0.5385/0.2608	0.9574/0.1177
17.	MET31	0.7051/0.2032	0.9359/0.1465

summarized in Tables 6−7 by class labeled according to respective TF, for $\omega = 1, 10$.

Notice that for $\omega = 10$, none of the TF classes are detected for change, for either factor. When $\omega = 1$ is employed, gene classes defined for TF's HAP2, SIMP1, THI2, CBF1 are GCR2 are discovered for change as a class by the shift term at the $FDR = .05$ level.

The results for BLM2 (2.4), treating θ as arising from a mixture are displayed in Tables 8−9. The effect of increasing the level of borrowing is to increase the detection rates for similar gene classes. The rejection rule for each TF c and factor j, $r_\gamma^{cj}(y) = 1$, is $P(\theta_{cj} > 0|Y) >= \gamma$. The local $bFDR$ rate reported below is $\min_\gamma(FDR(\gamma)|r_\gamma^{jc}(y) = 1)$. The local $bFNR$ rate is also reported.

As with BLM2 (2.3), inference on the TF classes improves for an increase in borrowing between the genes. For $\omega = 1$ and $bFDR \leq= .05$, the following TF gene classes are detected: HAP2, NDT80, NRG1, THI2, CBF1 and GCR2. The analysis is not without limitations. Only genes for which information is available are included in the analysis. There may be genes in transcriptional pathways that exhibit periodicity attributable to DO levels. It would be a mistake to make stronger conclusions than warranted, beyond the information given.

II.6. Renal Cell Carcinoma

In the next application, we turn to a microarray study of Renal clear cell carcinoma (RCC). Let us begin with some background of RCC. RCC is a deadly and complex disease. The American Cancer Society expects about 38,890 newly diagnosed

Table 8. BLM2 (2.4) $\omega = 10$, FDR/TNR

	TF	*Shift*	*Sin Wave*
1.	Dal82	0.5215/0.4427	0.4836/0.4913
2.	FLO8	0.5178/0.4659	0.4836/0.4913
3.	GAT1	0.5073/0.4741	0.4715/0.5005
4.	HAA1	0.5178/0.4659	0.4337/0.5121
5.	HAP2	0.5073/0.4741	0.4609/0.5042
6.	HPC2	0.5178/0.4659	0.4715/0.5005
7.	MOT3	0.5178/0.4659	0.4836/0.4913
8.	NDT80	0.5073/0.4741	0.4545/0.5097
9.	NRG1	0.5178/0.4659	0.4147/0.558
10.	PHO4	0.5193/0.4508	0.4715/0.5005
11.	RDR1	0.5073/0.4741	0.4545/0.5097
12.	SMP1	0.5073/0.4741	0.4759/0.4941
13.	THI2	0.5012/—	0.4759/0.4941
14.	CBF1	0.5012/—	0.4836/0.4913
15.	GCR2	0.5012/—	0.4609/0.5042
16.	HAP5	0.5178/0.4659	0.4715/0.5005
17.	MET31	0.5178/0.4659	0.4545/0.5097

Table 9. BLM2 (2.4) $\omega = 1$, FDR/TNR

	TF	*Shift*	*Sine Wave*
1.	Dal82	0.241/0.5041	0.398/0.5443
2.	FLO8	0.2565/0.4985	0.398/0.5443
3.	GAT1	0.174/0.5598	0.3901/0.5475
4.	HAA1	0.1923/0.5493	0.3563/0.5755
5.	HAP2	0.0057/0.7131	0.3436/0.5796
6.	HPC2	0.2236/0.5231	0.33/0.5891
7.	MOT3	0.2085/0.5379	0.4055/0.5385
8.	NDT80	0.0057/0.7131	0.3193/0.5914
9.	NRG1	0.0408/0.6204	0.3645/0.5677
10.	PHO4	0.2716/0.4863	0.4055/0.5385
11.	RDR1	0.1316/0.5818	0.3859/0.558
12.	SMP1	0.0759/0.6025	0.3859/0.558
13.	THI2	0.0126/0.6475	0.3859/0.558
14.	CBF1	0.0088/0.6929	0.3645/0.5677
15.	GCR2	0.0088/0.6929	0.336/0.5844
16.	HAP5	0.1068/0.5914	0.3859/0.558
17.	MET31	0.1536/0.5709	0.3563/0.575

cases of renal clear cell carcinoma (RCC) this year, with approximately 12,840 of these cases expected to end in death. Surgery is currently the primary treatment for RCC as many existing therapies have poor prognosis. There is a critical need for improved clinical investigations into therapies for renal cell carcinoma. Attempts to combat RCC would benefit greatly from improvements to the list of candidate genes associated with the disease. Many past microarray studies have failed to identify effective targets for treatment, although more promising results were shown by Lenburg *et al.* (2003), who compared normal renal to renal tumor gene expression on Affymetrix U133 chips. Identifying effective targets for treatment in high-throughput experiments such as Lenburg *et al.* microarray study is typically complicated by the uncertainty in the gene regulatory pathways, i.e. collections of genes that interact, responsible for cancer.

Lenburg *et al.* (2003) compared the results from their own analysis with seven previous microarray experiments to identify gene candidates associated with the multistep process of renal carcinogenesis. The original analysis included univariate gene detection. The authors identified 1,234 genes changing by > 3 fold. Among the up regulated genes, more than expected were found to be associated with functional gene classes: hypoxia, angiogenesis, necrosis factor, apoptosis, interferon, drug resistance and metastasis, by Fishers Exact Test. While offering notable results, their analysis did not make explicit use of historical pathways associated with tumorgenesis.

The MAS 5.0 expression values were obtained for the 17 patient samples available from the authors' supplementary website. The data for all samples on array were included in the analysis, normalized by setting the mean to 500 on each array as described in Lenburg et al., truncating expression values below by 1 and taking the

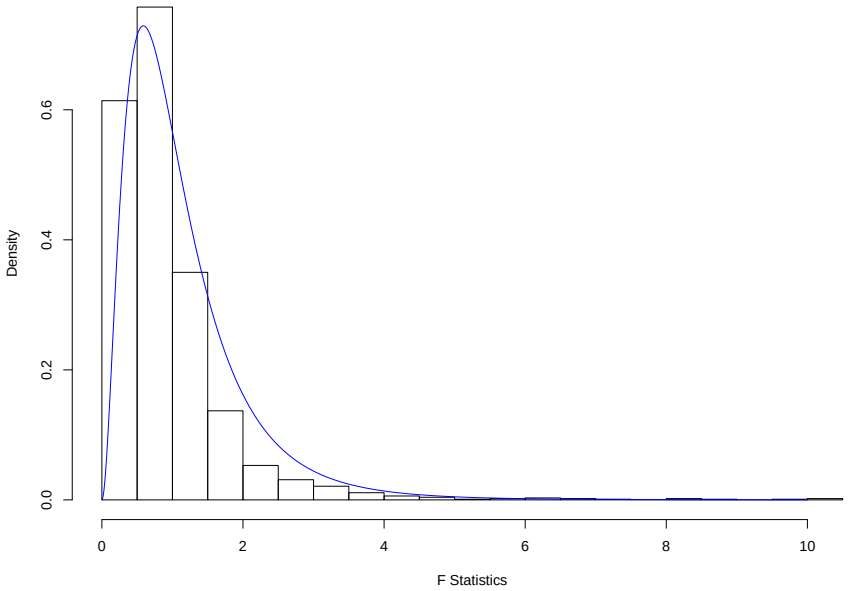

Fig. 5 Histogram of F statistics and (—) null, for testing variation due to patient.

base 2 logarithm. The original design was match paired, although one of the normal samples was discarded for poor quality. Likelihood ratio tests were performed on the log ratios, testing the variation by patient account for normal cancer differences. The distribution of the F statistics matched the theoretical null, see Figure 5.

Biocarta pathway annotations were obtained from the **D**atabase for **A**nnotation, **V**isualization, and **I**ntegrated **D**iscovery: DAVID (Dennis *et al.*, 2003), downloaded in September of 2006. The gene classes were defined from pathways of between 20-100 probes, 199 in total, as the smaller gene classes have relatively low detection power. To account for multiple probe sets mapping to the same Unigene cluster, these probes were equally down weighted by the terms in the adjacency matrix Z.

Gene expression was modeled by BLM3 for each Biocarta pathway separately. This seemed reasonable, and efficient, as few pathways shared many genes. Across the

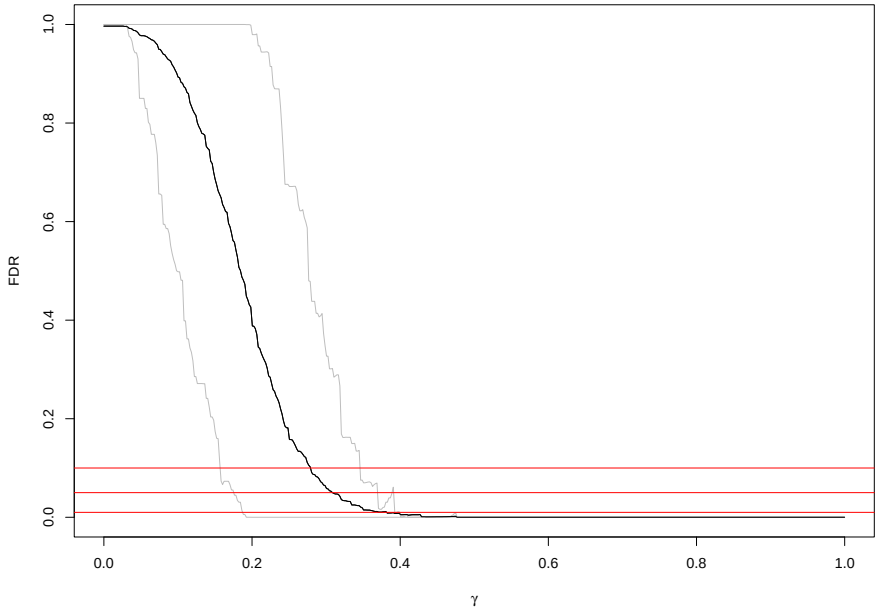

Fig. 6 FDR, with 90% C.I. bounds, $\omega = 1$. Horizontal lines at 0.1, 0.05 and 0.01.

arrays, the trend in the sample standard deviations of the respective probes was fit against their means with a regression spline. The fitted mean and standard deviation values were input as prior parameters for the intercept and scale in BLM3 in an Empirical Bayes manner. The effective prior sample size was $n = 5$. The FDR for each pathway was derived by reestimating the percent of genes changing, by (2.14), after removing the treatment effects with a plug in estimator. FDR's were derived for each Biocarta pathway with $W = 10, 1$. The FDR's are plotted against the γ cutoff regions for $W = 1$ in Figure 6. Notice the bands positioned at the 90% intervals about the FDR's.

For larger FDR's, the bands are wide apart, at $FDR = 1\%$ (0, 7%) and at $FDR = 5\%$, (0,30.12%) . The wide bounds on the FDR tend to be overlooked, although here it is quite apparent that the bounds are informative. Due to the wide bounds

Table 10. BLM3 $\omega = 10$, $FDR \leq .0001$

	Biocarta Pathway
1.	The 4-1BB-dependent immune response
2.	Oxidative Stress Induced Gene Expression Via Nrf2
3.	ATM Signaling Pathway
4.	CBL mediated ligand-induced downregulation of EGF receptors
5.	Cadmium induces DNA synthesis and proliferation in macrophages
6.	Repression of Pain Sensation by the Transcriptional Regulator DREAM
7.	Erythropoietin mediated neuroprotection through NF-kB
8.	METS affect on Macrophage Differentiation
9.	Free Radical Induced Apoptosis
10.	Inhibition of Cellular Proliferation by Gleevec
11.	Adhesion and Diapedesis of Granulocytes
12.	Segmentation Clock
13.	IGF-1 Signaling Pathway
14.	Insulin Signaling Pathway
15.	Internal Ribosome entry pathway
16.	Adhesion and Diapedesis of Lymphocytes
17.	Role of MEF2D in T-cell Apoptosis
18.	Acetylation and Deacetylation of RelA in The Nucleus
19.	TNF/Stress Related Signaling
20.	TPO Signaling Pathway

on the FDR's, gene classes were selected meeting $FDR \leq 0.0001$. A smaller value of ω produced more pathways for change: 20 for $\omega = 10$ and 37 for $\omega = 1$. These are listed in Tables 10−11. If in fact the algorithm is operating as the simulation study in Section 3.6 suggests, increasing borrowing improves specificity and sensitivity.

Among the signaling pathways detected as significant are those associated with genes CXCR4, D4-GDI, EGF, EPO, IGF-1, IL1R, IL 6, PDGF, T cell receptors and TNF. Many of these genes have been implicated or are known to be associated with cancer. For example, CXCR4 has been shown to have a pivotal role in cancer (Schrader *et al.*, 2002; Arya *et al.*, 2007), D4-GDI is a Rho GDP inhibitor, regulating breast cancer cell invasive activities (Zhang and Zhang 2006). EGF is known to regulate growth and metastasis in tumors and PDGF regulates autocrine stimulation of cancer cells (George, 2003). EPO promotes red blood cell formation, and is commonly given to cancer patients to relieve fatigue (Brower 2006). Some pathways detected are as-

Table 11. BLM3 $\omega = 1$, $FDR \leq .0001$

	Biocarta Pathway
1.	The 4-1BB-dependent immune response
2.	Agrin in Postsynaptic Differentiation
3.	Effects of calcineurin in Keratinocyte Differentiation
4.	CBL mediated ligand-induced downregulation of EGF receptors
5.	CD40L Signaling Pathway
6.	Cadmium induces DNA synthesis and proliferation in macrophages
7.	Cyclins and Cell Cycle Regulation
8.	CXCR4 Signaling Pathway
9.	D4-GDI Signaling Pathway
10.	The role of FYVE-finger proteins in vesicle transport
11.	EGF Signaling Pathway
12.	Eukaryotic protein translation
13.	EPO Signaling Pathway
14.	METS affect on Macrophage Differentiation
15.	Ghrelin
16.	Segmentation Clock
17.	Hypoxia-Inducible Factor in the Cardiovascular System
18.	IGF-1 Signaling Pathway
19.	Signal transduction through IL1R
20.	IL 6 signaling pathway
21.	Keratinocyte Differentiation
22.	The IGF-1 Receptor and Longevity
23.	Endocytotic role of NDK
24.	Ras-Independent pathway in NK cell-mediated cytotoxicity
25.	NFkB activation by Nontypeable Hemophilus influenzae
26.	PDGF Signaling Pathway
27.	Phosphoinositides and their downstream targets
28.	Influence of Ras and Rho proteins on G1 to S Transition
29.	Bone Remodelling
30.	Acetylation and Deacetylation of RelA in The Nucleus
31.	Sprouty regulation of tyrosine kinase signals
32.	Stathmin and breast cancer resistance to antimicrotubule agents
33.	TNF/Stress Related Signaling
34.	T Cell Receptor Signaling Pathway
35.	TNFR2 Signaling Pathway
36.	Toll-Like Receptor Pathway
37.	Control of Gene Expression by Vitamin D Receptor

sociated with genes known to be involved in other cancers were discovered, NFkB, RAS, as well as proliferation in macrophages. Interestingly, phosphoinositides have been studied in conjunction with progression of invasive cancers (Bertagnolo *et al.* 2007). Some of the detected pathways are associated with cell differentiation, such as Agrin in Postsynaptic Differentiation, Keratinocyte Differentiation, and METS affect on Macrophage Differentiation. Other pathways of interest are related to cell cycle or cell cycle transition.

We would like to know if BLM detects alterations in gene classes, defined from historical pathways, associated with renal carcinoma, that EA missed. Lenburg's analysis was repeated with EA, following Lenburg's approach for gene detection, first filtering out probe sets called absent by MAS 5.0, and calling a gene changed if: (1) all probe sets mapping to the gene had a geometric mean of two sample t-test p-values < 0.03, consistent with their 10 FDR calculation, and (2) a geometric mean of fold change > 3 or $< 1/3$. Fisher's Exact Test was calculated on the detected gene counts for each Biocarta pathway. The distribution of Fisher p-values was fit with $SPLOSH$ of Pounds and Cheng (2004) to estimate FDR's. None of the FDR's estimated from EA achieved the strict 0.0001 threshold for BLM. EA are listed in Table 12, ordered by FDR, for Biocarta Pathways with EA FDR's ≤ 0.01.

Ten pathways were detected for change by EA. Among these, there are some interesting pathways, related to surface adhesion, apoptosis and signaling. Only one is common to the list found by BLM3, D4-GDI Signaling Pathway, although several have processes in common related to hypoxia and EGF. Caspase activity in Apoptosis looks interesting, as well as Lymphocyte Cell Surface Molecules.

Table 12. Results EA, $FDR \leq .01$

	Biocarta Pathway	
1.	Neuroregulin receptor degredation protein-1 Controls ErbB3 receptor recycling	0.0010
2.	Actions of Nitric Oxide in the Heart	0.0011
3.	Nuclear Receptors in Lipid Metabolism and Toxicity	0.0017
4.	Caspase Cascade in Apoptosis	0.0028
5.	Catabolic Pathways for Arginine	0.0029
6.	D4-GDI Signaling Pathway	0.0040
7.	Hypoxia and p53 in the Cardiovascular system	0.0049
8.	B Lymphocyte Cell Surface Molecules	0.0057
9.	Role of EGF Receptor Transactivation by GPCRs in Cardiac Hypertrophy	0.0057
10.	Dendritic cells in regulating TH1 and TH2 Development	0.0097

II.7. Discussion

The problem gene class detection in noisy expression array high-throughput data was addressed with a Bayesian model, BLM, allowing for uncertainty at many levels with the ability to borrow information across the genes. The problem of detecting gene classes is very complex. BLM is flexible enough for many practical uses. In the case studies offered, as in simulation, borrowing of information was demonstrated to improve sensitivity and specificity. Rather than regressing genes on genes, which would increasing the number of parameters, a hierarchical scheme was adopted to impose gene-gene dependency, and borrow strength. This has the effect of reducing the effective degrees of freedom in the model (Spiegelhalter *et al.*, 2002) and the overall variation.

Several extensions of previously proposed methods designed to integrate prior information are offered here. The assumption that the connectivity matrix Z is fully known, Liao *et al.* (2007), is relaxed in a Bayesian framework by allowing for uncertainty in Z. In the special case that a complete pathway is expected to be activated, i.e. all of the genes show change, then accounting for the uncertainty in direction of

change with BLM2 may be sufficient. This assumption is relaxed further in BLM3, by allowing a subset of members of a gene class to change in an experiment. Like Bayesian Decomposition, BLM allows genes to share information, although BLM cannot discover new patterns. In the case that an experiment is run with well defined, fixed experimental factors, BLM is more flexible in how it lets genes in the same class share information. The mixture model of Parmigiani *et al.* (2002) also allows genes to borrow information, at the genomic level. It can be viewed as an adaptation of BLM3 without prior information. Also, BLM lends well to FDR and TNR analysis. The FDR is a rate, and as such tends to underestimate the variability in false detection for choosing genes or gene classes, as it is averaged across the genes. In advanced applications with RCC we observed this variability, knowledge of which was informative for selecting a threshold.

The main focus of attention here was on detecting historical pathways. We were interested in how historical information could improve feature selection, and how to make inferences on *a priori* gene classes. The heuristic approach is to consider gene and gene class selection separately. This is a reasonable approach to take, and in practice can work well. We propose a more flexible framework, to make explicit use of historical pathways. The methods offered here, although displaying some clear advantages, should not be regarded as replacing but rather supplementing a one-at-a-time gene analysis. In the simulations and case studies here, BLM was run in R on a Dell 8200 server with dual core Intel Xeon processors. Posterior simulation was completed in a reasonable amount of time, and could be implemented during regular analysis.

Pooling information between genes allows BLM to detect gene classes for treatment effects on arrays that traditionally may be considered uninteresting. Overall, it gives

the biologist focus and direction toward historical pathways and genes targets. Often during analysis, the focus is on the number of genes or size of changes within a pathway or ontology. However, in molecular biology, even when only one gene is changed, the knowledge of that gene (its function and its weight of importance within a specific pathway) is just as significant. These results suggest that there might be more to learn from high-throughput experiments than we might have expected, if we are careful to consider the fundamental limitations in the data and historical knowledge.

CHAPTER III

BAYESIAN CHANGE POINT ANALYSIS FOR BAC AND ACGH HIGH-THROUGHPUT ARRAYS

III.1. Introduction

Cytogenetics, the study of chromosome structure and anomaly, has long been recognized as important to the study of tumor development (Lengauer *et al.*, 1998). Chromosomal aberrations are characterized, at least in cancer, as chromosomal rearrangements, deletions, or amplifications selected over time that can evolve throughout tumor progression and or invasion. A variety of chromosomal rearrangements have been linked to cancer in patient populations, that might otherwise have been thought of as the same pathological disorder. Cataloging genetic aberrations in and between unhealthy subjects is a goal of personalized medicine, and is considered critical for developing strategies to treat highly diverse forms of cancers.

The mechanisms that govern chromosomal aberration are not well understood and new high-throughput technologies are helping to improve our understanding of the associations between chromosome structure and disease. One difficulty in modeling high-throughput human chromosomal data is that abnormalities can be subject specific and therefore disease populations are heterogeneous. Improvements to medical diagnostic and prognostic decision making, through informatics obtained at the whole genome level, would enable medicine to take this next step forward. At least this is the contention driving much of the latest research in high-throughput cytogenetics.

A Bayesian Change Point Analysis (BCPA) model is developed for borrowing strength and accounting for change point uncertainty in high-throughput aCGH experiments of heterogeneous patient populations. In simulation, BCPA shows a marginal trade off in sensitivity/specificity relative to the current standard for segmentation modeling, for detecting copy number changes across heterogeneous disease populations. The model is applied to Wilms Tumor BAC array data.

III.1.1. Advances in Cytogenetics

Advances in basic medical and clinical technologies are improving our understanding of the links between chromosomes and disease. Early work to catalog chromosomal aberration in disease used a method called karyotyping, made possible by staining and photographic techniques that allowed visualization of chromosome aberrations at low resolution. Later advances, such as Spectral Karyotyping (SKY) and Multiplex Fluorescence In Situ Hybridization (m-Fish), made it possible to visualize of all the chromosomes simultaneously in a different fluorescent color. The latest high-throughput methods, such as bacterial artificial chromosome (BAC) and oligonucleaotide array experiments, improved the resolution at which aberrations can be detected along the chromosome. BAC arrays were developed first, with much longer sequences than aCGH experiments. Since aCGH works with much shorter probe sequences than BAC arrays, the locations of chromosomal aberrations linked to disease can be detected with much better precision. The aCGH platform has been described in the literature as high-resolution for measuring regions of chromosomal copy gain or loss (Albertson *et al.* 2003). BAC and aCGH experiments tend to be similar in nature to two color fluorescent expression experiments, although the essential features involve detecting relative DNA copy gain or loss rather than relative changes in RNA expression. Self-

self aCGH designs to detect low-level genomic alterations have achieved as much as ±copy number gain of 2Mb resolution (Bilke *et al.*, 2005).

Box 2.1. Summary of high-throughput cytogenetics

- aCGH and BAC arrays are essentially two channel microarrays.
- The probes correspond to fixed locations along the chromosomes.
- DNA rather than RNA is collected and hybridized to the chips.
- aCGH and BAC arrays suffer from similar problems as expression arrays, e.g. dye bias, normalization, etc..

III.1.2. Historical Information

There is a unique body of verifiable prior information available on the human cytogenetics of many diseases. This information can be in the form of known ploidy, i.e. number of chromosome copies, or even more detailed information about copy gains or losses in specific regions, e.g. by karyotyping (Camps *et al.*, 2004). The Mitelman Molecular Biology and Clinical (MBC) Associations Searcher, maintained at the National Center for Biotechnology Information, is an expanding resource that allows investigators to record and publicly retrieve cytogenetic data on individuals with a range of diseases, so that specific genetic disorders are available for medical practitioners and pathologists. Other public databases are listed in Box 2.2.

Box 2.2. Public domain databases

- Mitelman Database (http://cgap.nci.nih.gov/Chromosomes/Mitelman)
- SKY/M-FISH & CGH Data (http://www.ncbi.nlm.nih.gov/sky/)
- Chromosomal Variation in Man (http://jws-edck.wiley.com:8096/)
- Atlas of Genetics and Cytogenetics in Oncology and Haematology (http://atlasgeneticsoncology.org/)

III.1.3. Analyzing BAC and aCGH Data

A number of different and sophisticated approaches have been proposed for analyzing BAC and aCGH data. Hidden Markov Models (HMM) were proposed (Fridlyand *et al.*, 2004) to exploit positional dependencies in clones along the chromosomal arms. A deficiency of the HMM, is that the distance between probes, within a chromosome, can vary and thus the stationarity assumption is violated. This tends to be a mild violation in practice, as the distances between the probes is generally approximately equal. More recently, a Bayesian version of HMM was introduced to make use of some of the reasonable assumptions about gain and loss (Guha, 2005). Another class of estimators treats identifying regions of gain or loss as a change point problem: Lasso (Picard *et al.*, 2005; Huang *et al.*, 2005), Gain and Loss Analysis of DNA (GLAD) (Hupe *et al.*, 2004) a Gaussian model based approach, and Circular Binary Segmentation (CBS) a nonparametric change point estimator (Olshen *et al.*, 2004). There is evidence that CBS can achieve higher sensitivity and specificity than HMM (Willenbrock and Fridlyand, 2005).

The historical methods suffer in that the results do not lend to obvious solutions for inferring aberrations. For example, CBS locates change points, given a user specified α-level for testing, although does not offer inference procedures for determining if a discovered segment has a gain or loss in copy number. Inference must be performed after the change points are determined. Other approaches, such as clustering (Wang *et al.*, 2005) and wavelets (Hsu *et al.*, 2005), have been proposed and compared although it is difficult to draw conclusions, since these methods do not offer standard ways to make statistical statements of significance (Lai *et al.*, 2005).

The main motivation for extending the basic concepts of GLAD, CBS or Lasso is to improve estimation of the rate of chromosomal aberration across heterogeneous patient populations. In diverse patient populations, the event and domain of a copy gain or deletion along a chromosome can vary between patients. Consider for example the simple case of two samples, depicted in Figure 7. Copy gain is illustrated along the chromosome in sample 1 between change points a_1 and a_2 and in sample 2 between change points b_1 and b_2. The common sub-region where both samples have a copy gain is shaded. Quantifying the probability of chromosomal amplification or deletion in common sub-regions of diverse patient populations can inform investigators of locations of conserved bio-markers.

Previously described methods do not allow borrowing of information across samples, nor offer holistic solutions for combining results in order to derive population level inferences. In populations where much is known about the locations of chromosomal aberrations, prior information, it is argued, should be useful for analysis. Such information must be integrated carefully though, into the analysis, as the change points

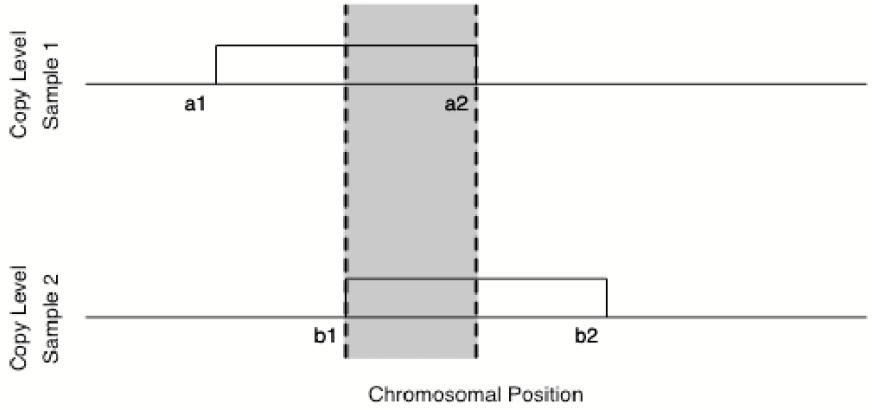

Fig. 7 Illustration of region of common gain, with uncommon change points.

and states differ between samples (Gelfand *et al.*, 1992). The goal of Bayesian Change Point Analysis (BCPA) is to provide a method capable of powerfully detecting unusual and sometimes subtle changes in the genome across diverse patient populations. This involves from a Bayesian point of view: (1) sharing information across samples in regions of common gain/loss, (2) accounting for uncertainty about the degree of similarity between patients, (3) integrating prior sources of information in the analysis, and (4) producing results easily interpreted by biologists. We describe BCPA in detail in Section 2. In section 3 we discuss posterior simulation. The results of sensitivity analysis with simulations, and application to a Wilm's Tumor dataset are provided in Section 4. In Section 5 we make concluding remarks, and give some future directions for BCPA.

III.2. Bayesian Change Point Analysis

BACs corresponds to DNA sequence of between 200,000−300,000 base pairs at known position along a chromosome. In disease samples, the chromosomes can carry extra

copies of certain BAC sequences that have been repeated, called copy gain or amplification, or deletions of the sequences, called copy loss. Human BAC array experiments are essentially two channel fluorescent microarray experiments, that are designed to measure chromosomal copy number variation by the relative fluorescent intensity between the channels. Typically, these experiments involve competitive hybridization of *cy5* labeled normal sample in the control channel and *cy3* labeled disease or tumor tissue in the experimental channel. Unlike expression arrays the samples are derived from DNA rather than mRNA. Like expression arrays, BAC arrays suffer from technological variation (Bolstad *et al.* 2003). For now, we assume that we have measurements from BAC arrays that have been normalized and log transformed. The methods outlined here may also be applied to the results of an aCGH experiment, the major difference being that the sequences, oligonucleotides, are much shorter than BACs, providing better resolution of the chromosome.

Before we discuss the model we need some notation. Let the regions of normal chromosomal copy number, and copy gain or loss, be referred to as segments along the chromosome. For modeling sake, the segments are assumed to be partitioned by change points. By definition the BACs within a contiguous segment have the same copy number. Let the BACs on an array be ordered by chromosomal position $i = 1, ..., N$ from one end of the chromosome to the other, and let ξ_{hi}'s indicate the segment k of each BAC i in sample $h = 1, ..., H$. The ξ_{hi}'s are defined as

$$\xi_{hi} = \sum_{k=1}^{M_h} k \cdot I\left(a_{h,k-1} < i \leq a_{h,k}\right)$$

given change points a_{hk} and segments $k = 1, ..., M_h$, where $I(\cdot)$ is the indicator function equal to 1 if the expression in the brackets is true and zero otherwise. The

variable $\xi_{hi} = k$ tells us that in sample h, position i belongs to the k-th segment. Thus we allow the change points and number of change points to differ between samples. Since we want to borrow strength across samples, we introduce an additional variable to indicate membership of each position i to the common sub-regions across all samples. For sub-regions $\tilde{k} = 1, ..., \tilde{M}$, we define $\tilde{\xi}_i$ analogously as

$$\tilde{\xi}_{hi} = \sum_{k=1}^{\tilde{M}} k \cdot I\left(\tilde{a}_{k-1} < i \leq \tilde{a}_k\right)$$

where the $\tilde{a}_k$ are the unique and ordered breakpoints across all samples. The motivation for including $\tilde{a}_k$ in the analysis will be come clearer ahead.

III.2.1. Likelihood

The normalized log 2 transformed fluorescent ratio in sample h at position i is denoted Y_{hi}. Given the segment membership of position i, as indicated by $\xi_{hi} = k$, is iid Gaussian

$$Y_{hi}|\mu_{hk}, \sigma^2_{hk}, \xi_{hi} = k \sim N\left(\mu_{hk}, \sigma^2_{hk}\right) \tag{3.1}$$

with mean μ_{hk} and variance σ^2_{hk}. Note that the segment means and variances are allowed to differ between samples. The prior distribution of the variance is Inverse Gamma.

$$\sigma^2_{hk} \sim \text{IG}\left(\frac{\gamma}{2}, \frac{\gamma\sigma^2_o}{2}\right) \tag{3.2}$$

where σ^2_o is a prior guess of the variance and the prior effective sample size γ.

III.2.2. Prior Mean

The prior for μ_{hk} depends on the state s of segment k, for three possible copy number states: $s = 1$ for copy number loss, $s = 2$ for normal copy and $s = 3$ for copy number gain. State is assigned by the discrete random variable $\psi_{hk} = s$, discussed in more detail below. Given $\psi_{hk} = s$, μ_{hk} is distributed as

$$\mu_{hk}|\sigma^2_{hk}, W, n_{hk}, \psi_{hk} = s \propto \exp\left\{-\frac{1}{2W\sigma^2_{hk}}\left[\sum_{\tilde{k}\in A_{hk}} n_{h\tilde{k}}(\mu_{hk} - \theta_{s,\tilde{k}})^2\right]\right\} \tag{3.3}$$

for the set $A_{hk} = \{\tilde{k} : I(\xi_{hi} = k \cap \tilde{\xi}_{hi} = \tilde{k}) = 1\}$ where $n_{h\tilde{k}}$ is the number of BACs in the $\tilde{k}$th subsegment of segment k in sample h. The hyperparameters $\theta_{s,\tilde{k}}$, defined below, are the mean levels in the subregions, $\tilde{k} = 1, ..., \tilde{M}$. Notice that the variance σ^2_{hk} is weighted by a global parameter W. The global tuning parameter W controls the level of borrowing between the samples. The discrete random variable $\psi_{hk} = s$ may also be expressed in an alternative form as a function, $\psi_{hk} = \psi_h(\xi_i = k) = s$, to indicate that for all positions i belonging to the k-th segment, the state equals s. We use this alternative notation below. The prior density of $\mu_{hk}|\cdot$ may conveniently be rexpressed as

$$\mu_{hk}|\sigma^2_{hk}, W, n_{hk}, \psi_{hk} = s \sim N\left(\eta_{hks}, \frac{W \cdot \sigma^2_{hk}}{n_{hk}}\right) \tag{3.4}$$

with

$$\eta_{hks} = \frac{\sum I\left(\psi_{hk} = s\right) I\left(a_{k-1} < \tilde{a}_{\tilde{k}} \leq a_k\right) \theta_{s\tilde{k}}}{\sum I\left(\psi_{hk} = s\right) I\left(a_{k-1} < \tilde{a}_{\tilde{k}} \leq a_k\right)} \tag{3.5}$$

a weighted average of the sub-region means, with weights proportional to the sub-

region sizes, $n_{h\tilde{k}}$. This form is useful for posterior sampling, described in the next section. In order to borrow strength across the samples, the prior mean η_{hks} of μ_{hk} is constructed as a linear function of hierarchical parameters $\theta_{s\tilde{k}}$, indexed by state s and sub-region $\tilde{k}$ across the population.

III.2.3. Hyper-prior Mean

The hierarchical population parameters $\theta_{h\tilde{k}}$'s in each sub-region $\tilde{k}$ and state $s = 1$ (deletion), $s = 2$ (normal copy) and $s = 3$ (amplification), account for the mean copy levels across the whole population. The prior densities for the θ's are, in the event of no prior information, are taken as improper

$$\begin{aligned} \pi(\theta_{1\tilde{k}}) &\equiv 1_{\{\theta_{1\tilde{k}}<0\}} \\ \pi(\theta_{2\tilde{k}}) &\equiv 1_{\{0\}} \\ \pi(\theta_{3\tilde{k}}) &\equiv 1_{\{\theta_{3\tilde{k}}>0\}} \end{aligned} \tag{3.6}$$

where 1_0 is a point mass at 0, indicating no change in the log ratios for normal copy state. This results in a posterior distribution for the hyper-mean copy level that is in the form of an truncated normal distribution. In the event that prior information is available, the priors are taken to be truncated normals

$$\begin{aligned} \pi(\theta_{1\tilde{k}}) &\equiv N(\theta_{1\tilde{k}}, \tau) \cdot I(\theta_{1\tilde{k}} < 0) \\ \pi(\theta_{2\tilde{k}}) &\equiv 1_{\{0\}} \\ \pi(\theta_{3\tilde{k}}) &\equiv N(\theta_{3\tilde{k}}, \tau) \cdot I(\theta_{3\tilde{k}} > 0). \end{aligned} \tag{3.7}$$

III.2.4. Hidden States

In reality, along a chromosomal segment within a cell from one sample there are a finite set of possible copy states corresponding to deletions, normal copy number or multiple amplifications. The exact levels of copy number are obscured in aCGH experiments, as heterogeneous cell mixtures from diseased tissue within same patient are hybridized to the experimental channel. Copy number can vary in diseased cells, and typically diseased tissue includes normal cells in varying proportions.

According to the model, we have 3 states: $s = 1$ for a copy loss, $s = 2$ for normal copy number, and $s = 3$ for a copy gain. The discrete random variable $\psi_{hk} = s$ assigns each segment k in sample h, and consequently any sub-regions overlapped by the segment, to a copy state. The prior distribution of the random variable $\psi_{hk} = s$ is assumed to be multinomial.

Posterior inference on the ψ_{hk}'s across all of the samples can provide investigators with information regarding regions of common deletion or amplification. Inference across samples between sub-regions can provide investigators with clues about associations between conserved chromosomal deletions or amplifications and ultimately to molecular sub-types of disease.

III.2.5. Change Points

The number of segments, one plus the number of change points, is assigned a truncated Poisson prior $M_h|N \sim \text{Poiss}(M_h|\beta_h) \cdot I(M_h < N)$, truncated to be less than or equal to the number of positions N. As of yet, little biological evidence suggests a strategy for specifying joint priors on the change points between samples. Ad

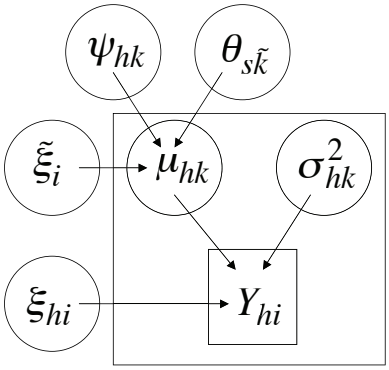

Fig. 8 Graphical model

hoc procedures for specifying informative joint priors for the break points deserves investigation. At present, all change point configurations are assigned uniform non-informative priors.

III.2.6. Graphical Summary

BCPA is depicted graphically in Figure 8. The observed log ratio Y_{hi} is assigned to segment k, by the indicator $\xi_{hi} = k$, with mean μ_{hk} and variance σ^2_{hk}. The state of segment k in sample h is assigned by the discrete random variable ψ_{hk}. The mean of μ_{hk} is a weighted average of hyperparameters $\theta_{s\tilde{k}}$, over the population level sub-regions, $\tilde{k}$'s, intersecting with segment k in sample h.

III.2.7. Posterior Simulation

The most difficult issue for fitting BCPA is developing a suitable strategy to account for uncertainty due to the change points. Given the discrete nature of the change point configurations, convergence can be very slow with adaptive methods. We return to

this issue, but for now consider the change point configuration as given. This allows us to proceed with Gibbs sampling as follows: first the segment variances and means are sampled, followed by the copy states, and lastly the sub-region population means, θ's.

III.2.8. Sampling Prior Means Variance

Given a change point configuration, the conjugacy of the model allows direct Gibbs sampling of all of the parameters. The full conditionals are listed in Appendix B.

III.2.9. Change Point Search Strategy

Several search strategies are offered to account for uncertainty in the chromosomal change point parameters. The reason for this is that the space of all possible change points is very large. Computationally, is would be very expensive to obtain the posterior densities for all possible combination of change points. Following these strategies, one can explore the marginal posterior probabilities for different change points of high posterior density. Let us begin by defining the change point configuration as the set $\underline{\xi} = \{\xi_h : h = 1, ..., H\}$. Suppose that we obtain a starting configuration $\underline{\xi}_o$ from some reasonable software. The way in which one will want to move from $\underline{\xi}_o$, to a new configuration, or between configurations, will depend on the data and the problem. In cases where the data can be very noisy, and one may want to jitter to the change points. In other cases, it may be enough to consider adding or deleting starting change points either systematically or at random. This strategy is summarized in Box 2.3.

Box 2.3. Accounting for uncertainty in change points

1. Choose reasonable starting configuration, $\underline{\xi}_o$
2. Choose a move to a new configuration $\underline{\xi}_c$ for $c = 1, ..., C$
3. Gibbs sampling of the full conditional posteriors given $\underline{\xi}_c$
4. Estimate the unnormalized marginal posterior of $\underline{\xi}$, $\tilde{\pi}\left(\underline{\xi}_c|\text{Data}\right)$ by Monte Carlo Integration
5. Numerically integrate quantities of interest over $\underline{\xi}_c$, for $c = 1, ..., C$

The unnormalized marginal posterior distribution of change point configuration $\underline{\xi}_c$ for configurations $c = 1, ..., C$, underscored to denote the variable class, may be estimated to a desired level of accuracy by Monte Carlo integration from the $t = 1, ..., T$ Gibbs samples as

$$\tilde{p}\left(\underline{\xi}_c|\text{Data}\right) \propto \sum_{t=1}^{T} \tilde{p}\left(\underline{\xi}_c, \underline{\mu}_c^{(t)}, \underline{\sigma}_c^{2(t)}, \underline{\psi}_c^{(t)}, \underline{\theta}_c^{(t)}|\text{Data}\right). \tag{3.8}$$

Here the model parameters are indexed by c, to denote dependence on $\underline{\xi}_c$. One could apply bootstrap resampling to obtain further samples, or use the $\tilde{p}\left(\underline{\xi}_c|\text{Data}\right)$'s to obtain posterior moments of other parameters in the model, accounting for uncertainty due to the change points. For example, suppose one is interested in learning about the state s of sample h at position i and let $(\bar{\psi}_{ch}(\xi_{ci})|\text{Data})$ be the conditional posterior mean state in sample h at position i over Gibbs samples $t = 1, ..., T$, given ξ_c. The conditional posterior mean of state, given all of the change point configurations considered, is estimated by

$$\bar{\bar{\psi}}_{hi} = \frac{\sum_{c=1}^{C} \bar{\psi}_{ch}(\xi_{ci})\tilde{\pi}(\xi_c|\text{Data})}{\sum_{c=1}^{C} \tilde{\pi}(\xi_c|\text{Data})}. \tag{3.9}$$

In order to survey the change point configurations, several search strategies are proposed. These may serve different purposes, depending on the problems and data sets. Search Strategy I is largely concerned with accounting for uncertainty in the locations of the change points. Search Strategies II and III are concerned with intelligently ways to search for plausible change point configurations.

• **Search Strategy I**

An initial change point configuration $\underline{\xi}_o$ is selected from a reasonable software. The change points are jittered, added or deleted at random at each iteration by user defined probabilities. A sensitivity analysis can be conducted to assess the uncertainty in selecting the right change points.

• **Search Strategy II**

Reproducible approaches for searching intelligently through combinations of change points efficiently are needed. One such approach relies on spatial hierarchical clustering. All of the BACs in each sample are initially clustered, by chromosome, with spatial hierarchical clustering. The only distinction between spatial hierarchical clustering and hierarchical clustering is that only clusters that are spatially adjacent may be combined. Once the samples are clustered, the heights $\mathcal{H}_d$ at each node d of respective dendrograms where a merge was performed, are combined and sorted in descending order, $\mathcal{H}_{(1)}, \mathcal{H}_{(2)}, ..., \mathcal{H}_{(D)}$. A starting maximum and ending minimum height are preselected from the sorted list, based on the number of initial and final groupings desired. The first change point configuration $\underline{\xi}_1$ is assembled by selecting

only those clusters separated by at least the the distance specified by the maximum height, $\mathcal{H}_{(1)}$. The next configuration $\underline{\xi}_2$ is assembled to include all of the clusters separated by at least the distance specified by the next biggest height, $\mathcal{H}_{(2)}$, and so on. In this way, the sample partitions are explored according to the relative order of the clusters, measured by the distance between clusters. Samples with less noise and more signal are partitioned earlier than samples with more noise.

• Search Strategy III

The BACs in each sample are clustered by recursive binary partitioning with a two sample t-statistic. In the first iteration, one change point is selected, dividing sample h into two segments, at the location corresponding to the smallest p-value for a two-sided t-test of the means between the segments. At each subsequent iteration, a new change point is selected, given the change points already selected, again at the location corresponding to the smallest p-value. This is repeated until sample h is divided into a user defined maximum number of segments. This approach is similar to CBS, actually a special case of CBS. In every sample, the Bayesian Information Criterion is computed for zero change points ($M_h = 1$), one change point ($M_h = 2$), etc., and the score, $S_{Mh} = BIC_{Mh}/\max(BIC_{Mh})$ is computed. All of the scores across all sample are combined in a vector and sorted. The first change point configuration $\underline{\xi}_1$ is assembled by selecting the change points in every sample meeting the best score overall, in this case with a value of 1. The process is repeated for all unique scores. In this way, attention is concentrated around the best change point configurations first.

In Search Strategy I, it is reasonable to report the mean of the marginal posterior results over the perturbed change point configurations, as the goal is to account for the uncertainty in identifying the right change points. In Search Strategies II and III

the goal is to choose, in a sense, the best prior model defined by the change point configuration with the highest marginal posterior density. These search strategies do not necessarily search more thoroughly through the higher dimensional spaces, i.e. with more segments. Averages of the marginal posterior results will be biased, under weighting the configurations with higher dimensionality, if these spaces are not searched in proportion to their size relative to the lower dimensional spaces. Therefore, in Search Strategies II and III, it better to choose the best change point configuration, and given that configuration, report posterior summaries of interest. These search strategies use the data to obtain configurations, and then evaluate the configurations with the posteriors. The data is used twice. More accurately, this is an Empirical Bayes analysis rather than a fully Bayesian analysis.

III.3. Simulation Studies

III.3.1. Simulation Study 1

Random chromosomes were simulated with the SegMix simulation model (Gaile *et al.*, 2006) in R. Abnormal segments are specified by user choice of the: (1) segment midpoint, (2) segment half width, randomly generated by a Poisson distribution, (3) segment mean copy level, which may be further perturbed by a user specified noise parameter and (4) probability of realizing an abnormal copy number in the segment, in each sample across the population. Noise about the segments is Gaussian $N(0, \sigma^2)$, with variance controlled by the user.

$H = 20$ random chromosomes were generated, of $N = 150$ equally spaced BACs,

with standard deviation $\sigma = 0.10$. Copy gains were specified at midpoints 50 and 95, and a loss at 135, with segment half-widths randomly generated from a Poisson distribution with mean 20, 10, and 5 respectively. The mean level for copy gain was set to $\mu = 0.30$, and for loss set to $\mu = -.50$. The rate of abnormal copy number was set to 0.90 for all aberrant segments.

Figure 9 shows the posterior means with 0.90 C.I.'s from BCPA to 4 of the 20 simulated chromosomes, given change points fit by DNACopy, the R library implementation of CBS (http://lib.stat.cmu.edu/R/CRAN/). DNACopy fits change points based on a user specified α-level threshold testing the likelihood ratio of recursively defined change point configurations, chosen in a binary fashion (Olshen *et al.* 2004). We set $\alpha = 0.01$.

In Figure 9, the credible bounds for the segment means reflect not only sample size variation, but also the uncertainty in copy number state. Overall, the change points fit by DNACopy for this simulated dataset were quite reasonable.

We performed sensitivity analysis to learn about change point uncertainty by perturbing the change points fit by DNACopy. We jittered the starting configuration 9,999 times according to the distribution proportional to $\exp|x - \hat{a}_{hk}|^{\nu} \times I(\hat{b}_{h(k-1)} < x \leq \hat{b}_{h(k)})$ for starting change points: $\hat{a}_{hk}$'s, and midpoints between the cuts: $\hat{b}_{hk}$'s. We sampled 2,999 configurations independently with $\nu = 2$, 3000 with $\nu = 1$, and 4000 with $\nu = .75$. For each configuration, 500 Gibbs samples were generated with C++ scripts on a Dell 8200 with dual core Intel Xeon processors. The total simulation, including all 10,000 configurations took under 48 hours. Figure 10 shows the resulting posterior probability of amplification/deletion by position with global parameter W

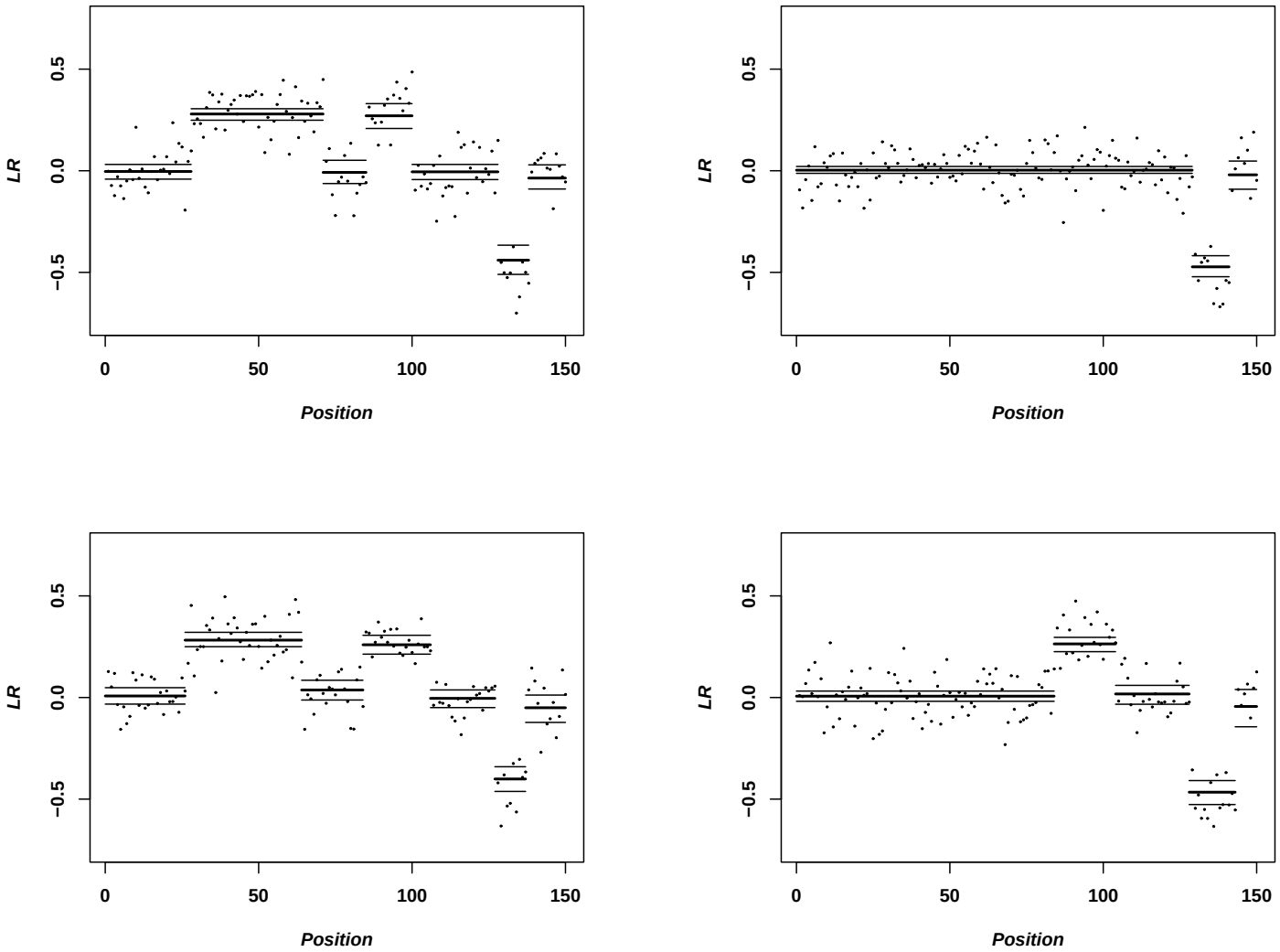

Fig. 9 Four SegMix simulated chromosomes

= 10, marginalizing over the uncertainty in the change points by (3.14) in Section 3. Accounting for the uncertainty in the exact locations of the change points induced little additional variability in the marginal posteriors of copy state. Similar results were observed with $W = 5$.

In order to inspect the behavior of posterior probabilities of amplification/deletion when a true change point is missed, we deleted the starting change points independently with probabilities .1, .2 and .5 and repeated the process 10,000 times. Deleting a true change point with probability .5 is extreme. We did so to get an idea of what would happen in such a heinous case. Figure 11 shows an increase in the variability in the posterior probability of amplification/deletion by position if an important change point is omitted. The additional variability is concentrated in the troughs between regions of common copy amplification/deletion. These results were fit with $W = 10$ although similar results were observed with $W = 5$. Figure 12 shows the results of adding change points at random to the starting values, again for 10,000 iterations. Marginalizing over the uncertainty in the additional change points induced little variability in the posterior probability of amplification/deletion. Based on our simulations, we believe that it is better to err on the side of too many starting change points rather than too few.

In order to assess the utility in borrowing strength, we decreased the simulated segment means μ_{hk}'s for amplification or deletion in the previous example to ± 0.10, and again fit starting breakpoints with DNACopy, but this time with a more liberal choice for $\alpha = 0.10$. It should be noted that the realized copy level change is below most detection criteria, especially viewed in the context of the commonly accepted rule of thumb of .225 (Nakao *et al.*, 2004).

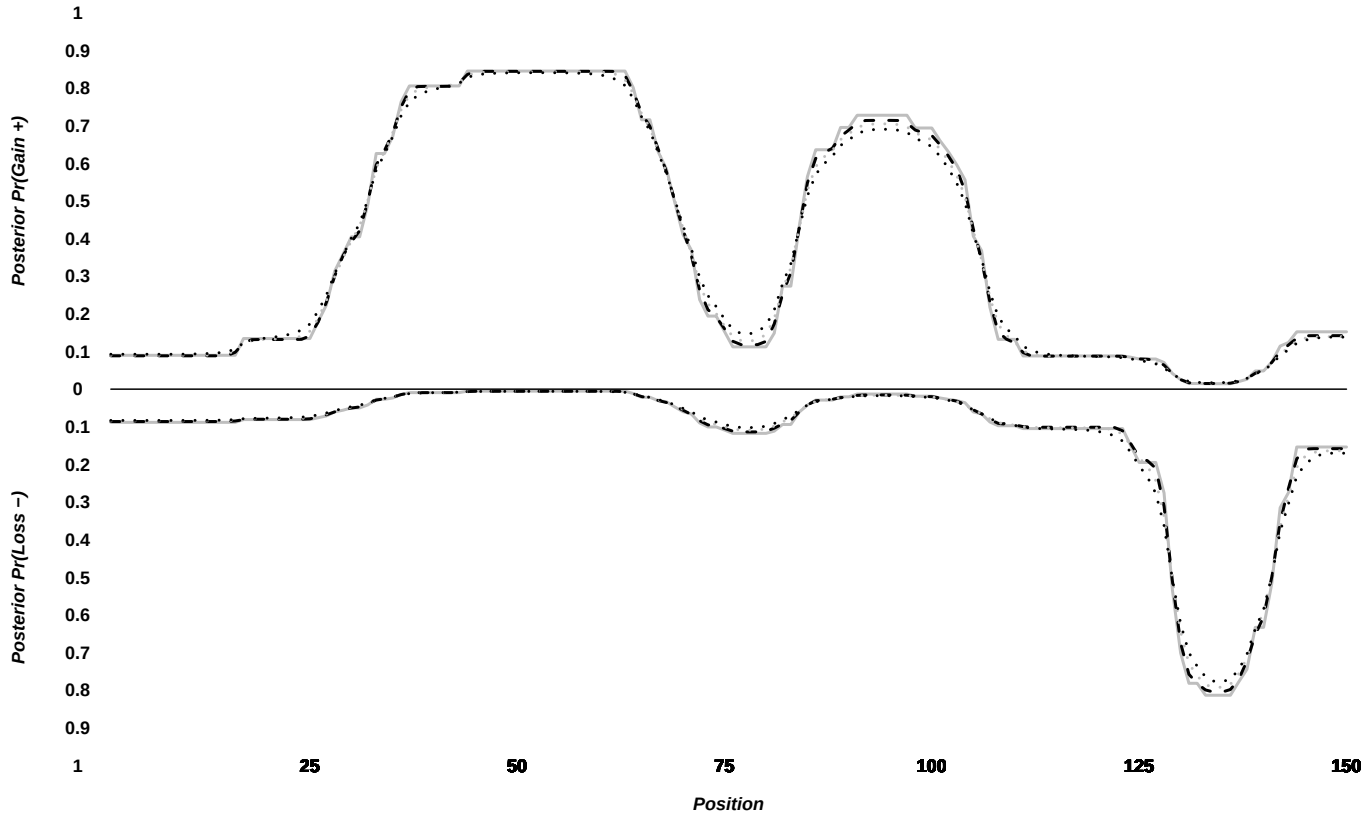

Fig. 10 Posterior copy gain/loss with change points jittered at random, 10,000 iterations: (−) starting configuration, (- -) $\nu = 2$, (**- -**) $\nu = 1$ and $(\cdots)$ $\nu = .75$.

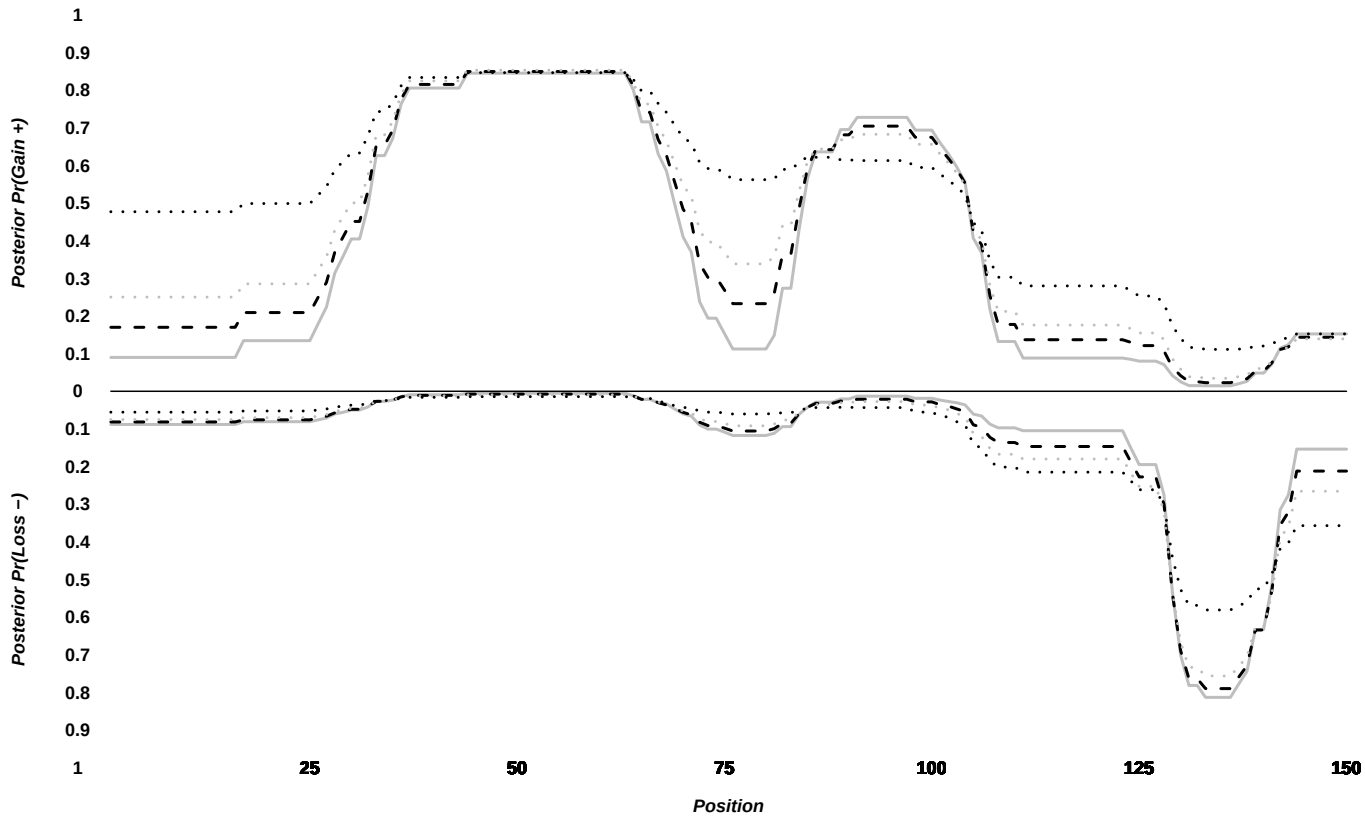

Fig. 11 Posterior copy gain/loss, change points deleted at random, 10,000 iterations: (−) starting configuration, (**- -**) deletion probability .1, (grey $\cdots$) .2 and ($\cdots$) .5.

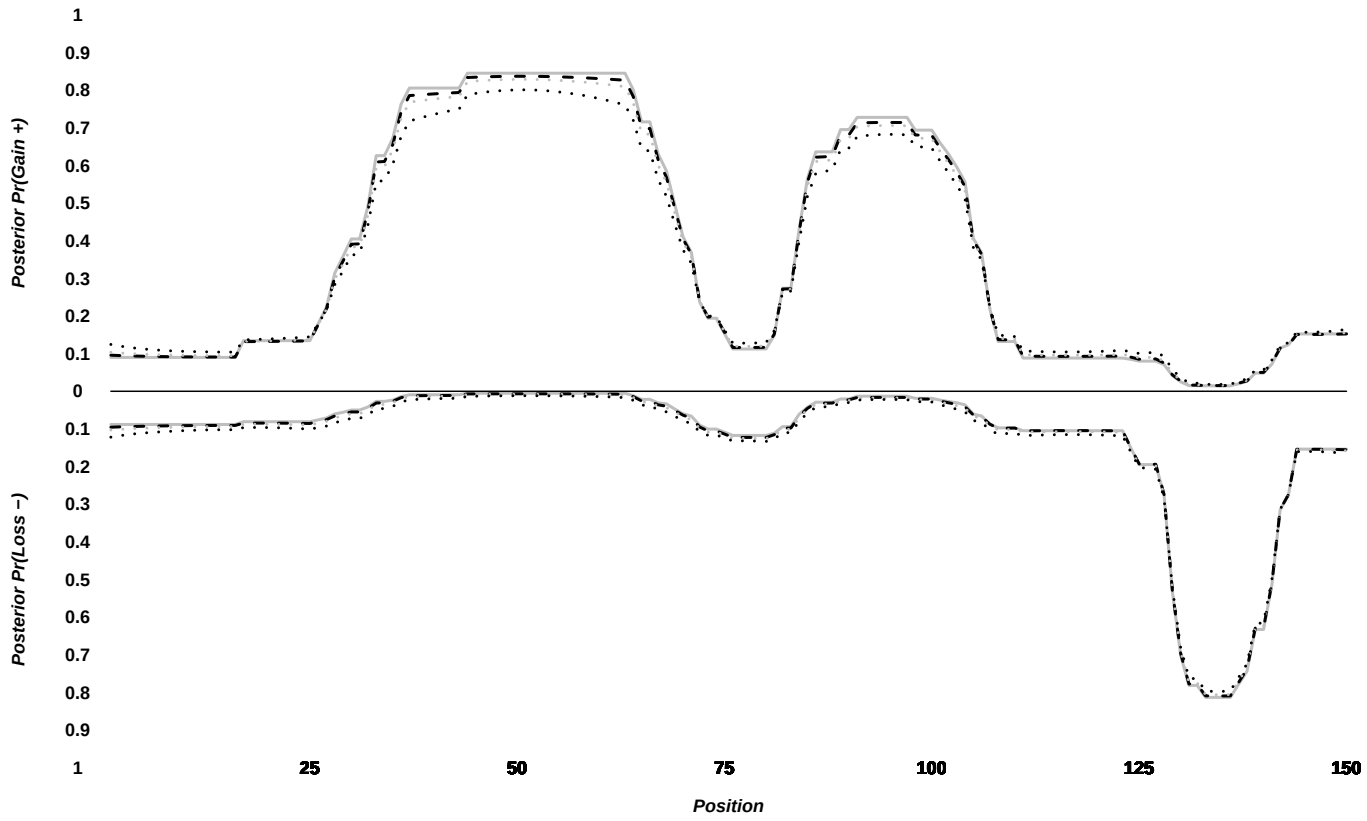

Fig. 12 Posterior copy gain/loss, change points added at random, 10,000 iterations: (−) starting configuration, (**- -**) addition probability .1, (grey $\cdots$) .2 and ($\cdots$) .5.

Figure 13 shows the posterior probability of amplification or deletion at each position for $W = 10$, 5, and 1, marginalized over 10,000 iteration of jittered starting change points. Here the low but consistent true levels of change, and domains of aberrant segments, are virtually undetectable by eye. We suspected that jittering the starting change points in this example might have a more profound effect, as it is very difficult for any algorithm to find change points accurately at such low levels of signal to noise. The real question is whether BCPA discovered biology. Fortunately, borrowing strength across the samples increases the chance of discovering regions of common copy gain/loss. Noticeably the posterior probabilities of gain/loss are less than 0.85. More samples are needed for better accuracy.

III.3.2. Simulation Study 2

A similar simulation was repeated in order to compare the sensitivity and specificity of BCPA, with borrowing, against CBS using a t-test rule. Simulated samples from SegMix were generated, $H = 100$ random chromosomes of $N = 200$ equally spaced BACS with standard deviation $\sigma = 0.10$. Copy gains were specified at midpoints 30 and 95 and 155 with Poisson half-width means of 5, 20 and 15, and segment means of $\mu = 0.25, -0.15, 0.10$. The rate of abnormal copy number was set to 90% for all aberrant segments. For each of ten simulated data sets, BCPA was run, at W = 1, using Search Strategy III. The search was conducted in R, and the change point configurations were piped to C++ where Gibbs sampling was conducted. Parallel simulations were run on a Dell 8200 with dual core Intel Xeon processors, taking under 24 hours per data set. The results were compared with CBS, run at $\alpha = 0.10$. The sensitivity and specificity were compared across all samples, between methods,

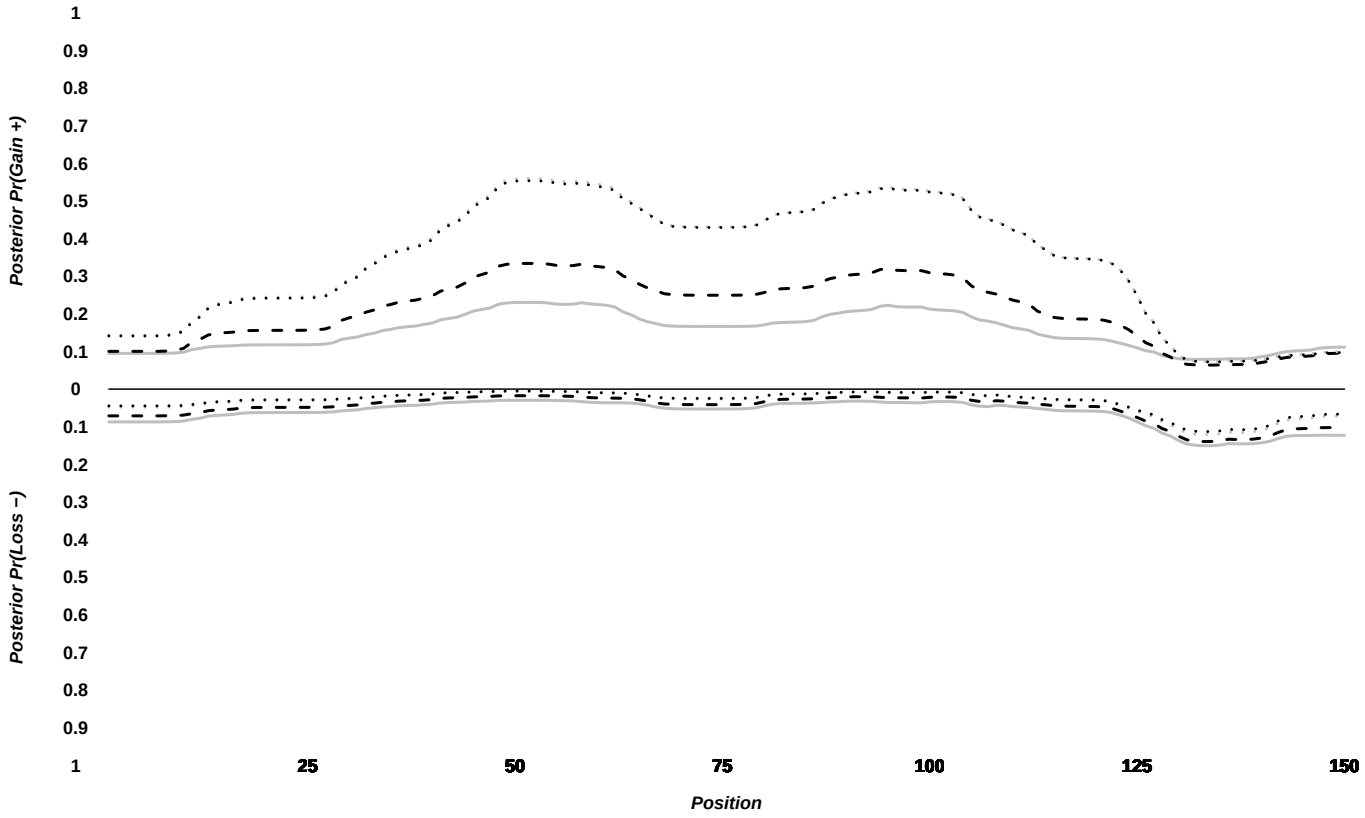

Fig. 13 Posterior copy gain/loss, reduced aberrant segment means, $\mu_{hk} = \pm 0.10$: $(-)$ $W = 10$, (- -) $W = 5$ and $(\cdots)$ $W = 1$.

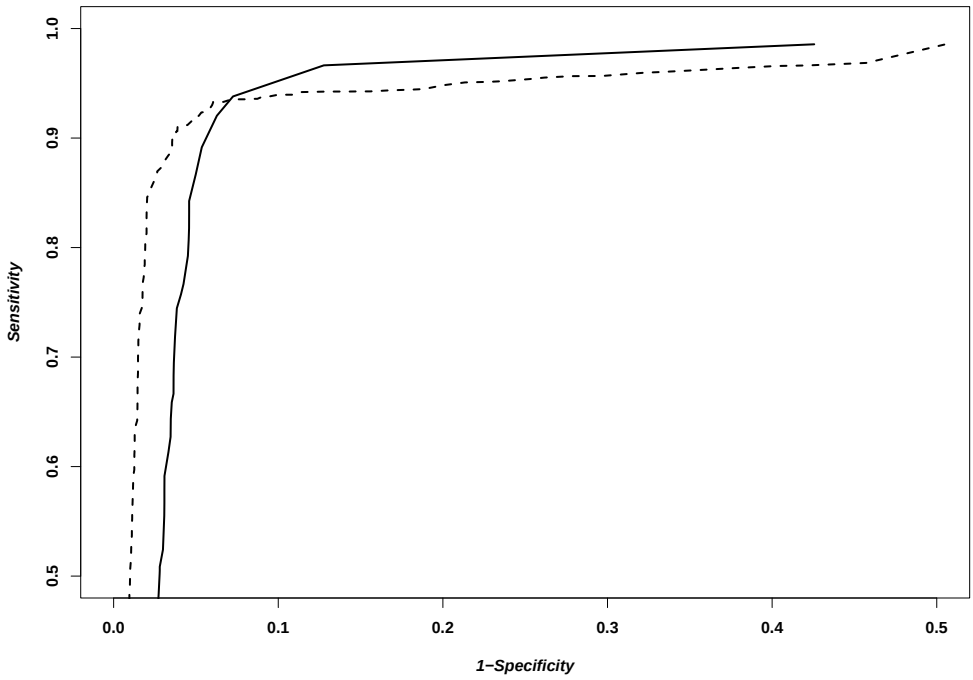

Fig. 14 ROC curve, (—) BCPA, (· · ·) CBS.

for identifying locations within aberrant segments.

In each sample, a position was called aberrant by BCPA if the posterior probability of an aberrancy exceeded the threshold $\gamma*$, and in CBS if the t-test p-value of the respective segment of which the BAC is member is at or below p^*. The results are summarized in ROC curves.

Figure 14 shows the Sensitivity/(1- Specificity) trade off for both BCP and CBS with one of the ten simulated data sets. Both methods did very well with the data sets, although BCPA showed a trade off in higher sensitivity. At (1-Specificity) = 0.1, the Sensitivity was compared between both methods. BCPA showed consistently higher

sensitivity, on average 3% higher, with a one way t-test p-value of 2.995e-05. CBS did well at detecting the bigger aberrations, with midpoints 30 and 95, although with a strict $\alpha = .01$ level failed to identify the weak aberration at location 155. At a $\alpha = .1$ level, CBS did detect breaks near 155 but had a difficulty locating the segment end points.

III.4. Wilms Tumor BAC Arrays

Wilms tumor BAC arrays, $H = 164$, from a Wilms Tumor experiment run at Roswell Park Cancer Institute were analyzed with BCPA, with attention on chromosome 1 (Natrajan *et al.*, 2006). As in the simulations, CBS was compared with BCPA. MCMC samples were generated in C++ scripts on a Dell 8200 with dual core Intel Xeon processors. Each configurations took on average under five minutes. BCPA was fit with the change points corresponding to the best BIC score as described above, for at most ten segments/sample, at $W = 100, 10$ and 1. The effective prior sample size for the variation was set to a a non-informative level of $n_0 = 10$. The prior for θ was modified to allow for a small margin about 0, $(-0.05, 0.05)$ in the normal copy state, rather than the strict mass at 0.

$$\begin{aligned} \pi(\theta_{1\tilde{k}}) &\equiv N(\theta_{1\tilde{k}}, \tau) \cdot I(\theta_{1\tilde{k}} < -\delta) \\ \pi(\theta_{2\tilde{k}}) &\equiv N(\theta_{2\tilde{k}}, \tau) \cdot I(-\delta < \theta_{2\tilde{k}} < \delta) \\ \pi(\theta_{3\tilde{k}}) &\equiv N(\theta_{3\tilde{k}}, \tau) \cdot I(\theta_{3\tilde{k}} > \delta) \end{aligned} \tag{3.10}$$

In large part, the posterior effect is to select the no copy state more often for small deviations observed in the means about 0. The parameter δ was chosen to be 0.05. The

reason for this modification was to account for centering error in the normalization. This modification will be studied with future data sets, utilizing better normalization strategies, but for now the data is taken as given.

Figures 15−17 show the posterior state, median with 50% C.I.'s, as measured across all $H = 164$ Wilm's Tumor samples, for $W = 100, 10$ and 1. The 50% C.I.'s clearly shrink with borrowing, although notice the difference in resolution concerning the peaks and valleys in the median state between W's. At $W = 10$ there may be more information conveyed about local 'hot spots' in the population in contrast to $W = 100$. This must be validated. In Figures 18−19, the posterior probabilities of gain and loss have tighter intervals as well, as borrowing is increased. There is clearly an advantage for making population level inferences concerning state changes with reduced variability. This benefit is realized at the sample level also. The posterior distribution of copy state may be used to assess not only significant changes within samples, but also a level of uncertainty is attached to those conclusions, i.e. the entire posterior distribution of the state ϕ at each BAC is available. As the distribution of ψ narrows, an overall trend is resolved. It is reassuring that this trend is similar for all three values of W.

Contrast these results with the expected states discovered by CBS. Here, CBS was fit with a liberal $\alpha = 0.01$. In many of the samples this resulted in more than ten segments. Smaller levels of α were tried, with similar results. A segment was called a copy gain/loss within a sample if the two-sample t-test had a p-value less than 0.0001, and the mean was positive/negative. The frequency over all samples was estimated by averaging across each BAC. The trend in the expected state is similar, in the first half of the data the loss is about 10%, and in the last half the gain is at about 40%.

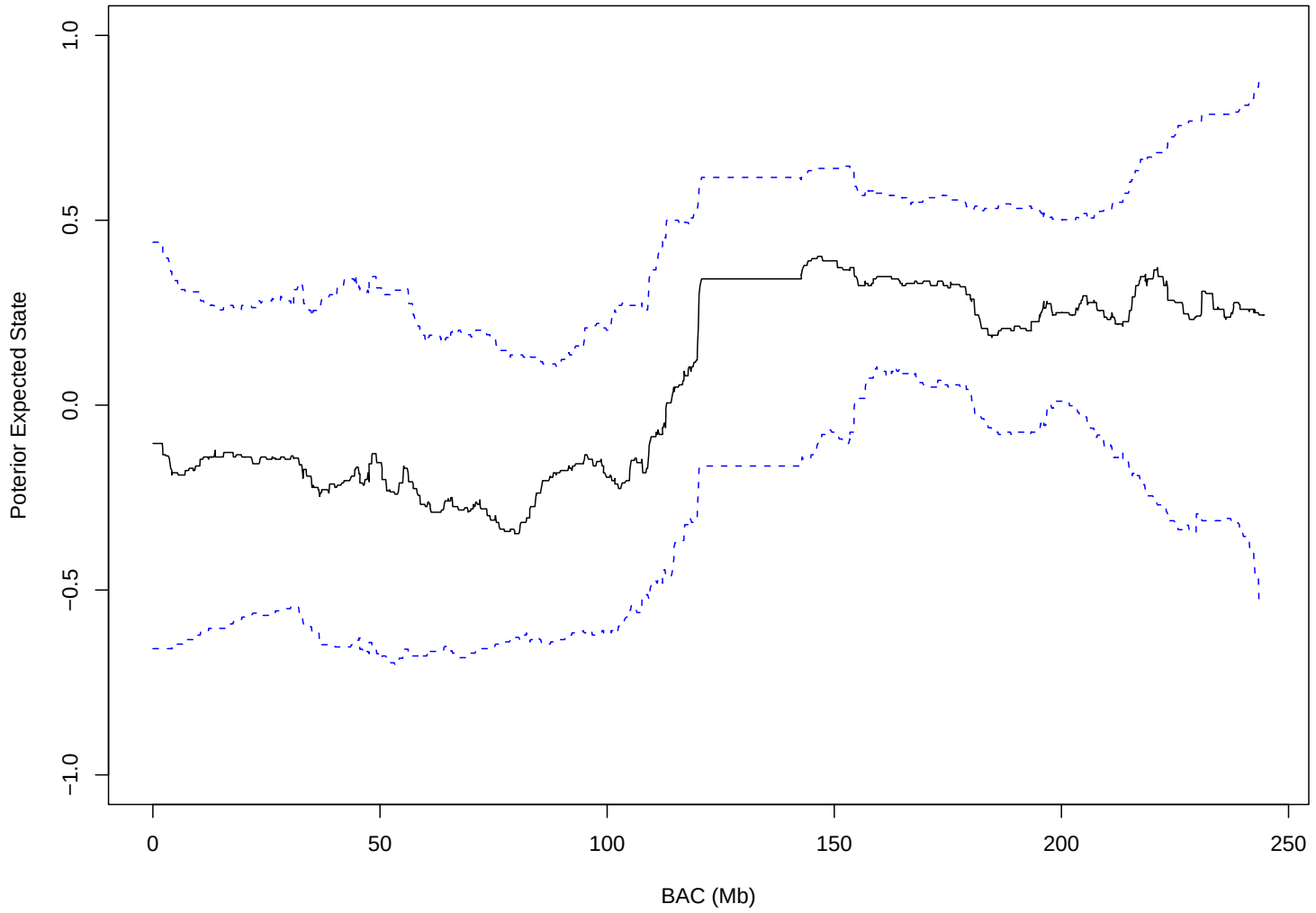

Fig. 15 BCPA $W = 100$, $-$ posterior median state, $\cdots$ 50% C.I..

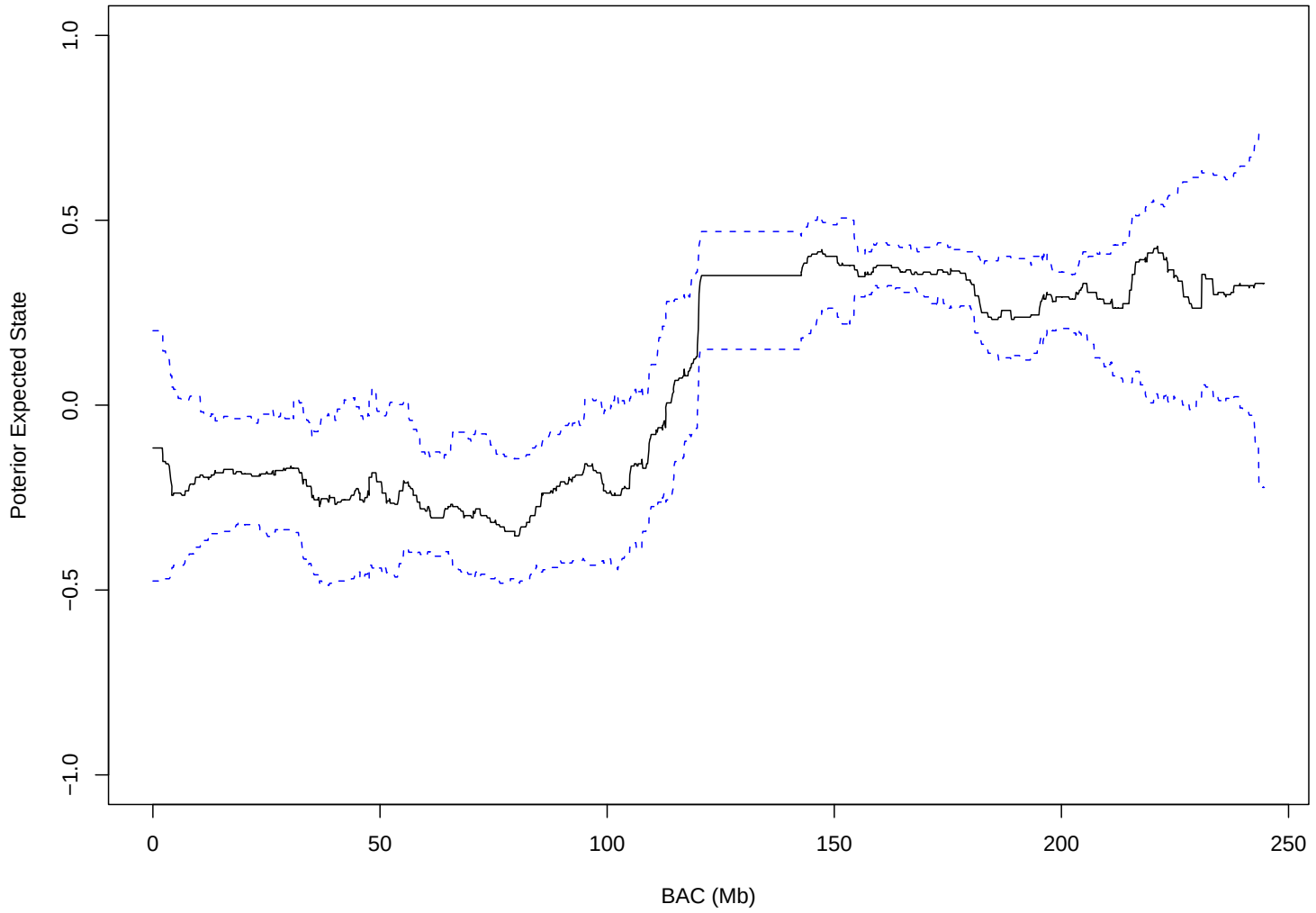

Fig. 16 BCPA $W = 10$, $-$ posterior median state, $\cdots$ 50% C.I..

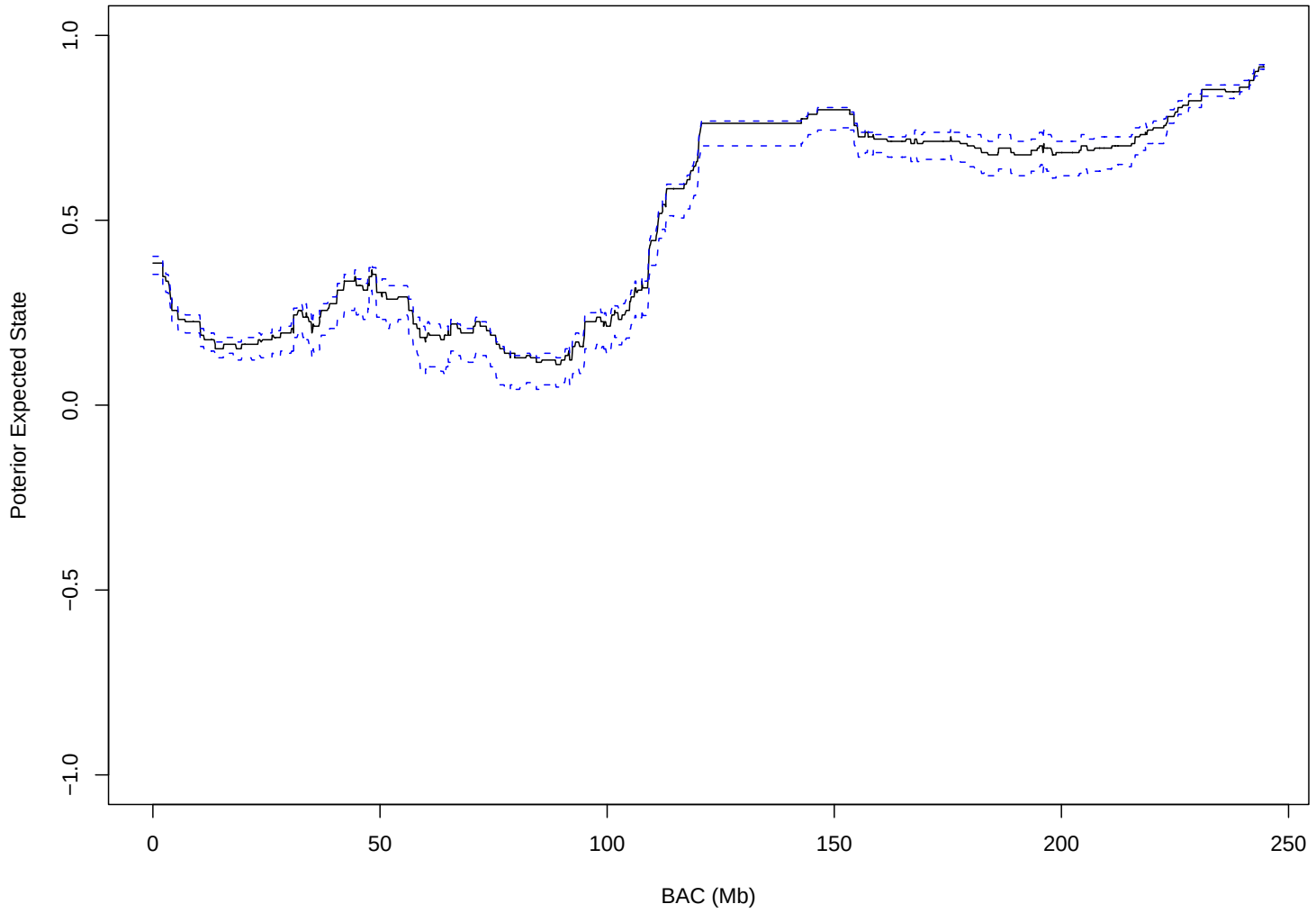

Fig. 17 BCPA $W = 1$, $-$ posterior median state, $\cdots$ 50% C.I..

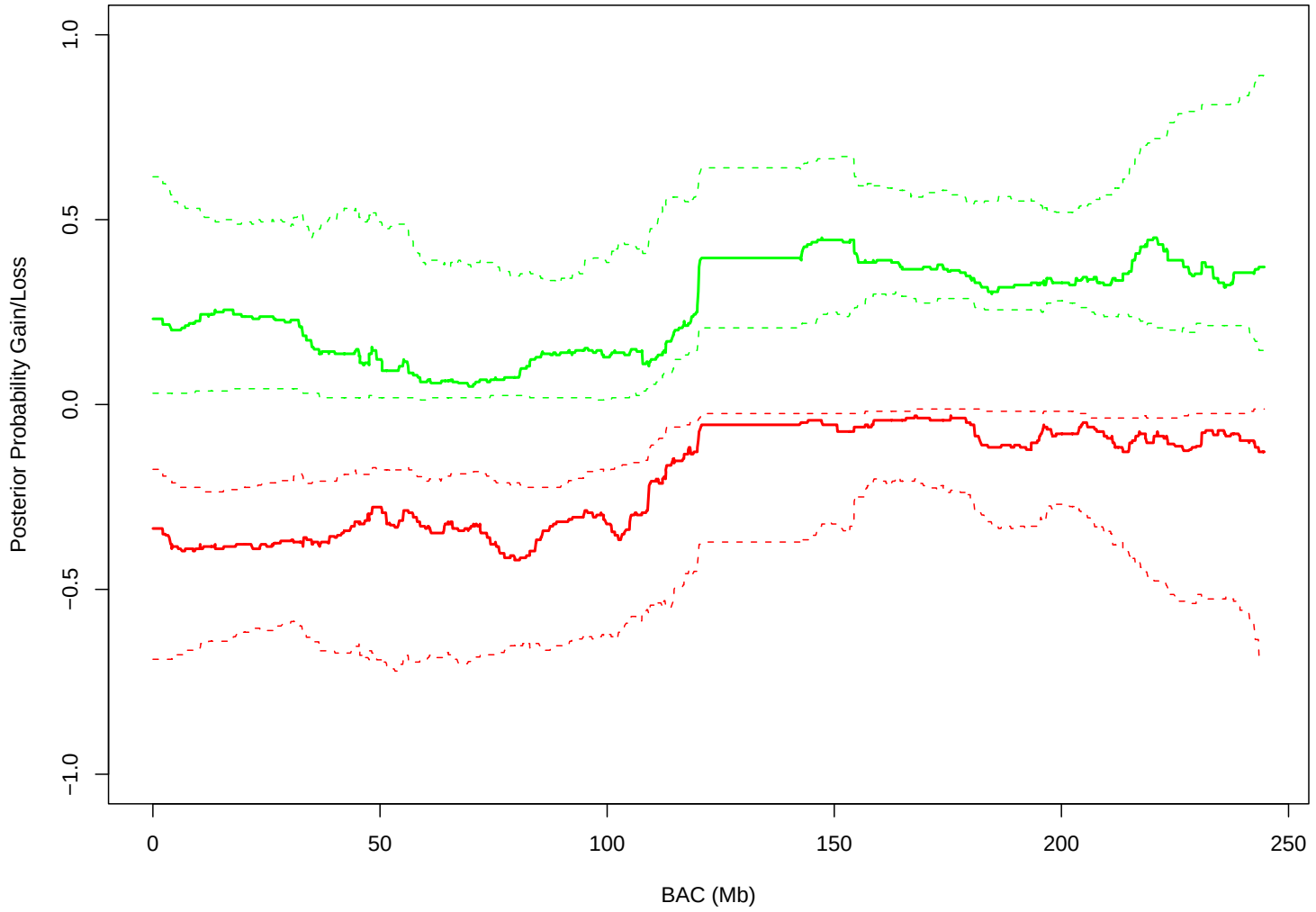

Fig. 18 BCPA $W = 100$ posterior probability gain (positive ordinate) and loss (negative ordinate), $-$ median and $\cdots$ 50% C.I..

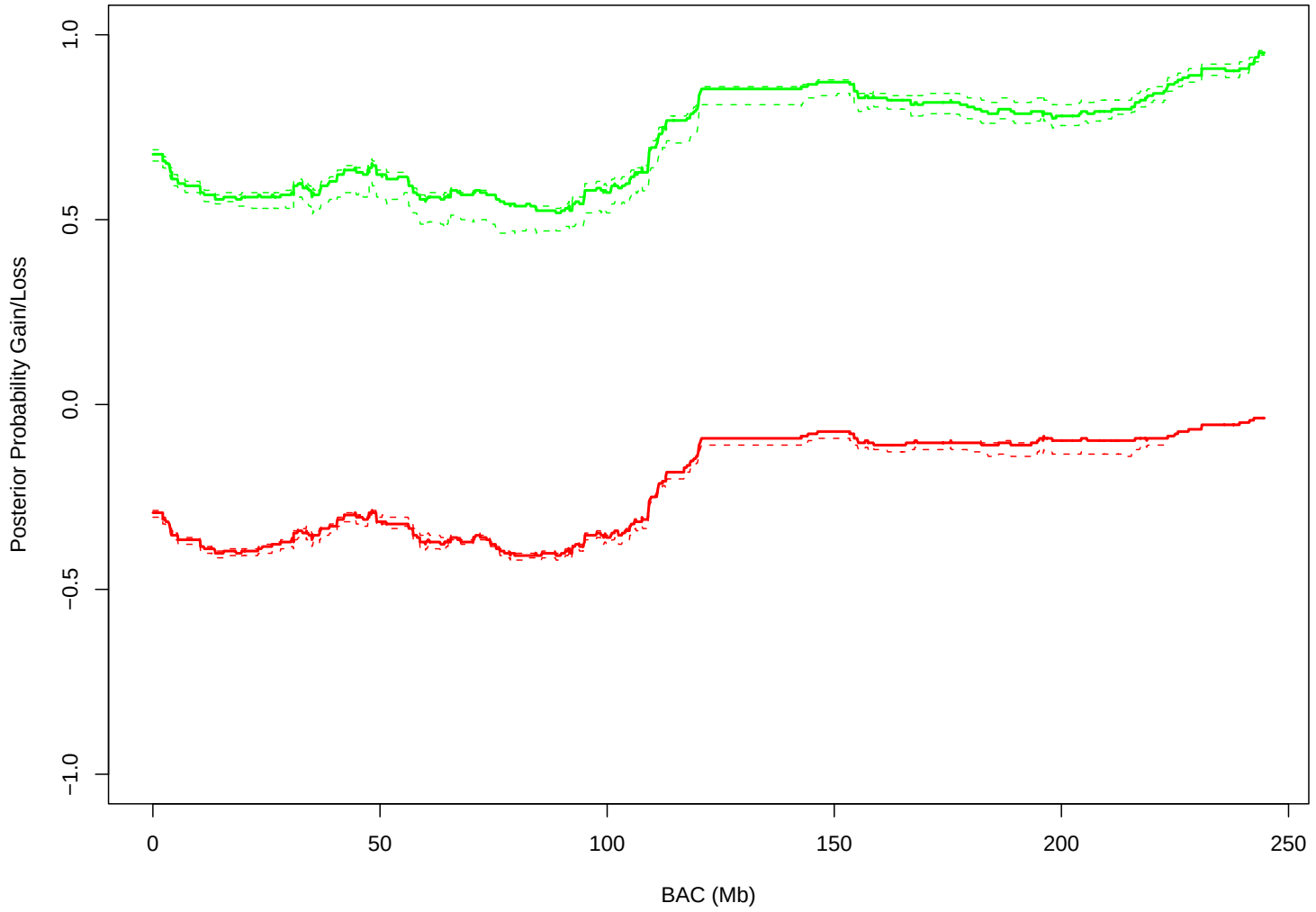

Fig. 19 BCPA $W = 1$ posterior probability gain (positive ordinate) and loss (negative ordinate), $-$ median and $\cdots$ 50% C.I.

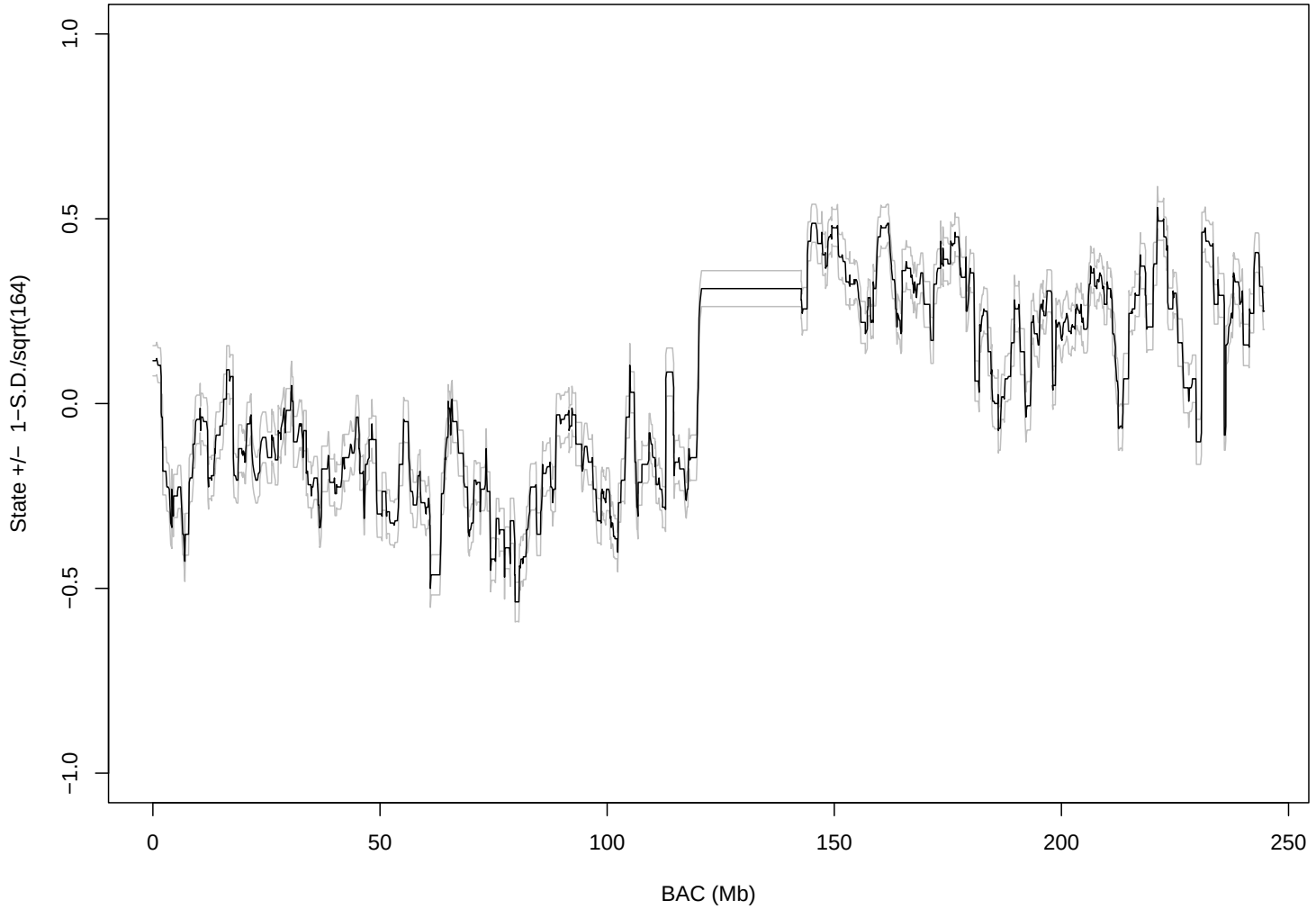

Fig. 20 CBS, mean state (black) +/- 1 sd/$\sqrt{164}$ (gray)

Although the trend in the frequency of the state computed from CBS is much noisier, Figure 20. CBS may be more sensitive to smaller changes along the chromosome. This deserves to be validated. Note that other p-value cutoffs were tried with similar results. Admittedly, the results here are based on an attempt to produce the very best of each algorithm. Optimizing each algorithm entails possible error.

III.5. Discussion

BCPA was successfully developed, and compared against the current segmentation standard, CBS, for analysis of BAC and aCGH arrays. We demonstrate that a Bayesian approach for learning about chromosomal copy gain and loss is feasible up to a change point strategy. Uncertainty due to the change points is a source of variability that has received little attention in the literature. Our simulations suggest that software such as DNACopy offer quite reasonable starting values, as long as the starting values are chosen liberally. Missing real change points can increase variability and reduce power. Our strategy is to account for the uncertainty in the change points by marginalization. This offers a feasible approach to increase the power for detecting chromosomal copy gain and loss in heterogeneous patient populations.

Our simulations suggest a marginal though significant trade off in sensitivity with BCPA. The simulated data was not difficult to classify, as both methods did very well, and BCPA is computationally intensive. The larger simulations, run in C++ on a Dell 8200 with dual core Intel Xeon processors took under 48 hours. While the marginal gain does not appear worth the cost in time, there are many examples of Bayesian methods that perform competitively with the best Frequentist method, in the absence of prior information.

An important next step is to consider the utility of prior information. In many cases, the ploidy of a cell type is partially known or may be obtained. Karyotyping may serve to inform the high-resolution experimental analyses. Exploiting this prior knowledge may help build more intelligent models for detecting copy gain or loss on aCGH allowing for replicates and normal controls. Future work includes summarizing the benefits to using prior information for aCGH analysis.

Bayesian modeling does offer the potential to improve probability estimates of the rate of chromosomal instability in heterogeneous disease populations. While our modeling approach is general enough to handle many scenarios, it comes with trade offs. These models require skill in prior specification and also computational efficiency. Competitive models can adapt. BCPA shows promise as a useful and novel approach for BAC and aCGH analysis.

CHAPTER IV

BAYESIAN DYNAMIC NETWORK INFERENCE WITH REAL-TIME GENE EXPRESSION

IV.1. Introduction

The dynamics of gene regulation are fundamental to Genomics and Bioinformatics. Understanding the pathology of cells requires a thorough understanding of the order of molecular reactions that alter cellular states. Microarrays, the evolutionary tool to measure changes in gene expression, offer insight into the variation in gene expression. The problem with microarrays is that the experimental designs are limited, lacking the temporal resolution necessary to make detailed inferences about interactions between genes. Moreover, inferences are based on samples extracted from ex-vivo cells. The latest experiments, using green fluorescent proteins (GFP)s, offer finer temporal resolution of gene expression in live cells. These experiments are more suitable for learning about pathways, providing real-time data in live cells.

Regulation of gene expression is a dynamic and complex process, evolving under many pressures, one of which is the competitive pressure for survival. The state of a cell is the result of a collective sequence of programmed molecular reactions. A chronic disease state, for example, is the result of a chain of events that lead to an un-sustainable course for the organism. How are genes related temporally? How do collections of genes react to stimuli over time? These are some of the questions behind much of the interest in gene pathways.

Medical research is fueled with great interest in molecular pathogenesis. Prognosis often worsens as disease progresses, and therefore early detection is crucial for successful therapy. Our understanding of the processes by which healthy cells are transformed to unhealthy states, may better inform our ability to formulate therapies for clinical research. Such investigation naturally leads to further hypotheses, such as how genes respond to treatments. The relevance is that pathogenesis, treatments and gene interactions, all share a temporal component that is poorly understood. Knowledge of underlying key mechanisms that sequentially alter the dynamics in gene expression, it is widely believed, would improve chances for early detection and better inform treatment options.

Pathogenesis in living organisms is not always characterized as an instantaneous process, although within a single cell, the final events that lead to the irreversible stages of a chronic disease can be sudden. One way to think about progression, is modulation in the structural dynamics relating the genes. For example, consider the special case of a Linear Gaussian Switching State Space model

$$\begin{aligned} X_t &= A_t X_{t-1} + W_t \\ Y_t &= B X_t + E_t \end{aligned} \tag{4.1}$$

where $W_t \sim N(0, \Omega)$, $E_t \sim N(0, \Sigma)$ and

$$A_t = \begin{cases} A_o & t \leq t_0 \\ A_1 & t > t_0 \end{cases}.$$

In the above model, the observed variable Y_t, representing gene expression, depends on the (continuous) state of the network, X_t, through the matrix B. The state at

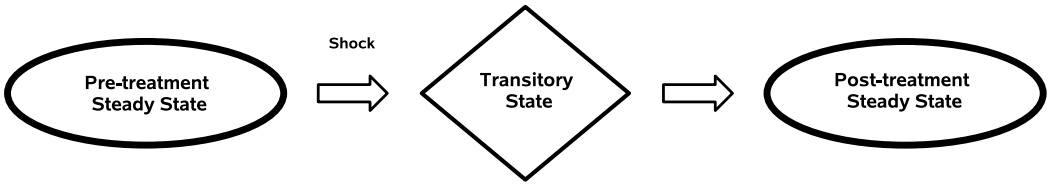

Fig. 21 State progression model - following a shock, genes progress from the pre-treatment steady state to a transitory state until the post-treatment steady state is reached.

time t, X_t, depends on the state at time $t-1$ through the transition matrix A_t, that depends on whether t_0 is reached. Time t_0 may be thought of as the time of modulation, or in medical terms, if disease onset is the focus, pathogenesis.

IV.1.1. The State Progression Model

Figure 21 depicts progression of a cell, or population of cells, from a pre-shock steady state through a transitory state to a post-shock steady state (Grodins, 1963). In the transitory state, there are several possibilities for the morphology of the system. One is that the nature in the way the genes are related, is modified. This is possible due to, for example, a physical mutation of the organism's DNA. Another possibility is that the physical architecture of the system remains intact, although the expression levels of some genes are intensified and others suppressed, in reaction to the shock. A third possibility is that the both the physical architecture of the system and gene expression are modified. In the last post-treatment state, the organisms has surpassed the transition and has reached a post-shock steady state which may be different from the pre-shock steady state.

Some relevant questions are how to characterize differences between the pre- and

post-shock steady states. Another question relevant for transition is what control steps were achieved to guide the system towards and through the transitory state? Specifically in health related research, modulation of the structural dynamics relating genes is of great interest, as it may lead to discoveries of new therapies that can turn chronic diseases into acute disorders.

A class of theoretical models is proposed to perform inference with GFP experimental data. The theory is motivated by historical work from mathematical biology. Experimental design issues and modeling assumptions are thoroughly discussed. Posterior inference is conducted with a case study of the publicly available S.O.S. gene network data (Friedman *et al.*, 2005). Several modeling extensions are discussed and model comparison is performed in a Bayesian framework.

IV.1.2. Real-Time Gene Expression Experiments

New experiments are making it possible to control and monitor gene expression in live cells in real time. Reporter cells are cloned with a sequence that includes the upstream promoter motif of a gene of interest. However, the down stream protein sequence is replaced by the code for green fluorescent protein. When the gene is promoted, e.g. in response to an experimental factor or treatment, the gene encoding GFP is also promoted. GFP is monitored real-time, while in production. Fluorescence acts as a surrogate for expression (Ronen *et al.*, (2002); Zaslaver *et al.*, 2004).

Although GFP has been used for some time, clone libraries are just beginning to undergo rapid expansion suitable for high-throughput experimentation (Zaslaver *et al.* 2006). The latest approaches make it possible to monitor more proteins in ever big-

ger experiments. These experiments are not on the scale of microarray experiments, perhaps on the order of at most 100 genes.

Living cell arrays (LCAs), a relatively new technology (King *et al.*, 2006), make it possible to study how genes behave and interact over time at much higher resolution than expression array experiments, with the benefit of a reduction in cost over alternative experiments. The array is fabricated into 100 cell chambers connected by microfluidic channels in a grid. Each chamber on the array contains a sample of reporter clonal cells, designed to monitor the expression of a unique gene by transfection with a reporter DNA plasmid encoding a protein that can easily be measured, i.e. green fluorescent protein. Micofluid channels on the array control extracellular stimuli delivered into each chamber. Once in a chamber, the stimuli can bind to cell surface receptors, signaling transcription factor proteins to travel into the cell nucleus and bind to promoter regions of the genes with the consensus promoter motif. Transcription of genes in the respective signaling pathway are activated. The reporter DNA plasmid encoding GFP is transcribed. Microscopic images are recorded at fixed time points over the array. Expression is monitored at fine temporal resolution in living cell arrays.

During an experiment, a homogeneous sample of approximately 100-200 cells is seeded into each chamber, depending on the cells, from a cell line transfected with a sequence which includes the upstream promoter sequence of a gene of interest, and a coding sequence for green fluorescent protein (GFP). Extracellular stimuli can be seeded into the chambers through microfluidic channels. Promotion of each gene of interest is reported by the respective GFP, which emits a fluorescence when irradiated. The data is collected by photographing the cell chambers at time intervals, and quantifying the

relative fluorescent signal in the chambers. These quantifications offer real-time gene expression profiles in live cells at the cell and chamber level.

IV.1.3. Analysis of Dynamic Gene Expression

Much interest is centered around gene network modeling, as high-throughput experiments and expansion of pathway databases are making it ever more possible. Models that can explain variation in the structural dynamics of gene expression could potentially help us understand how gene products work together to maintain order in the cells and how deviations from that order lead to pathological states. Real time gene expression experiments provide a glimpse into the complicated fabric of dynamic molecular systems. Microarray experiments are limited in that temporal resolution, as even in fine temporal sampling at best the samples contain mixtures of ex-vivo cells. Unlike expression data, GFP reporter experiments provide much finer resolution, to evaluate the non-linear relationships between the genes.

Graph theory is central to many of the concepts for modeling networks. Lauritzen (2002) reviews directed acyclic graphs (DAGs). A key feature of gene network modeling is the translation of graphs to models and the inverse problem of deriving a graph from a a model. Much of the gene network methodology developed recently centers on estimating simultaneous systems of equations. Consider the case where the measured gene expression values are assumed to be linearly related. A matrix of gene expression values Y, arranged by sample in the columns and genes in the rows, follows the form $Y = BY + E$ for a known connectivity matrix B and $E \sim MNV(0, \Sigma)$. It follows that $Y \sim MNV(0, U^T \Sigma U)$ for $U = (I - B)^{-1}$. Methods for estimating B from data, and theoretical results discussing consistency, were offered by Liao *et al.*

(2003). A very useful model for this construct is the Conditional Gaussian Autoregressive (CAR) model. In the context of static time microarray experiments $\mathbf{Y}$ is a matrix of observed gene expression values, with columns Y_i distributed as

$$Y_i|\mathbf{Y}_{-i}, \theta, w, d \sim N\left(\theta_i + w_i^{-1}\sum_{j\neq i} a_{ij}(Y_j - \theta_j), d_i^2\right) \tag{4.2}$$

A symmetry condition is imposed, requiring that $a_{ij}d_i^2 = a_{ji}d_j^2$. When the symmetry condition is satisfied and conditional variances are equal to $d_i^2 = \sigma^2/w_i$, by factorization we have that

$$\mathbf{Y} \sim N\left(\theta, \sigma^2(W - A)^{-1}\right) \tag{4.3}$$

where $A = [a_{ij}I(i \neq j)]$ and $W = \text{diag}(w_1, ..., w_n)$.

These models may be made dynamic by allowing for a temporal aspect, allowing one or more of the model parameters to vary with time. In a Markov scheme,

$$P(\theta_t|\theta_{t-1}, \theta_{t-2}, ..., \theta_1) = P(\theta_t|\theta_{t-1}) \tag{4.4}$$

the conditional probability of observing θ_t at time t, only depends on θ_{t-1}. The Markov assumption is very flexible and popular for its simplicity, although not always appropriate for network modeling. Semi-Markov models, allowing for a duration effect, may be more realistic. For example, the conditional distribution of θ_t depends not only on θ_{t-1}, but also on the length of time θ_{t-1} has been in state A. Zoa (2005)

extended the simultaneous linear system to a more general non-linear system, for learning gene pathways from array data. Tamada (2002) demonstrated the utility of prior information in fitting nonlinear systems of equations to yeast array data, in a heuristic scheme. Dobra *et al.* (2006) developed methodology for learning sparse graphical networks from expression data. Perrin (2003) describes linear State Space Modeling of time series gene expression data.

Much of the past statistical work for modeling gene networks was designed for modeling gene expression from microarray data with relatively poor temporal resolution. Dojer *et al.* (2006) derived a method of differential equation modeling, with posterior scoring of gene networks flexible enough to include information on known transcription factors and protein degradation. Zhoa *et al.* (2006) offers a flexible modeling scheme for modeling gene networks with minimum description length scoring. In the Bayesian framework, Geweke and Tanizaki (2001) discuss MCMC methods for fitting a general class of non-linear state space models and Roberts *et al.* (2000) for fitting Hidden Markov Models with Reversible Jump MCMC. Ghahramani (1997) has a full review of Bayesian Dynamic networks. This work was certainly groundbreaking, and offered important advances, although for modeling gene expression levels as measured with the kind of temporal resolution capable with GFP's, more flexible models, and better assumptions are required.

IV.1.4. Real-time Gene Expression Analysis

Gene network modeling requires greater biological background, than say in bio-marker discovery with microarrays. Since LCA's are relatively new experiments, there is limited statistical understanding for modeling the fine temporal resolution and subtle

gene interactions. One of the fundamental differences of microarray results is that with fine temporal resolution, more background information is needed to inform about the plausible biological sequence of events, to draw valid modeling assumptions. In many systems, even the best understood, there is uncertainty attached to the sequences of genomic events, i.e. gene A regulates gene B, followed by gene C, and so on. Modeling assumptions also rest on the negative feedback between genes and their own protein products, i.e. negative feedback loops over time. Understanding of positive and negative feed-back loops is incomplete, although these are thought to be essential in order for cells to achieve steady states. Plausible modeling assumptions gene-gene interrelationships and loops are not established yet, due to uncertainties about the biology and even limitations of the experiments.

Chen *et al.* (1999) reviewed a class of models treating a gene network as a system of differential equations. Chen *et al.* discussed modeling RNA and protein concurrently. In the present context, both are not observed, however, some of the basic feature of the Langevin Equations (Coffey *et al.*, 1996), i.e. first order differential equations discussed below, will be useful. It is helpful to familiarize with biological concepts in modeling at the subcellular level. Thorough reviews of mathematical models developed for molecular and cellular biology are presented in Segel (1980), and Wu (2001), and Schneider *et al.* (1975). A basic understanding of control theory (Grodins, 1963) is also useful as many of the concepts in modeling biological systems at the subcellular level have been adopted from control theory. Detailed development of statistical approaches for modeling gene expression dynamics were discussed by Ronen *et al.* (2002). These methods suffer from a number of shortcomings, at one end a lack of interpretation, and at the other end, under-fitting the data.

A class models is needed to link the dynamic dependencies between the genes: (1) that makes biological sense, (2) that makes modeling sense and (3) can provide results interpretable for inference about the genes. Each component of the three fold goal is essential, as we must be able to interpret the results. Otherwise, we cannot make inferences. Therefore, if we cannot make sense of the data, then we cannot perform network inference. Nonetheless, some of the methods that have been used to model gene expression networks with microarray data may prove helpful, if not of direct use. The information provided by LCA's combined with historical pathways offers hope for learning about real-time gene regulation.

Modeling dynamic gene expression presents many statistical challenges. Some challenges that are quite relevant here are related to: (1) experimental design, (2) normalization, (3) modeling assumptions and inference. Gene expression dynamics have been characterized in other fields, and as such may serve as prior information for modeling LCA's. Defining the relationships among genes will be an ongoing problem and will require help from biologists, in order to formulate modeling assumptions. We are interested in inference, and as such there may be no best way in general to summarize LCA results, although strategies may be implemented and studied to do well. Inevitably, the design of the experiments will determine the fate of the research. As such, the unknown sources of major technical variation will need become the focus of attention, in order to improve the designs. The basic science of these experiments and designs need to develop before we can better handle the tougher questions. For now many of the relevant questions for statistical research are related to the sources of variation in the experiments.

IV.1.5. Historical Pathways

Flexible modeling strategies are required for LCA's, in order to account for uncertainty in the prior information as well as the data. The prior information is available in the peer reviewed articles, but also in the public domain in pathway databases (www.pathguide.org). The pathways are commonly illustrated on line by static cartoon like figures depicting the temporal relationships between the genes (nodes) by directed arrows (interactions). It is unlikely that a model will ever be specified perfectly, despite the wealth of information available about many genomes, due to the plenitude of molecular interactions that are possible in living organisms, many of which are at best partially understood and still, and quite complex. In practice, it is likely that at least some of the pathway information will be incomplete. How to model incomplete pathways with incomplete information are relevant statistical questions.

IV.2. Experimental Design

As a consequence of the experimental design, reporter genes are measured separately. It is unrealistic to expect to obtain reporter expressions from may different genes in the same cell. It is possible to conduct an experiment with multiple reporters using different dyes, although dye effects and biases may distort conclusions. The consequences of this design feature, for learning about expression profiles, are that certain changes, dominant trends, might be resolved by known or learned gene-gene dependencies, although more specific features may not be explained.

An additional consequence of the design is that genes observed in the same replicate experiment may have very little within experiment correlation. Consider two dependent genes, gene A and gene B. Measurements on A may be no more temporally

related to gene B if measured in the same experimental replicate, than in another experimental replicate. This suggests that averaging of the gene expression profiles across experiments may improve inferences, if little is lost by modeling across experiment. For the time being, we are mainly interested in the dominant features in the profiles that characterize broad changes associated with biological factors of interest and so averaging replicates appears sensible.

Another important aspect of the experiment is the temporal design. Consider an experiment that introduces an extra-cellular stimuli at time t_0, at the outset of the experiment. Time elapses until time t_1 and GFP reporters are measured. If the scientific question involves comparing the pre-stimulus steady state to the transitory state and further to the post-stimulus steady state, then obviously the design is inadequate. A more suitable design would measures GFP for a time lapse from $t_{-\delta}$ to t_0 before the stimulus is introduced. Comparing the variability in gene expression between phases, pre-stimulus, transition and post-stimulus is a worthwhile goal as each may offer some new insight into the mechanisms important to morphology in the cell. Good principles of experimental design are essential in order to overcome confounding.

IV.3. Normalization

GFP is monitored in each respective cell, in each respective chamber or well. In order to measure trends across genes, fluorescent levels must be averaged across cells in a well. A sensible approach outlined by Friedman *et al.* (2005) is to approximate promoter activity by weighting fluorescent signal by cell volume. Standard imaging software may be used for this. In practice, since the levels can vary between experiments, there is a need to standardize. A straightforward approach adopted for the

work presented here is to normalize all of the profiles in each replicate experiment to have min = 0 and max = 1. This crude measure puts all of the experiments on equal footing. This delicate aspect needs to be investigated, as changes in the normalization may alter inferences and conclusions.

IV.4. The Transition State Model

A class of dynamic network models is proposed, to flexibly allow for inference at many levels in a temporal gene expression experiment. A clear advantage of modeling in a Bayesian framework is the flexibility to perform posterior inference accounting for all of the uncertainty. Frequentist nonlinear modeling comes with its own special nuances, one of them being that the inferences are based on normal approximations. In high dimensions, this can lead to faulty conclusions. This is a more critical issue here as there are sharp nonlinear changes in the data that need to be modeled. Although not all features can be modeled, given the incomplete information, the dominant interactions, that stand out abruptly, are of specific interest. At this early stage of modeling, it is useful to work with well known pathways for model validation. In this complex framework, potential confounding and modeling uncertainty will limit our understanding.

Let $y_i(t)$ be the level of GFP measured by a reporter of gene expression for gene i at time t. It is assumed that $y_i(t) \sim N(\mu_i(t), \sigma_i^2)$ is iid Gaussian with mean $\mu_i(t)$ depending on time. In the presence a promoter gene j, with mean expression level $\mu_j(t)$ at time t, the partial derivative of $\mu_i(t)$ with respect to time is assumed to be of the form

$$\frac{\partial \mu_i(t)}{\partial t} = V_i f_i(\mu_j(t)) - U_i \mu_i(t) \tag{4.5}$$

where the coefficient $U_i > 0$ accounts for the degradation in expression as the abundance, $\mu_i(t)$, rises. Negative auto-feedback is well documented in biology. There are many examples of a gene product forming a dimer that negatively auto-regulates its own production. $V_i > 0$ governs the relationships between change in the expected gene expression in gene i and $f_i(\mu_j(t))$. The set of first order differential equations governed by (4.5) is known as the Langevin Equations, studied by Langevin in conjunction with particle physics (Coffey *et al.*, 2004). The fundamental Langevin model is adopted here (Chen *et al.*, 1999) for modeling dynamic gene expression.

In molecular biology the choices for $f_i(\cdot)$ suggested are $f_i(\mu_j(t)) = \mu_j(t)/(K_i + \mu_j(t))$ and $f_i(\mu_j(t)) = \mu_j(t)^a/(K_i + \mu_j(t)^a)$, $a > 0$ (Ronen *et al.* 2002; Chen *et al.* 1999; Wu 1999). The choice adopted here is linear $f_i(\mu_j(t)) = \mu_j(t)$. The solution to (4.5) with a linear link is

$$\mu_i(t) = ce^{-U_i} + V_i \int_0^t \mu_j(s) ds \tag{4.6}$$

where the constant c is determined by the initial conditions at $t = 0$: $\mu_i(0)$ and $\mu_j(0)$.

The development of the proposed model is motivated with the publicly available case study example taken from the S.O.S. gene network experiment (Friedman *et al.*, 2001). We will discuss the experiment in more detail in the case study section, but for now consider Figure 22, of the temporal profile of the recA gene in the S.O.S. gene network experiment. At time t_0 a treatment is introduced. Notice in the figure

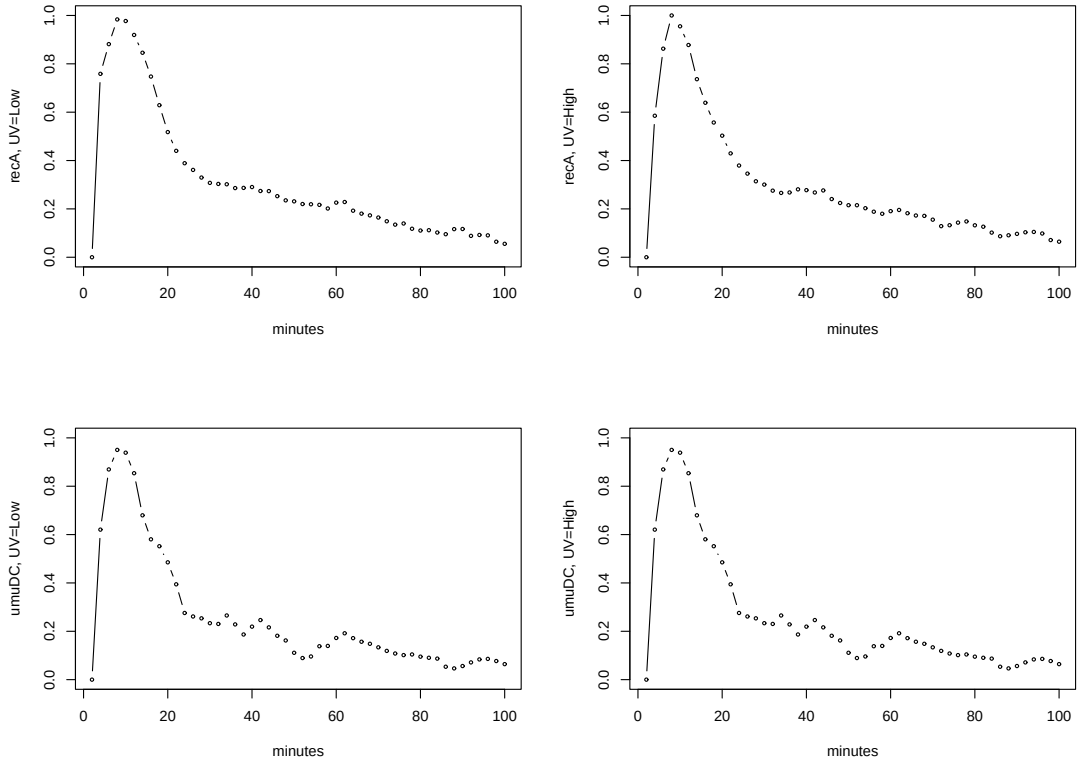

Fig. 22 Temporal profiles in S.O.S. genes

that the expression initially accelerates upwards and then declines. This profile is reminiscent of the temporal profiles commonly seen in GFP experiments (King *et al.*, 2006).

The class of models for capturing trends in expression during the transition state, in response to a stimulus, as

$$\mu_i(t) = (a_i + b_i(t))\lambda_i^t \tag{4.7}$$

for $0 < \lambda_i < 1$ and $-\infty < a_i, b_i(t) < \infty$. The partial derivative of $\mu_i(t)$ with respect

to time is

$$\frac{\partial \mu_i(t)}{\partial t} = \frac{\partial b_i(t)}{\partial t}\lambda_i^t + (a_i + b_i(t))\log(\lambda_i)\lambda_i^t. \tag{4.8}$$

Notice that by combining (4.8) and (4.9) we have

$$\frac{\partial \mu_i(t)}{\partial t} = \frac{\partial b_i(t)}{\partial t}\lambda_i^t + \mu_i(t)\log(\lambda_i). \tag{4.9}$$

where the second term preserve the negative relationships between changes in gene expression and expression level. The coefficient $\log(\lambda_i) < 0$ acts as a degradation constant. This model has the form of the Langevin equation (4.5). The function $b_i(t)$ of time represents the trend in expression due to the integration of promoter/suppressor activity. Different functional forms for $b_i(t)$ are considered. In the case that gene i has one promoter element, expressed at a constant level, it is assumed that the form of $b_i(t)$ is

$$\frac{\partial b_i(t)}{\partial t} = b_i, \tag{4.10}$$

This model has the form

$$\mu_i(t) = (a_i + b_i t)\lambda_i^t \tag{4.11}$$

with partial derivative in time

$$\frac{\partial \mu_i(t)}{\partial t} = b_i\lambda_i^t + \mu_i(t)\log(\lambda_i). \tag{4.12}$$

The effect of increasing b_i is to increase the initial slope by which gene expression is promoted. Changes in expression are related to b_i damped over time by the term λ_i^t. Assuming that the promoter element is maintained at a constant level, this is a reasonable assumption and will serve a practical purpose in many experiments, in which the actual level of the promoter element is not observed. The term $\frac{\partial b(t)}{\partial t}$ may supply surrogate information for an unobserved regulator, as incomplete information about pathways is likely to be common in many such experiments. In the case that the actual level of the promoter element is not observed, the linear term is an estimate of the average increase in gene i due to regulation. The assumption modeled is that the average level of change serves as a reasonable estimator for the true unobserved change. Model (4.11) is plotted for different values of a, b, and λ in Figure 23.

In the above expression, $\frac{\partial \mu_i(t)}{\partial t} \to 0$ as $t \to \infty$. In the S.O.S. data, Figure 22, there is a sharp decline in promotion up to a knee, followed by gradual linear evacuation. During the subsequent gradual evacuation period, periodicity is observed. The initial shock that increased gene expression, it is hypothesized, forced all of the cells' cycles to be synchronized. After the dominant effects of the treatment were realized, the cells resumed their respective cycles. Including a term in (4.9), in the form of an intercept, $\alpha_i > 0$, can account for the subsequent gradual linear decay, where in this case

$$\frac{\partial \mu_i(t)}{\partial t} = b_i \lambda_i^t + \mu_i(t) \log(\lambda_i) - \alpha_i \tag{4.13}$$

the slope of expression during evacuation is $-\alpha_i$.

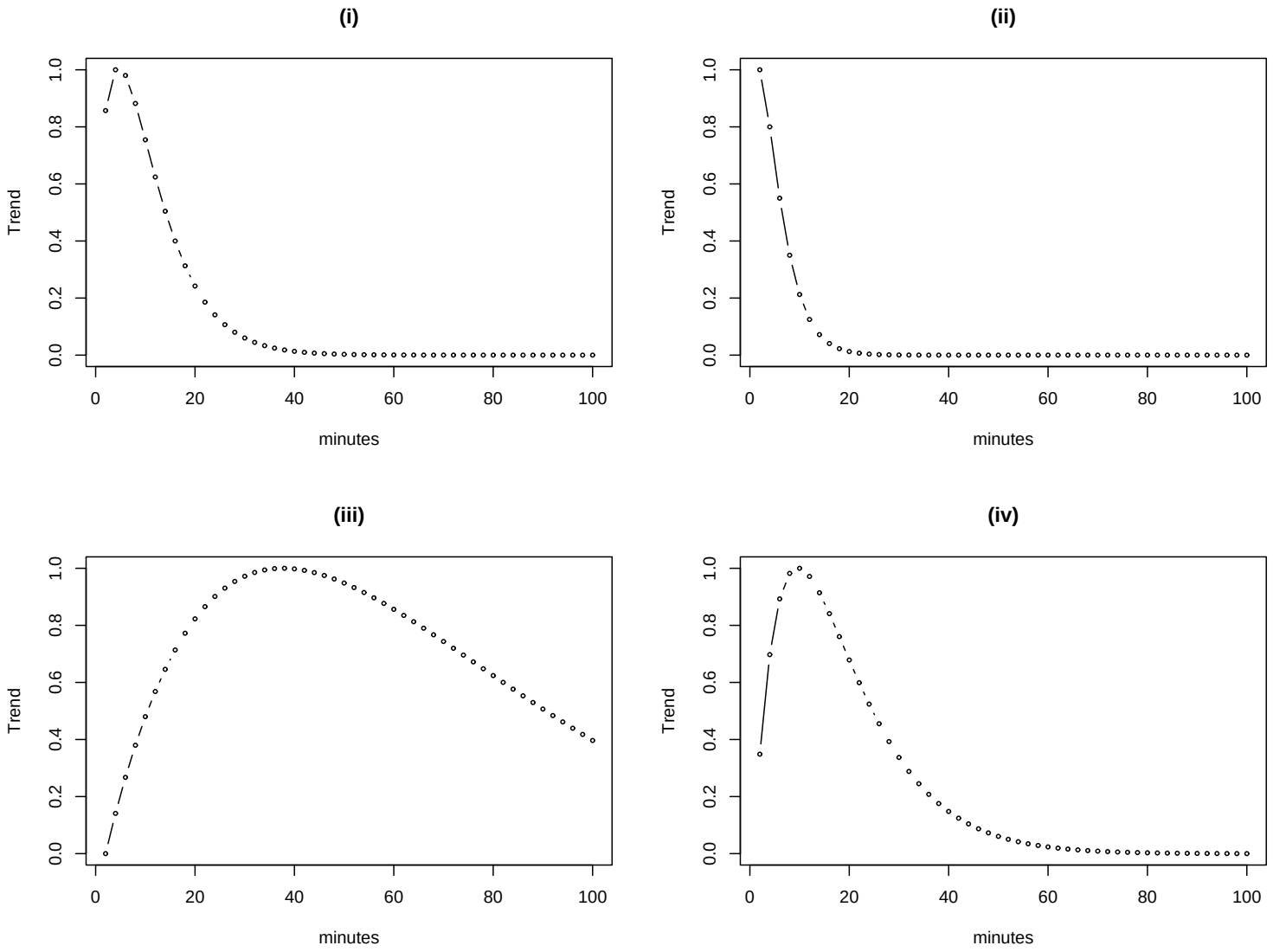

Fig. 23 Model (4.12): (I) a = 1, b =2 , $\lambda = 0.70$, (II) a =1 , b = 1.5, $\lambda = 0.50$, (III) a = 1, b = 1.5, $\lambda = 0.95$, (IV) a = -1, b = 3, $\lambda = 0.80$.

IV.4.1. Promoter and Suppressor Activity

Suppose that a promoter gene k and repressor gene j expression are observed in conjunction with gene i. The term $\frac{\partial b_i(t)}{\partial t}$ is assumed to be linear in

$$\frac{\partial b_i(t)}{\partial t} = \beta_{ik}\mu_k(t) - \beta_{ij}\mu_j(t) \tag{4.14}$$

$\mu_k(t)$ and $\mu_k(t)$. The term $b_i(t)$ is

$$b_i(t) = \beta_{ik}\int_0^t \mu_k(s)ds - \beta_{ij}\int_0^t \mu_j(s)ds \tag{4.15}$$

Based on the additional information provided by promoter gene k and repressor gene j, the functional form of $\frac{\partial \mu_i(t)}{\partial t}$ is assumed to be

$$\frac{\partial \mu_i(t)}{\partial t} = \left(\beta_{ik}\mu_k(t) - \beta_{ij}\mu_j(t)\right)\lambda_i^t + \mu_i(t)\log(\lambda_i) - \alpha_i \tag{4.16}$$

This form implies that the change in expression in gene i due to degradation is linear in $\mu_i(t)$, while the change due to promotion or repression is non-linear due to the decay in effect through λ^t. The above assumptions require the sign of the slope to be inversely related to the level of gene expression.

A practical concern is that over time gene expression exhibits temporal periodicity, as seen with actual data, Figure 22. Periodic functions are a common phenomena in may processes where a forcing function can explain oscillations (Grodins, 1963). In this case the cells may be trying to achieve homeostasis, or some other state in response to a shock. The mean is adapted to include the function $\theta_i(t)$, assumed to be a mixture of sin waves of different amplitude, scale and an unknown number of

components S.

$$\theta_i(t) = \sum_{s=1}^{S} h_s \sin(f_{is}(t)) \tag{4.17}$$

The partial derivative of $\mu_i(t)$ with respect to time is

$$\frac{\partial \mu_i(t)}{\partial t} = \frac{\partial b_{1i}(t)}{\partial t}\lambda_i^t + \mu_i(t)\log(\lambda_i) - \alpha_i + \frac{\partial \theta(t)}{\partial t}. \tag{4.18}$$

Under the assumption that $\theta(t)$ is assumed to be a mixture of sine waves, the derivative is a mixture of cosine waves

$$\frac{\partial \theta(t)}{\partial t} = \sum_{s=1}^{S} h_s \frac{\partial f_s(t)}{\partial t} \cdot \cos(f_s(t)). \tag{4.19}$$

A reasonable deduction is that

$$\frac{\partial \theta_i(t)}{\partial t} \approx \beta_{\theta_i}\theta_i(t - \Delta_i) \tag{4.20}$$

for some $\Delta_i \geq 1$.

IV.4.2. Delay Differential Equations

The effects of promotion and repression may not be immediately observed, as in the theoretical model. In fact, delay in physical systems is a commonly observed phenomenon (Segel, 1984; Rosenfeld and Alon, 2003; Bellen and Zennaro, 2003). In the case of one promoter gene, gene k, the model re-expressed with a term for time delay

$$\frac{\partial \mu_i(t)}{\partial t} = \beta_{ik}\mu_k(t-\delta_{ik})\lambda_i^t + \mu_i(t)\log(\lambda_i) - \alpha_i + \frac{\partial \theta_i(t)}{\partial t}, \tag{4.21}$$

$0 < \delta_{ik} < M_{ik}$. The decision to include a delay of course depends on the biology, and the experimental lapse between the time points. In experiments with small gaps between measurements, it may be considered quite reasonable to include a term for delay, although in other settings unsatisfactory. Reasonable constraints must be placed on delta, based on prior biological knowledge, for model identifiability.

IV.5. The S.O.S. Gene Network

The S.O.S. real time gene expression study (Friedman *et al.*, 2005) was conducted on a homemade microscope. Reporter e-coli cells were grown. Eight genes were monitored under two conditions, UV radiation low and high. Two replicate experiments were conducted for each UV level. Prior information on the S.O.S. gene network was obtained from Ronen *et al.* (2002), see Figure 24.

DNA damage induces an increase in expression and product of the recA gene, which cleaves to and suppresses lexA. In normal steady state, the lexA gene represses expression of downstream genes umuDC, urvA, urvD, uvrY, ruvA and polB. Increases in recA suppress lexA, allowing these genes to be expressed and carry out a programmed DNA damage response. Notice in the graph that no information is available on for promotion, only the repression.

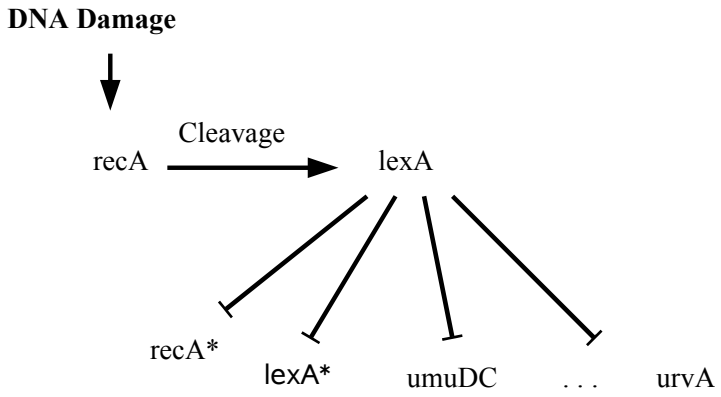

Fig. 24 S.O.S. gene network

IV.5.1. Model Fitting

A series of modeling assumptions were used to fit the S.O.S. data of Friedman *et al.* (2005), in a case study of model performance. Constrained priors are defined for all of the parameters. The constraints embody prior information about the relationships between the genes, i.e. relative influence such as promotion or degradation. The goal is to make inference concerning the relationships between the genes given prior assumptions about their relationships.

The differences $d_i(t)$,

$$d_i(t) = \frac{y_i(t+1) - y_i(t-1)}{2} \tag{4.22}$$

are modeled by Euler's method as

$$d_i(t) \sim N\left(\frac{\partial \mu_i(t)}{\partial t}, \sigma_i^2\right) \tag{4.23}$$

where

$$\begin{aligned}\frac{\partial \mu_i(t)}{\partial t} &= (b_i - \beta_{ij} y_j(t-\delta_{ij}))\, \lambda_i^t \\ &\quad + \mu_i(t)\log(\lambda_i) - \alpha_i + \frac{\partial \theta(t)}{\partial t}. \end{aligned} \tag{4.24}$$

The priors were specified as

$$\begin{aligned}\sigma_i^2 &\sim \text{InverseGamma}(0.0001, 0.0001) \\ \alpha_i &\sim \text{Uniform}(0,1) \\ \beta_{ik} &\sim \text{Uniform}(0,3) \\ \beta_{ij} &\sim \text{Uniform}(0,3) \\ \lambda_i &\sim \text{Uniform}(0,1) \\ \delta_{ik} &\sim \text{DiscreteUniform}(0,10) \\ \delta_{ij} &\sim \text{DiscreteUniform}(0,10). \end{aligned} \tag{4.25}$$

The constraints were discovered by trial and error. The parameters are updated by Metropolis in Gibbs, evaluating each MH step by $n_{MH} = 50$ iterations.

Fitting the periodicity in the data is complicated by the irregularity of the oscillations. A model was fit first to $y_i(t)$, in an MCMC framework. Residuals obtained at each iteration $r_i^{(b)}(t) = Y_i(t) - \mu_i^{(b)}(t)$ for $b = 1, ..., B$. The residuals, $r_i^{(b)}(t-2)$, included as linear regressors in fit to $d_i(t)$ lagged by one time unit of two minutes. Figures 25−26 show the numerical derivative of recA and lexA, over an interpolated version of $d_i(t)$. The overlay are the residuals $\hat{r}_i(t-1)$, derived from a spline regression fit to an interpolated version of the original data, $y_i(t)$. The residuals are rescaled to $-.5$.

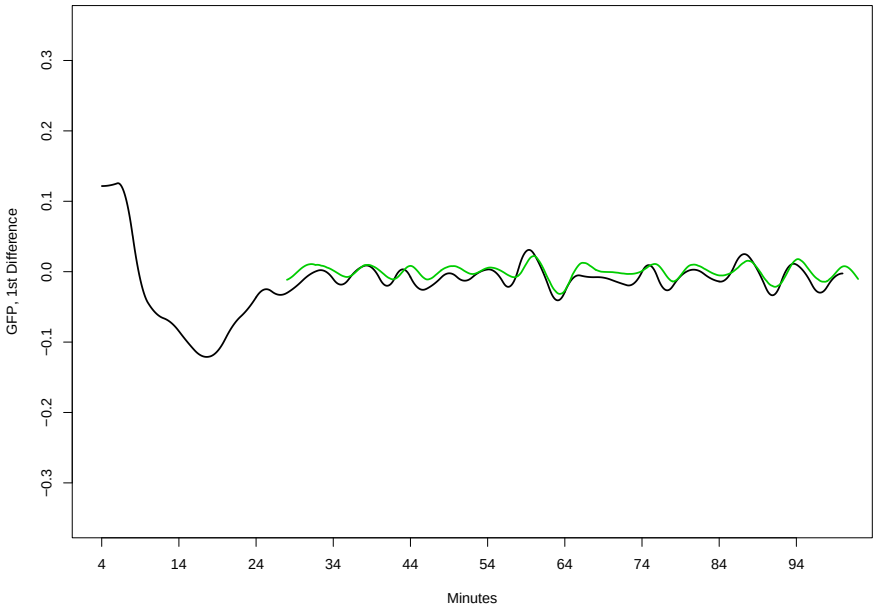

Fig. 25 Numerical derivative of recA (interpolated). Overlaid are rescaled and lagged residuals fit to promoter activity.

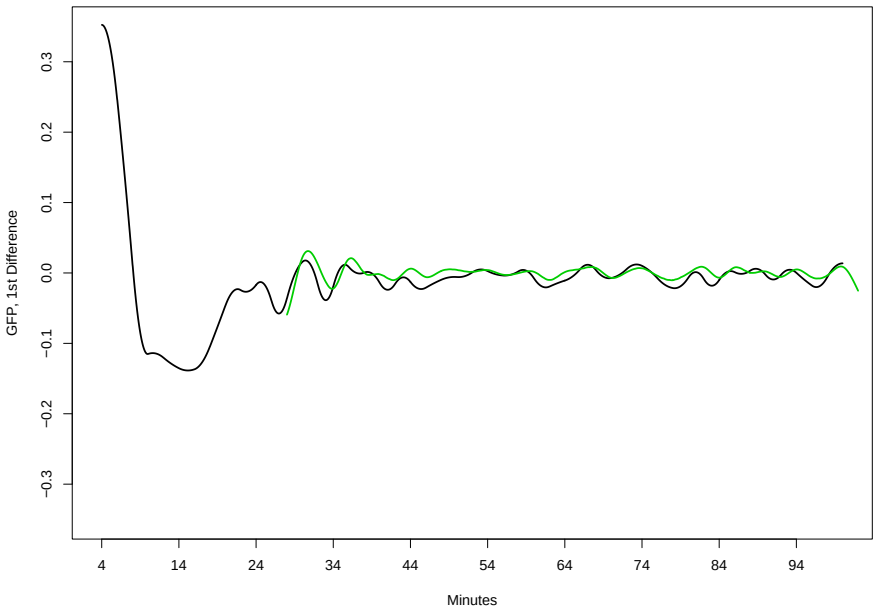

Fig. 26 Numerical derivative of lexA (interpolated). Overlaid are rescaled and lagged residuals fit to promoter activity.

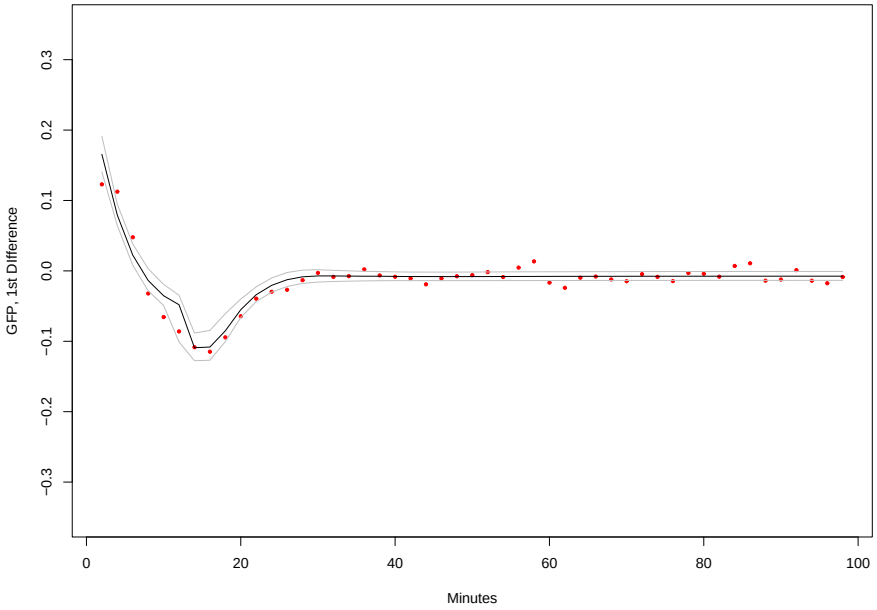

Fig. 27 1^{st} difference recA, UV = low, fitted and 90% C.I.

IV.5.2. S.O.S. Results

Model (4.13) was to fit the observed differences in recA, $d_i(t)$, averaged across both experiments with low UV radiation. A constant term was included for promotion activity b and a negative regression coefficient $-\beta$ was fit to the lexA series, allowing for a temporal delay. The fit was repeated for recA, averaged over both UV high experiments.

Figure 27 shows the fitted means with 90% credible intervals to recA with UV-Low. Notice that overall, the model is capturing the trend in the changes in promoter activity, although there is unaccounted for periodicity resembling irregular oscillations. Figure 28 shows the fitted posteriors for coefficients a, b, λ and β.

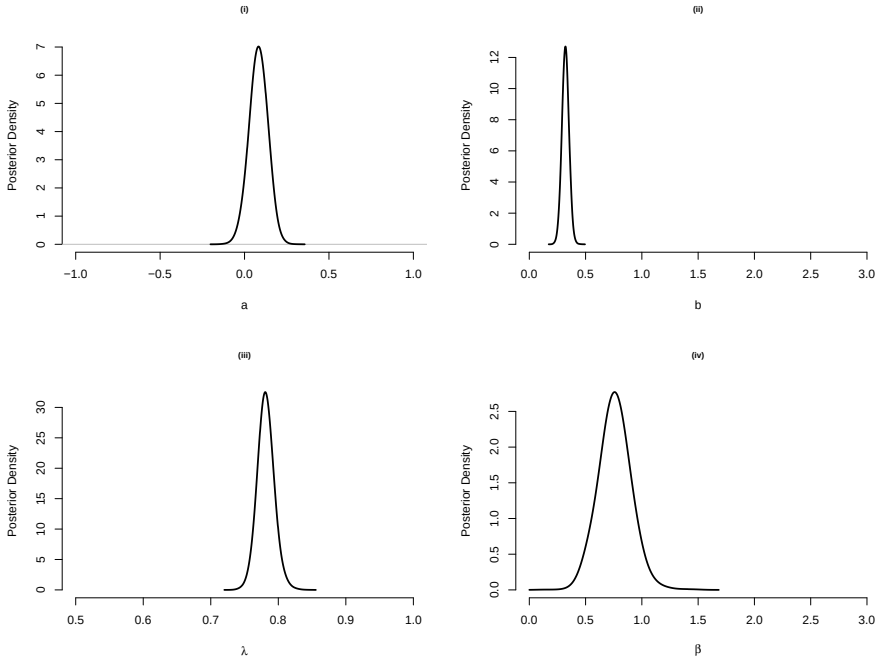

Fig. 28 recA, UV = low, posterior (i) α, (ii) b, (iii) λ, (iv) β.

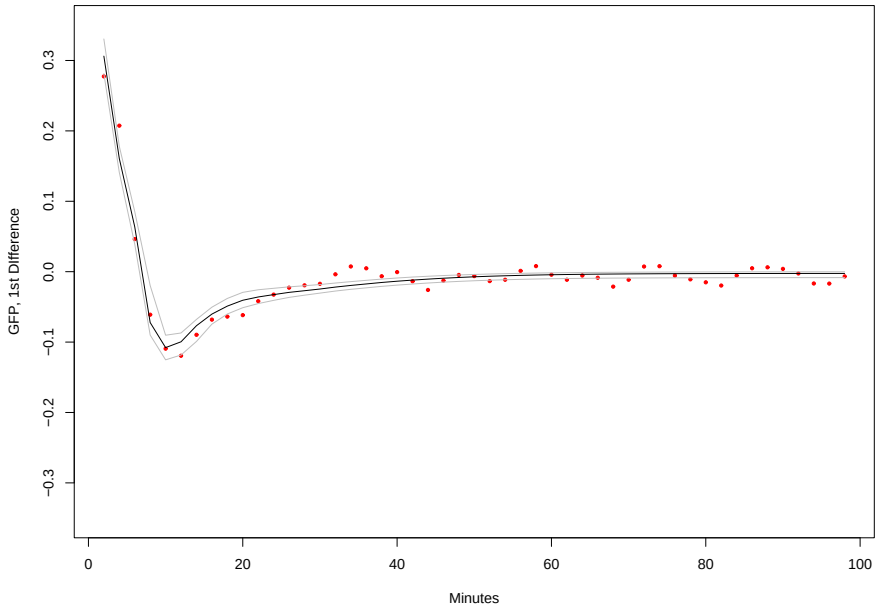

Fig. 29 1^{st} difference recA, UV = high, fitted and 90% C.I.

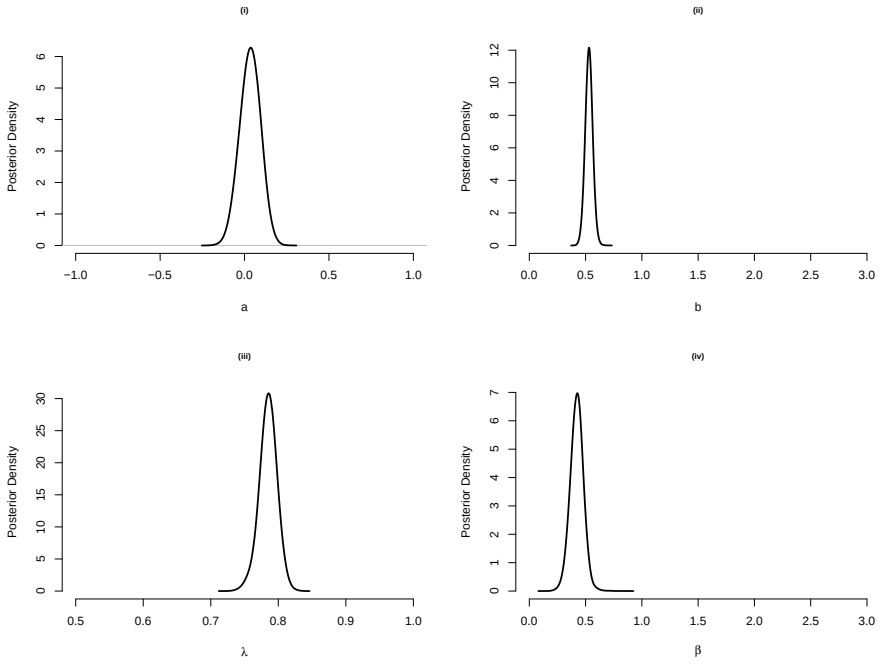

Fig. 30 recA, UV = high, posterior (i) α, (ii) b, (iii) λ, (iv) β.

Figures 29−30 show the posterior fits to recA for the UV high experiments. Notice that the parameter b accounting for an increase in promoter activity is shifted up in distribution with UV = High. The parameter accounting for the inhibition due to lexA promoter activity, appears to have a tighter posterior and a reduced center. The posteriors for a and λ are similar in both sets of experiments. Several important inferences are drawn. The level of decay in promoter activity due to recA levels is fairly constant in both sets of experiments, as seen in the posteriors for λ.

One of the critical questions is whether or not the interactions between lexA and treatment are real or the result of unaccounted for periodicity. The success of this approach rests on our ability to accurately fit and remove the residuals of $y_i(t)$. Figures 31−32 show the fit to the 1st difference in umuDC and urvD, $d_i(t)$, averaged

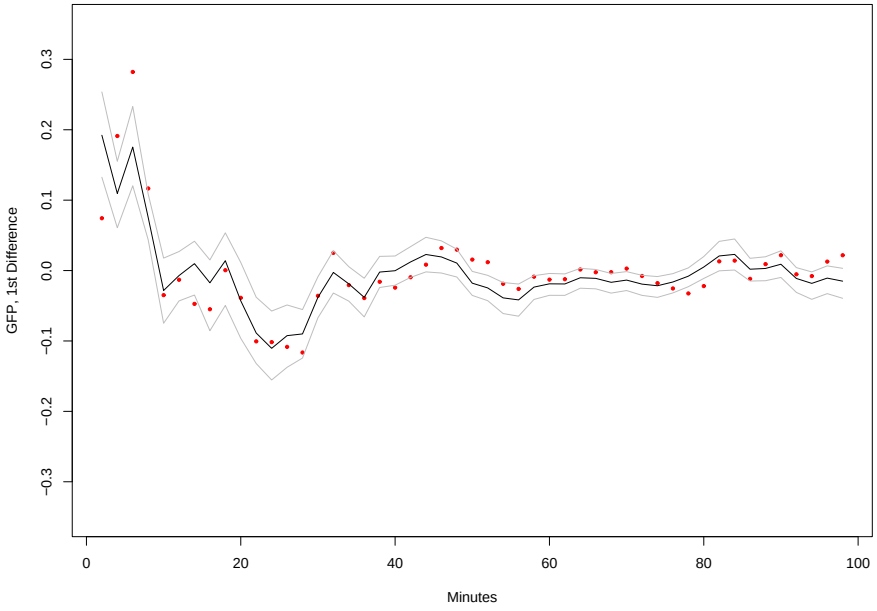

Fig. 31 1^{st} difference umuDC, UV = low

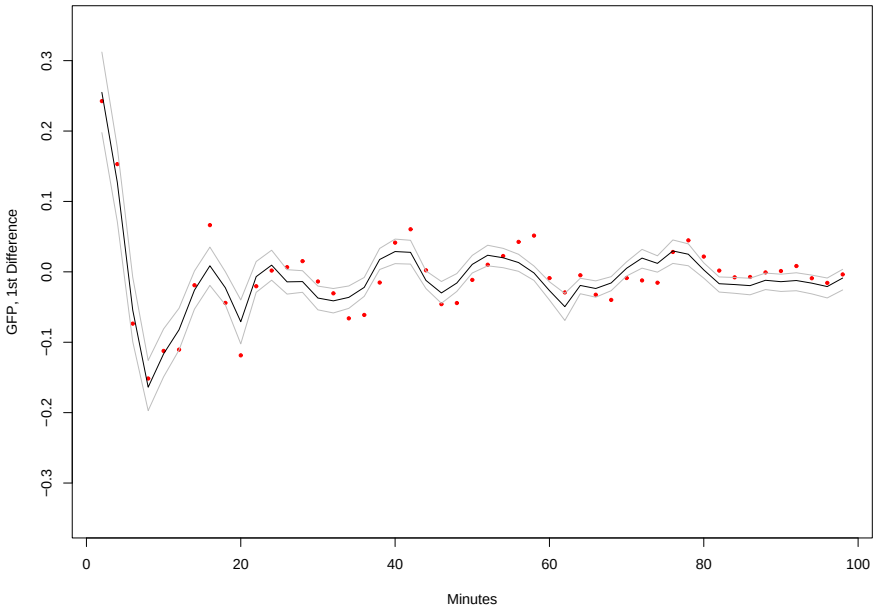

Fig. 32 1^{st} difference urvD, UV = low

across replicate experiments allowing for periodicity.

It appears that the trend in umuDC is ft reasonably well from these figures, both due to treatment and internal oscillatory behavior. The assumption that the periodicity is analogous to a sum of sin waves appears sensible. It may even be argued based on these figures that the residuals be lagged by more than one time unit, or place a prior on the lag. There are several important reasons why including the terms for periodicity in the model in this way. The results may vary due to unexplained variation, and therefore we would like to compare the results with and without accounting for periodicity in the differences. The periodicity is a function of the cells in the well, unrelated to the cells in the inhibitor gene, lexA, for instance, and therefore should depend on only the respective gene being modeled. The last reason is that once the effect of the treatment subsides, the cells go out of phase with one another, and difference in the cell cycles appear, as the trend in periodicity in internal promoter activity dominates.

Figures 33−34 shows the posteriors distributions for recA and umuDC, at UV = LLow, with and without accounting for periodicity. Figures 35−36 show the QQ-normal plots of the residuals fitting recA with and without accounting for periodicity. In recA the distribution for β_i attributable to the inhibition by lexA is shifted down slightly in distribution, while the effect of degradation λ_i is somewhat higher. If these results are accurate, then accounting for periodicity can make a difference for inference.

In order to make stronger inferences on the parameters, constrained models were fit to each of the genes, accounting for delayed inhibition from lexA and periodicity. In model 1 the parameters are assumed to be the same at both levels of UV radiation. In

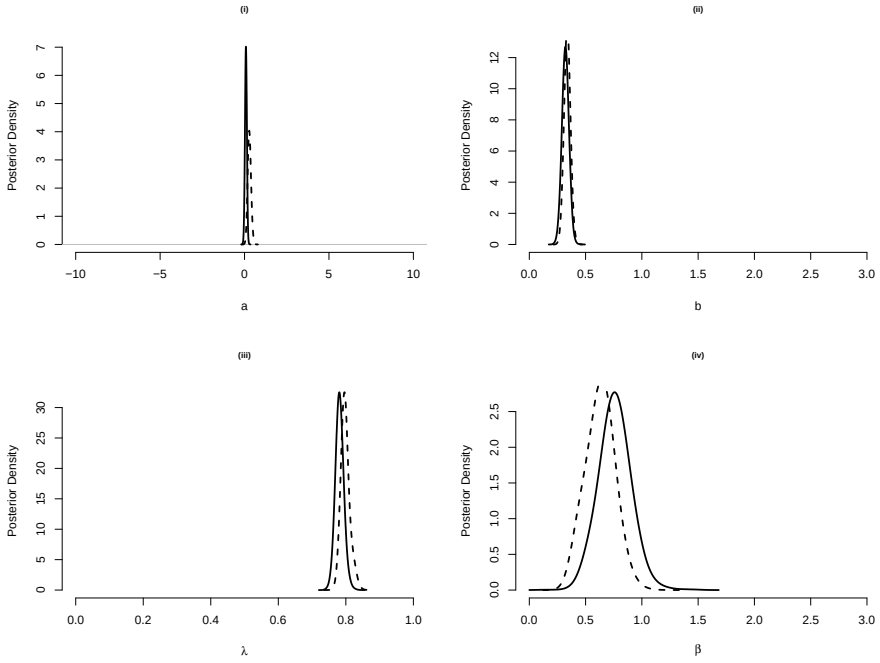

Fig. 33 recA (—) no periodicity and (- -) periodicity, UV = low: (i) a, (ii) b, (iii) λ, (iv) β.

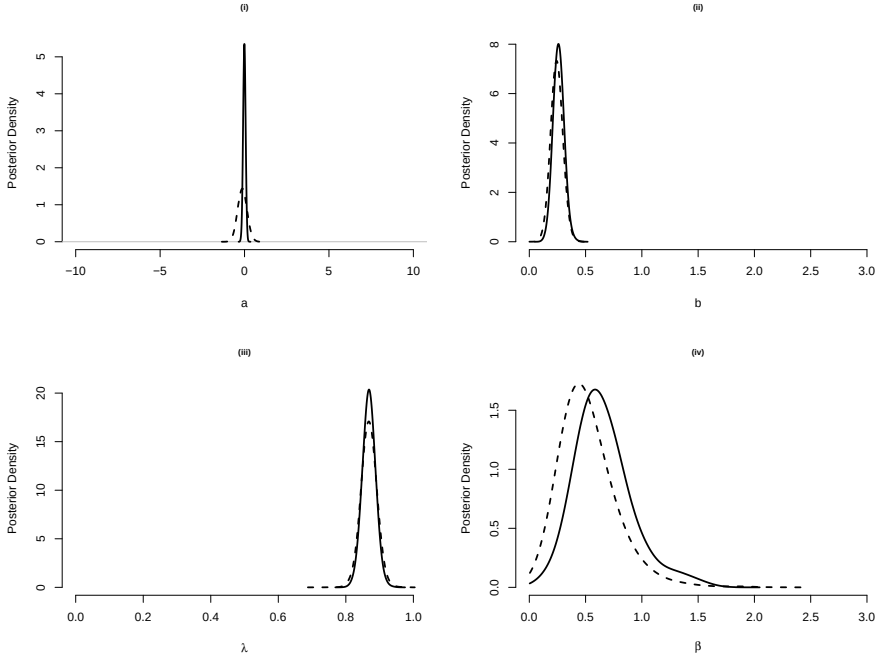

Fig. 34 umuDC (—) no periodicity and (- -) periodicity, UV = low: (i) a, (ii) b, (iii) λ, (iv) β.

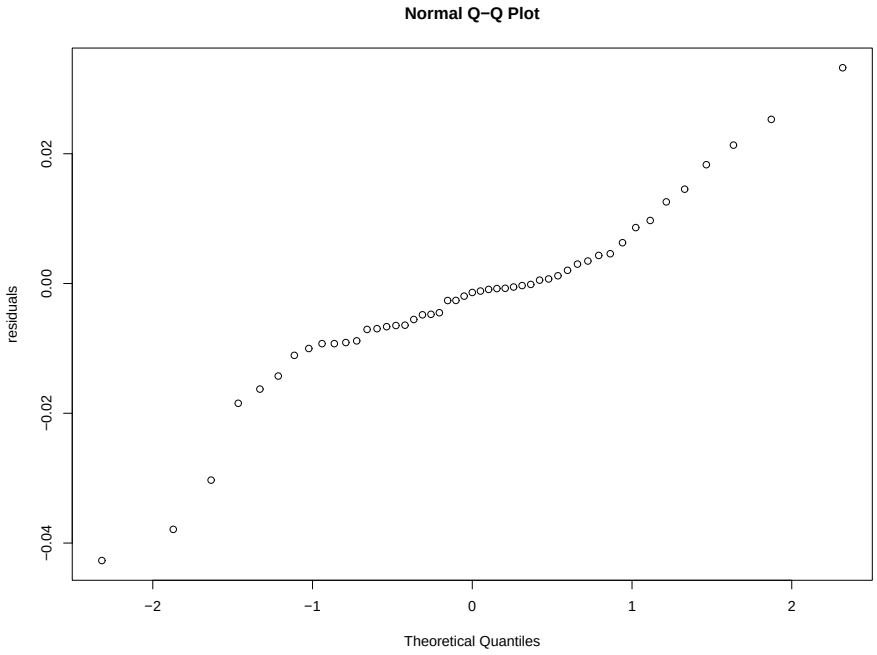

Fig. 35 QQ plot - recA residuals, UV = low, ignoring periodicity.

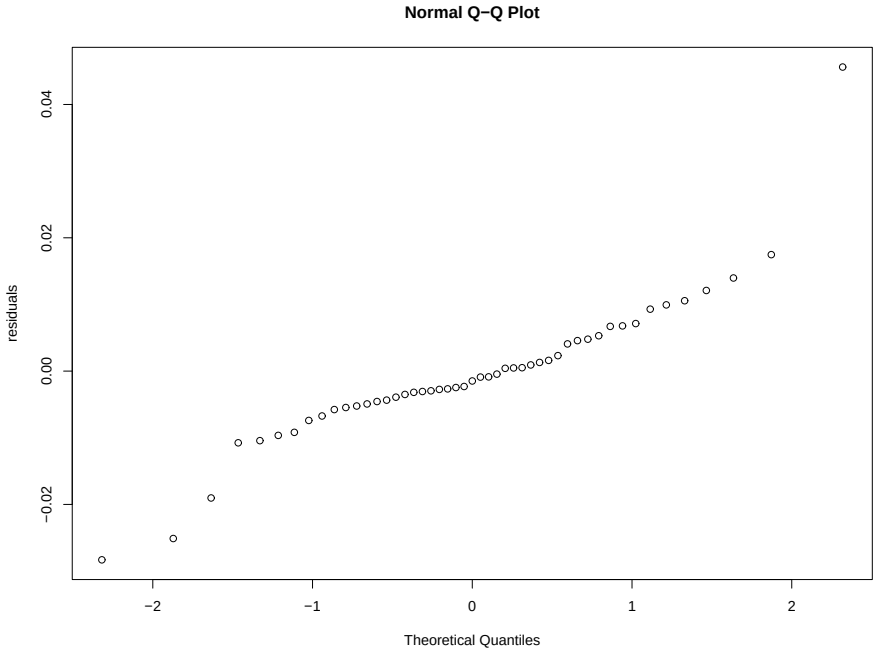

Fig. 36 QQ plot - recA residuals, UV = low, accounting for periodicity.

Table 13. Summary of fit, log marginal posterior

	Model 1	Model 2	Model 3
1. recA	97.33	112.48	135.19
2. umuDC	66.60	83.55	100.65
3. urvD	66.49	74.85	88.93

model 2, the terms for inhibition and degradation, β and λ, were restricted to be the same between UV-levels. Model 3 is the unconstrained model, fit separately between UV levels. The log of the marginal posterior fits to $d(t)$, obtained by MC integration, are shown in Table 13 for genes recA, umuDC, and urvD. There is strong evidence in favor of the unconstrained model, consistent with the historical S.O.S. gene pathway that depicts structural dynamics between all of the genes.

Summaries of the posterior densities, from the unconstrained model, accounting for periodicity and allowing for a delay in inhibition, are listed in Table 14. For each gene, 10,000 MCMC samples were generated in under ten minutes in R on a Dell 8200 with dual core Intel Xeon processors. There is a sharp contrast in promoter activity between UV levels in many of the genes. The effect of inhibition by lexA, as measured in the posterior density of β_i, is slightly less pronounced upon recA moving from low to high UV, while the time delay for inhibition is shorter. The inhibition effect of lexA on umuDC is not observed to be significantly different between UV levels, although the time delay is on average shorter, 18 minutes versus 8, between UV levels. The effect of lexA on urvD can be much stronger for high UV, as observed in the longer posterior right tail for β, and much stronger for urvA as well. This is an important result, consistent with the empirical evidence of a need for a time lag term in dynamic models of gene expression. Generally the term for degradation λ showed a very tight distribution, similar between UV radiation levels. This is reassuring as

Table 14. Fitted posterior, median and 90% C.I.

recA	UV Low	UV High
b	0.3372 (0.2939,0.3809)	0.5406 (0.4875,0.5946)
λ	0.7973 (0.7797,0.8239)	0.7820 (0.7492,0.8065)
β	0.6346 (0.4187,0.8325)	0.3990 (0.2823,0.5057)
δ	12 (10,12)	6 (6,8)
umuDC	UV Low	UV High
b	0.2435 (0.1617,0.3295)	0.5520 (0.4738,0.6348)
λ	0.8683 (0.8322,0.9044)	0.7495 (0.7125,0.7907)
β	0.4826 (0.1874,0.9717)	0.4805 (0.2117,2.5519)
δ	18 (12,20)	8 (6,18)
urvD	UV Low	UV High
b	0.4619 (0.3458,0.6048)	0.3465 (0.1595,0.5694)
λ	0.7894 (0.715,0.85)	0.7012 (0.4621,0.8568)
β	0.6064 (0.4386,0.8393)	0.7412 (0.3073,2.5142)
δ	4 (4,4)	4 (4,16)
uvrA	UV Low	UV High
b	0.5754 (0.5242,0.6274)	0.5903 (0.5283,0.6534)
λ	0.7309 (0.7124,0.7461)	0.789 (0.7694,0.8085)
β	1.2918 (0.8918,2.2655)	0.6994 (0.6178,0.7812)
δ	12 (12,14)	6 (6,6)
lexA	UV Low	UV High
b	0.6026 (0.5397,0.6763)	0.6885 (0.6197,0.751)
λ	0.7862 (0.7459,0.8105)	0.7521 (0.731,0.779)
β	0.5343 (0.436,0.6554)	0.6013 (0.4781,0.7166)
δ	6 (6,8)	6 (4,6)
uvrY	UV Low	UV High
b	0.2062 (0.0238,0.5772)	1.1318 (0.0935,2.522
λ	0.6718 (0.6321,0.7219)	0.2049 (0.0412,0.5485)
β	2.4348 (1.7974,2.9089)	1.4261 (0.1549,2.8243)
δ	0 (0,0)	0 (0,0)
ruvA	UV Low	UV High
b	0.1884 (0.0211,0.6372)	0.3198 (0.0514,1.1193)
λ	0.362 (0.2711,0.5043)	0.4761 (0.0474,0.7916)
β	2.6834 (1.3452,2.968)	1.4111 (0.1561,2.8347)
δ	0 (0,8)	12 (0,20)
polB	UV Low	UV High
b	0.1674 (0.0652,0.2668)	0.2096 (0.1225,0.3038)
λ	0.8488 (0.7707,0.9576)	0.8301 (0.7835,0.8773)
β	0.8067 (0.2987,1.9481)	0.9377 (0.4064,2.0766)
δ	10 (8,12)	18 (16,20)

we did not expect the effect of degradation to depend on UV level. The lexA gene appears to be inhibited more by recA at high UV radiation. The effects are not significant, as the posterior overlap substantially. For genes uvrY, ruvA and polB there appear to be UV-level effects in promotion.

IV.6. Discussion

A Bayesian model is proposed for inference with real-time gene expression data. The model was fit with MCMC methods in relatively short time with R scripts run in parallel on a Dell 8200 with dual core Intel Xeon processors. In the S.O.S. gene network data, the models offered generally were shown to capture the dynamic trends in gene expression over time. Treatment effects appear likely, although the benefit of accounting for internal periodicity within each respective reporter gene-cellular culture, is not yet clear. Interactions are present between lexA and treatment in several of the genes. Including time lag effects demonstrated superior data fitting overall, and as shown in the results tends to be significantly different from 0. Treatment effects and gene interactions were not starkly different accounting for periodicity, although under fitting was reduced dramatically.

Future directions include model validation studies. The results here may inform future experimental designs. Design issues are far from worked out, and will be the topic of future discussion for some time. The work here focused exclusively on fitting the theoretical dynamic models to each responder gene independently of the rest of the genes. Another future goal is to fit complete networks with the Bayesian model, in order to learn collectively of potential interactions between the genes and effects of interest.

In summary, theoretical models like the one proposed here offer great utility for learning in systems biology. As basic biological science advances, methodologies will be needed to discern the results and inform conclusions. Interpretable models do not always fit the data the best, although offer the clear advantage of producing results amenable to inference. Future studies will serve to validate and evolve theoretical models useful for GFP experiments.

CHAPTER V

SUMMARY AND CONCLUSIONS

As Bioinformatics experiments become more sophisticated, modeling assumptions will no doubt require change, affecting modeling strategy and choice. This is a very serious challenge for systems biology research, as the field is evolving so rapidly. Bayesian statistics offers many advantages for systems biology research. The flexibility of Bayesian methodologies alone, as demonstrated with the simulations and case studies offered here, show tremendous advantages for assessing the utility of prior information and modeling assumptions, in a framework that allows us to account for more uncertainty.

As we observed, the problem of detection of significant gene classes, defined from historical pathways, with noisy expression array data is complicated at many levels. At a theoretical level, the number of dependencies that may be modeled directly between genes is limited by sample size requirements. At the practical level there is uncertainty in the historical pathways. BLM allows for uncertainty at many levels, with the ability to borrow information across the genes in a class without sample size limitations. In the simulation and case studies, borrowing demonstrated improved sensitivity and specificity. A limitation of borrowing in this context is the question of how much to borrow. Sensitivity analysis was offered to better understand the role of borrowing between genes. Preliminary evidence with simulation and public data demonstrates a promising advantage for integrating historical pathway knowledge explicitely into microarray analysis.

A Bayesian change point model was proposed for learning about chromosomal aberration in heterogeneous patient populations. It was shown that the Bayesian approach is feasible, up to a needed strategy for exploring the posterior space of change point configurations. Borrowing information does improve probability estimates of the rates of chromosomal instability, through a variance reduction. We learned about the effect of missing real change points, and robustness of over-specification of the change points. In simulation, BCPA demonstrated a marginal though significant trade off in sensitivity with CBS.

Real-time gene expression experiments were introduced, offering the potential for learning the structural dynamics in gene expression. This technology is still in its infancy, and making sense of the experiments will require a broad set of skills. There is much uncertainty surrounding real-time gene expression experiments. Experimental design issues were discussed, as well as normalization. A mathematical model was introduced and inference performed in a Bayesian framework. In case study with the S.O.S. gene data (Friedman *et al.*, 2005), the model fit the temporal variation in the gene expression well and inferred treatment effects and gene interactions. Future model validation is needed, although the model appears promising.

In summary, theoretical models like the one proposed here offer great utility for learning in systems biology. As basic biological science advances, methodologies will be needed to discern the results and inform conclusions. Interpretable models do not always fit the data the best, although offer the clear advantage of producing inferences. Future studies will serve to validate the models. Data mining in high-throughput experimental results is a messy task, but the unique structure of the biological systems

we are interested in, we argue, can and should inform our search. Bioinformatics is challenged to integrate diverse sources of information, and provide the semantics where none exist for analysis. The Bayesian learning paradigm is a natural one in this context, to critically assess information for its utility and improve the power to learn in heterogeneous disease populations.

REFERENCES

Albertson, D.G. and Pinkel, D. (2003) Genomic microarrays in human genetic disease and cancer. *Human Molecular Genetics* **12**, 145-152.

Allocco, D.J., Kohane, I.S. and Butte, A.J. (2004) Quantifying the relationship between co-expression, co-regulation and gene function, *BMC Bioinformatics*, **5**.

Arya, M., Ahmed, H., Silhi, N., Williamson, M. and Patel, H.R. (2007) Clinical importance and therapeutic implications of the pivotal CXCL12-CXCR4 (Chemokine Ligand-Receptor) interaction in cancer cell migration. *Tumor Biol.*, **18**, 123-131.

Barry, W.T., Nobel, A.B. and Wright, F.A (2005) Significance analysis of functional categories in geneexpression studies: a structured permutation approach. *Bioinformatics*, **21**, 1943-1949.

Bellen, A. and Zennaro, M. (2003) *Numerical Methods for Delay Differential Equations*. New York: Clarendon.

Bharjwaj, N. and Lu, H. (2005) Correlation between gene expression profiles and protein protein interactions within and across genomes. *Bioinformatics*, **21**, 2730-2738.

Bhattacharjee, M., Pritchard, C.C., Nelson P.S. and Arjas, E. (2004) Bayesian integrated functional analysis of microarray data. *Bioinformatics*, **20**, 2943-2953.

Bilke, S., Chen, Q-R, Whiteford, C.C. and Khan, J. (2005). Detection of low level genomic alterations by comparative genomic hybridization based on cDNA micro-arrays. *Bioinformatics*, **21**, 111-1145.

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

Brower, V. (2006) Erythropoietin may impair, not improve, cancer survival. *Nat. Med.*, **9**, 1439-1439.

Camps, J., Morales, C., Prat, E., Ribas, M., Capella, G., Egozcue, J., Peinado, M.A. and Miró, R. (2004). Genetic evolution in colon cancer KM12 cells and metastatic derivatives. *Int. J. Cancer*, **110**, 869-874.

Chen T., He, H.L. and Church, G. (1999) Modeling Gene Expression with Differential Equations. In *Pac. Symp. Biocomp.*, 29-40.

Coffey, W.T., Yu, P. and Waldron, J.T. (2004) *The Langevin Equation: With Applications to Stochastic Problems in Physics, Chemistry and Electrical Engineering, Second Edition*. London: World Science Publishing Co., Pte. Ltd.

Conlon, E.M., Song, J.J. and Liu J.S. (2006) Bayesian models for pooling microarray studies with multiple sources of replications. *BMC Bioinformatics*, **7**.

Curtis, R.K., Oresic, M. and Vidal-Puig, A. (2005) Pathways to the analysis of microarray data. *Trends in Biotechnology*, **23**, 429-435.

Dennis, G., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C. and Lempicki, R.A. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.*, **4**, 3.

Dobra, A., Jones, B., Hans, C., Nevins, J. and West, M. (2004) Sparse graphical models for exploring gene expression data. *J. of Mul. Anal.*, **90**, 196-212.

Dojer, N., Gambin, A., Mizera, A., Wilczyski, B. and 3 Tiuryn, J. (2006) Applying dynamic Bayesian networks to perturbed gene expression data. *BMC Bioinformatics* **7**.

Fridlyand, J., Snijders, A.M., Pinkel, D., Albertson, D.G. and Jain, A.J. (2004). Hidden Markov Models approach to the analysis of array CGH data. *J. Mul. Anal.*, **90**, 132-153.

Friedman, N., Vardi, S., Ronen, M. Alon, U. and Stavans, J. (2005) Precise temporal modulation in the response of the SOS DNA repair network in individual bacteria. *Plos. Bio.*, **3**, e238.

Gaile, D.P., Miecznikowski, J., Conroy, J. and Nowak, N.J. (2006) Fitting array comparative genomic hybridization data with a segmented mixture model, Technical report 06-03. Buffalo, NY: Department of Biostatistics, State University of New York.

Gelfand, A.E., Carlin, B.P. and Smith, A.F.M. (1992). Hierarchical Bayesian analysis of changepoint problems. *Appl. Statist.*, **41**, 389-405.

Geller, S.C., Gregg, J.P., Hagerman, P. and Rocke, D.M. (2003) Transformation and normalization of oligonucleotide microarray data. *Bioinformatics*, **19**, 1817-1823.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004) *Bayesian Data Analysis*, Boca Raton, Fla. : Chapman & Hall/CRC.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. on Patt. Anal. and Mach. Intell.*, **6**, 721-741.

George D. (2003) Targeting PDGF receptors in cancer–rationales and proof of concept clinical trials. *Adv. Exp. Med. Biol.*, **532**, 141-51.

George, E.I. and McCulloch, R.E. (1993) Variable selection via Gibbs sampling. *J. Am. Statist. Ass.* **88**, 881-889.

Geweke, J. and Tanizaki, H. (2001) Bayesian estimation of state-space models using the Metropolis-Hastings algorithm within gibbs sampling. *Comp. Statist. and Data Anal.*, **37**, 151-170.

Ghahramani, Z. (1997) Learning dynamic Bayesian networks. In *Adaptive Processing of Sequences and Data Structures . Lecture Notes in Artificial Intelligence* pp. 168-197. Berlin: Springer-Verlag.

Gold, D.L., Wang, J., Coombes, K.R. (2005) Inter-gene correlation on oligonucleotide arrays: how much does normalization matter? *The Am. J. of Pharmacogenomics*, **5**, 271-279.

Gold, D.L., Coombes, K.R., Wang, J. and Mallick, B. (2007) Enrichment analysis in high-throughput genomics – accounting for dependency in the NULL. *Brief. in Bioinformatics*, **8**, 71-77.

Grodins, F.S. (1963) *Control Theory and Biol. Systems*. New York: Columbia University Press.

Guha, S., Li, Y. and Neuberg, D. (2005). Bayesian Hidden Markov modeling of obscured spatial transitional data with application to the analysis of array CGH data, Harvard University Biostatistics Working Paper Series, Harvard University Paper 24.

Hodges, J.S. (1998). Some algebra and geometry for hierarchical linear models, applied to diagnostics.*J. R. Statist. Soc. B*, **60**, 497-536.

Hsu, L., Self, S.G., Grove, D., Randolph, D., Wang, K., Delrow, J.J., Loo, L. and Porter, P. (2005). Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, **6**, 211-226.

Huang, T., Wu, B., Lizardi, P. and Zhao, H. (2005). Detection of DNA copy number alterations using penalized least squares regression. *Bioinformatics*, **21**, 3811-3817.

Hupe, P., Stransky, N., Thiery, J.P., Radvanyi, F. and Barillot, E. (2004). Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**, 3413-3422.

King, K.R., Wang, S., Irimia, D., Jayaraman, A., Toner, M. and Yarmush, M.L. (2007) A high-throughput microfluidic real-time gene expression living cell array. *The R. Soc. of Chem.*, **6**, 1-10.

Klevecz, R.R., Bolen, J., Forrest, G. and Murray D.B. (2004). A genomewide oscillation in transcription gates DNA replication and cell cycle. *Proc. Natn. Acad. Sci. USA*, **101**, 1200-1205.

Lai, W., Johnson, M.D., Kucherlapati, R. and Park, P. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, **21**, 3763-3770.

Lauritzen, S. (1996) *Graphical Models*. Oxford Statistical Science Series **17**, Oxford: Clarendon Press, 1996.

Lenburg, M.E. Liou, L.S., Gerry, N.P., Frampton, G.M., Cohen, H.T. and Christmans M.F. (2003) Previously unidentified changes in renal cell carcinoma gene expression identified by parametric analysis of microarray data. *BMC Cancer*, **3**.

Lengauer, C., Kinzler, K.W. and Vogelstein, B. (1998) Genetic instabilities in human cancers. *Nat.*, **396**, 643-649.

Lewin, A., Richardson, S., Marshall C., Glazier A. and Aitman T. (2005) Bayesian modelling of differential gene expression. *Biometrics* (in press).

Liao, J.C., Boscolo, R., Yang, Y-L, Tran, L.M., Sabatti, C. and Roychowdhury V.P. (2003) Network component analysis: Reconstruction of regulatory signals in biological systems. *Proc. Natn. Acad. Sci. USA*, **100**, 15522-15527.

Lindley, D.V. and Smith, A.F.M. (1972) Bayes estimates for the linear model. *J. R. Statist. Soc. B*, **34**, 1-42.

Lu, Y., Liu, P., Xiao, P. and Deng, H. (2005) Gene expression Hotellings T2 multivariate profiling for detecting different expression in microarrays. *Bioinformatics*, **21**, 3105-3113.

Moloshok, T.D., Klevecz, R.R., Grant, J.D., Manion, F.J., Speier IV, W.F. and Ochs, M.F. (2002) Application of Bayesian decomposition for analyzing microarray data. *Bioinformatics*, **18**, 566-575.

Nagaraj, V., O'Flanagan, R.A., Bruning, A.R., Mathias , J.R., Vershon, A.K. and Sengupta, A.M.(2004) Combined analysis of expression data and transcription factor binding sites in the yeast genome. *BMC Bioinformatics*, **5**.

Nakao, K., Mehta, K.R., Fridlyand, J., Moore, D.H., Jain, A.N., Lafuente, A., Wiencke, J.W., Terdiman, J.P. and Waldman, F.M. (2004). High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. *Carcinogenesis*, **25**, 1345-1357.

Natrajan, R., Williams, R.D., Hing, S.N., Mackay, A., Reis-Filho J.S., Fenwick, K., Iravani, M. Valgeirsson, H., Grigoriadis, A., Langfor, C.F., Dovey, O., Gregory, S.G., Weber, B.L., Ashworth, A., Grundy, P.E., Pritchard-Jones, K. and Jones, C. (2006). Array CGH profiling of favorable histology Wilms Tumours reveals novel gains and losses associated with relapse. *J. of Path.*, **210**, 49-58.

Olshen, A.B. and Venkatraman, E.S. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557-572.

Pan, W. (2006) Incorporating gene functional annotations in detecting differential gene expression. *J.R. Statist. Soc. C*, **55**, 301-316.

Parmigiani, G., Garrett, E.S., Anbazhagan, R. and Gabrielson, E. (2002) A statistical framework for expression-based molecular classification in cancer. *J.R. Statist. Soc. B*, **64**, 717-736.

Perrin, B., Ralaivola, L., Mazurie, A., Bottani, S., Jacques, M. and dAlch-Buc, F.(2003) Gene networks inference using dynamic Bayesian networks. *Bioinformatics*, **19**, 138-148.

Picard, F., Robin, S., Lavielle, M., Vaisse, C. and Daudin, J.J. (2005). A statistical approach for array CGH data analysis. *BMC Bioinformatics* **6**.

Pounds, S., Cheng, C. (2004) Improving false discovery rate estimation. *Bioinformatics*, **20** 1737-1745.

Raftery, A.E., Madigan, D. and Hoeting J.A. (1997) Bayesian model averaging for linear regression models. *J. Am. Statist. Ass.* **92**, 179-191.

Roberts, C.P., Rydan, T. and Titterington, D.M. Bayesian inference in hidden Markov models through jump Markov chain Monte Carlo. *J.R. Statist. Soc. B* **62** , 57-75.

Ronen, M., Rosenberg, R., Shraiman, B.I., and Alon, U. (2002) Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Natn. Acad. Sci. USA*, **99**, 10555-10560.

Rosenfeld, N. and Alon, U. (2003) Response delays and the structure of transcription networks. *J. Mol. Bio.*, **329**, 645-654.

Schrader, A., Lechner, O., Templin, M., Dittmar, K., Machtens, S., Mengel, M., Probst-Kepper, M., Franzke, A., Wollensak, T., Gatzlaff, P., (2002) CXCR4/CXCL12 expression and signalling in kidney cancer. *Br. J. Cancer*, **86**, 1250-1256.

Segel, L.A. (1984) *Mathematical Models for Molecular and Cellular Biology*. Cambridge, MA: Cambridge University Press.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde A. (2002) Bayesian measures of model complexity and fit. *J.R. Statist. Soc. B*, **64**, 583-639.

Sun, N., Carroll, R.J. and Zhao, H. (2006) Bayesian error analysis model for reconstructing transcriptional regulatory networks. *Proc. Natn. Acad. Sci.*, **103**, 7988-7993.

Tadesse, M., Ibrahim, J.G., Gentelman, R., Chiaretti, S., Ritz, J. and Foa R. (2005) Bayesian error-in-variable survival model for the analysis of GeneChip arrays. *Biometrics*, **61**, 488-497.

Tamada, Y., Kim, S.Y., Bannai, H., Imoto, S., Tashiro, K., Kuhara, S. and Miyano, S. (2003) Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*, **19**, 227-236.

Van der Laan, M. and Bryan, J. (2001), Gene expression analysis with the parametric bootstrap. *Biostatistics*, **2**,445-461.

Wang, P., Kim, Y., Pollack, J., Narasimhan, B. and Tibshirani, R. (2005). A method for calling gains and losses in array CGH data. *Biostatistics*, **6**, 45-58.

Whittemore, A.S. (2007) A Bayesian false discovery rate for multiple testing. *J. of Appl. Statist.* **34**, 1-9.

Wilkinson, D.J., and Yeung, S.K.H. (2002). Conditional simulation from highly structured Gaussian systems,with application to blocking-MCMC for the Bayesian analysis of very large linear models. *Statist. and Comput.*, **12**, 287-300.

Willenbrock, H. and Fridlyand, J. (2005). A comparison study: Applying segmentation to array CGH data for downstream analyses, *Bioinformatics*, **21**, 4084-4091.

Wu, T.T. (2001) *Analytical Mol. Biology*. Boston: Kluwer Academic Publishers.

Yang, Y.H. and Speed, T. (2002) Design issues for cDNA microarray experiments. *Nat. Rev. Genet.*, **3**, 579-88.

Zaslaver, A., Mayo, A.E., Rosenberg, R., Bashkin, P., Sberro, H., Tsalyuk, M., Suretter, M.G. and Alon U. (2004) Just-in-time transcription program in metabolic pathways. *Nat. Gen.*, **5**, 486-491.

Zaslaver, A., Bren, A., Ronen, M., Shalev, I., Kikoin, I., Shavit, S., Liebermeister, W., Surette, M. and Alon, U. (2006) A comprehensive library of fluorescent transcriptional reporters for Escherichia coli. *Nat. Meth.*, **3**, 623-628.

Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, **67**, 301-320.

APPENDIX A

As discussed in Chapter II, the full conditional posterior distributions for BLM with extentions are offered here. In order to avoid redundancy, only the posterior distributions that are different between the respective extentions are listed. Let us begin with the the original form, BLM1. In the original form, Z is assumed completely known. In this case, the conditional posteriors of σ_i^2 and β_i for $i = 1, ..., n$ are

$$\sigma_i^{2(t)} | \theta^{(t-1)}, \beta^{(t-1)}, Y \quad \sim \quad \text{IG}(\gamma_{1i}^*, \gamma_{2i}^*) \tag{A.1}$$

for,

$$\begin{aligned} \gamma_{1i}^* &= \frac{\gamma 1i + J \cdot K}{2} \\ \gamma_{1i}^* &= \frac{\gamma 2i + \sum_{jk} \left(Y_{ijk} - X_{jk} \beta_{ij}^{(t-1)} \right)^2 + \omega^{-1} \sum_j \left(\beta_{ij}^{(t-1)} - Z_i \theta_j^{(t-1)} \right)^2}{2} \end{aligned} \tag{A.2}$$

and,

$$\beta_i^{(t)} | \theta^{(t-1)}, \sigma_i^{2(t)}, Y \quad \sim \quad N(\beta_i^*, \Sigma_i^*) \tag{A.3}$$

for,

$$\begin{aligned} \Sigma_i^* &= \sigma_i^{2(t)} \left(X'X + \omega^{-1} \right)^{-1} \\ \beta_i^* &= \left(X'X + \omega^{-1} \right)^{-1} \left(X'Y_i + \omega^{-1} Z_i \theta^{(t-1)} \right). \end{aligned} \tag{A.4}$$

Define the matrix T at iteration t as $T^{(t)} = \text{diag}(\omega\sigma_1^{2(t)}, \omega\sigma_2^{2(t)}, ...)^{-1}$. The parameter vectors θ_j, for $j = 1, ..., J$, are updated at iteration t as

$$\theta_j^{(t)}|\beta^{(t)}, T^{(t)}, Y \sim N\left(\theta_j^*, \Omega_j^*\right) \tag{A.5}$$

for,

$$\begin{aligned} \Omega_j^* &= \left(Z'T^{(t)}Z + \Omega_o^{-1}\right)^{-1} \\ \theta_j^* &= \left(Z'T^{(t)}Z + \Omega_o^{-1}\right)^{-1}\left(Z'T^{(t)}\beta_j^{(t)} + \Omega_o^{-1}\theta_o\right). \end{aligned} \tag{A.6}$$

For the case of uncertainty on the sign of Z, BLM2, Gibbs sampling proceeds as above, except with the inclusion of auxiliary variables ξ_i, for $i = 1, ..., n$. Without loss of generality let $J = 2$ indicating just one treatment effect versus a control. For genes $i = 1, ..., n$ at replication t the ξ_i's are updated, and subsequently the mixture weights, as

$$\begin{aligned} P_i^{(t)} &= \frac{\pi_i^{(t-1)} N\left(\beta_i^{(t-1)}|Z_i\theta_i^{(t-1)}, Y\right)}{\pi_i^{(t-1)} N\left(\beta_i^{(t-1)}|Z_i\theta_i^{(t-1)}, Y\right) + \left(1 - \pi_i^{(t-1)}\right) N\left(-\beta_i^{(t-1)}|Z_i\theta_i^{(t-1)}, Y\right)} \\ \xi_i^{(t)}|P, Y &\sim \text{Bern}\left(P_i^{(t)}\right) \\ \pi_i^{(t-1)}|\xi, Y &\sim \text{Beta}\left(\nu_1 + \xi_i^{(t)}, \nu_2 + (1 - \xi_i^{(t)})\right). \end{aligned} \tag{A.7}$$

A new variable q_i, defined as $sign(Z_i)$, is updated automatically at iteration t as $q_i^{(t)} = 2\xi_i^{(t)} - 1$. The parameters σ_i^2, β_i and θ_j may now be updated conditionally as

$$\sigma_i^{2(t)}|\theta^{(t-1)}, \beta^{(t-1)}, Y \sim \text{IG}(\gamma_{1i}^*, \gamma_{2i}^*) \tag{A.8}$$

for,

$$\begin{aligned}\gamma^*_{1i} &= \frac{\gamma 1i + m \cdot K}{2} \\ \gamma^*_{2i} &= \frac{\gamma 2i + \sum_{jk}\left(Y_{ijk} - X_{jk}\beta^{(t-1)}_{ij}\right)^2 + \omega^{-1}\sum_j\left(\beta^{(t-1)}_{ij} - q^{(t)}_i Z_i \theta^{(t-1)}_j\right)^2}{2}\end{aligned} \quad \text{(A.9)}$$

and

$$\beta^{(t)}_i | \theta^{(t-1)}, \sigma^{2(t)}_i, Y \sim N(\beta^*_i, \Sigma^*_i) \quad \text{(A.10)}$$

for,

$$\begin{aligned}\Sigma^*_i &= \sigma^{2(t)}_i \left(X'X + \omega^{-1}\right)^{-1} \\ \beta^*_i &= \left(X'X + \omega^{-1}\right)^{-1}\left(X'Y_i + \omega^{-1} q^{(t)}_i Z_i \theta^{(t-1)}\right)\end{aligned} \quad \text{(A.11)}$$

and

$$\theta^{(t)}_j | \beta^{(t)}, T^{(t)}, Y \sim N\left(\theta^*_j, \Omega^*_j\right) \quad \text{(A.12)}$$

$$\begin{aligned}\Omega^*_j &= \left(Z'T^{(t)}Z + \Omega^{-1}_o\right)^{-1} \\ \theta^*_j &= \left(Z'T^{(t)}Z + \Omega^{-1}_o\right)^{-1}\left(Z'T^{(t)}Q(t)_i\beta^{(t)}_j - \eta\right)\end{aligned} \quad \text{(A.13)}$$

where $Q(t)_i = \text{diag}\{q^{(t)}\}$.

In the final extension, BLM3, an additional variable selection step is included. The

coefficients, β_{ijk} are assumed follow one of 3 states, a normal distribution with mean $Z_i\theta$, a point mass at 0 or a normal distribution with mean $-Z_i\theta$. The states variables Ψ_{ih} and mixture weights p_{ih}, for $h = 1, 2, 3$, are updated as

$$\begin{aligned}
P_{i1}^{(t)} &\propto \prod_k N(Y_{ijk}; XB_{ij}^{(t)}, \sigma_i^{(t)2}) \cdot N(B_{ij}^{(t)}; Z_i\theta_{ij}^{(t)}, \omega\sigma_i^{(t)2}) \\
P_{i2}^{(t)} &\propto \prod_k N(Y_{ijk}; 0, \sigma_i^{(t)2}) \\
P_{i3}^{(t)} &\propto \prod_k N(Y_{ijk}; XB_{ij}^{(t)}, \sigma_i^{(t)2}) \cdot N(B_{ij}^{(t)}; -Z_i\theta_{ij}^{(t)}, \omega\sigma_i^{(t)2}) \\
\psi_i^{(t)}|P, Y &\sim \text{Multi}\left(3; P_i^{(t)}\right) \\
p_i^{(t-1)}|\psi_i, Y &\sim \text{Dir}\left(cA + R^{(t)_i})\right). \qquad \text{(A.14)}
\end{aligned}$$

for $R_i^{(t)}$, a vector of dimension 3, with $R_{ih}^{(t)} = 1$ where $\psi_i^{(t)} = h$, and $R_{ih'}^{(t)} = 0$ all $h' \neq h$. The conditional posteriors are updated similarly as before, except in this case $\beta_{ij}^{(t)} = 0$ if $\psi_i^{(t)} = 2$.

APPENDIX B

As described in Chapter II, the full conditional posterior distributions for the means, variances and hyper means are listed here. Given a change point configuration, denoted $\underline{\xi}$, the full conditional posterior distribution of the variance parameters σ^2_{hk}, for $h = 1, ..., H$ and $k = 1, ..., M_h$, are updated at iteration t as

$$\sigma^{2(t)}_{hk}|\cdot \quad \sim \quad 1/\text{Gamma}\left(\gamma^*_{1hk}, \gamma^*_{2hk}\right) \tag{B.1}$$

$$\begin{aligned} \gamma^*_{1hk} &= \frac{1}{2}\left(\gamma_1 + n_h k + 1\right) \\ \gamma^*_{2hk} &= \frac{1}{2}\left(\gamma_2 + \sum_i I\left(\xi_{hi} = k\right)\left(Y_{hi} - \mu^{(t-1)}_{hk}\right)^2 + W^{-1}\left(\mu^{(t-1)}_{hk} - \eta^{(t-1)}_{hks}\right)^2\right) \end{aligned} \tag{B.2}$$

The full conditional posterior distribution for the means μ_{hk}'s, are updated at iteration t by

$$\mu^{(t)}_{hk}|\cdot \quad \sim \quad N\left(\mu^*_{hk}, \sigma^{2*}_{hk}\right) \tag{B.3}$$

for

$$\begin{aligned} \mu^*_{hk} &= \frac{W \cdot \bar{Y}_{hk} + \eta^{(t-1)}_{hks}}{W+1} \\ \sigma^{2*}_{hk} &= \frac{W}{1+W} \cdot \frac{\sigma^{2(t)}_{hk}}{n_{hk}}, \end{aligned} \tag{B.4}$$

where

$$\eta_{hks}^{(t-1)} = \frac{\sum_{\tilde{k}=1}^{\tilde{M}} n_{h\tilde{k}} \theta_{s\tilde{k}}^{(t-1)}}{n_{hk}}. \tag{B.5}$$

The state parameters, ψ_{hk}'s, are updated at iteration t as

$$\begin{aligned} P(\psi_{hk}^{(t)} = 1) &= \frac{F_{hk}^{(t)}(0)}{1 + f_{hk}^{(t)}(0)} \\ P(\psi_{hk}^{(t)} = 2) &= \frac{f_{hk}^{(t)}(0)}{1 + f_{hk}^{(t)}(0)} \\ P(\psi_{hk}^{(t)} = 3) &= \frac{1 - F_{hk}^{(t)}(0)}{1 + f_{hk}^{(t)}(0)} \end{aligned} \tag{B.6}$$

where $F_{hk}^{(t)}(0)$ and $f_{hk}^{(t)}(0)$ are the CDF and PDF of the normal distribution evaluated at 0, with mean and variance $\left(\mu_{hk}^{(t)}, W \cdot \sigma_{hk}^{2(t)}/n_{hk}\right)$.

The population wide sub-region level means θ's are updated conditionally as

$$\begin{aligned} \theta_{1\tilde{k}}^{(t)} &\sim N\left(\frac{\phi_{1\tilde{k}}^{(t)}}{\nu_{1\tilde{k}}^{(t)}}, \frac{1}{\nu_{1\tilde{k}}^{(t)}}\right) \cdot I\left(\theta_{1\tilde{k}}^{(t)} < 0\right) \\ \theta_{3\tilde{k}}^{(t)} &\sim N\left(\frac{\phi_{3\tilde{k}}^{(t)}}{\nu_{3\tilde{k}}^{(t)}}, \frac{1}{\nu_{3\tilde{k}}^{(t)}}\right) \cdot I\left(\theta_{1\tilde{k}}^{(t)} > 0\right) \end{aligned} \tag{B.7}$$

where,

$$\begin{aligned} \nu_{1\tilde{k}}^{(t)} &= \sum_{hi} I\left(\psi_h(\xi_{hi} = k \cap \tilde{\xi}_i = \tilde{k}) = 1\right) / \left(W \cdot \sigma_{hk}^{2(t)}\right) \\ \phi_{1\tilde{k}}^{(t)} &= \sum_{hi} I\left(\psi_h(\xi_{hi} = k \cap \tilde{\xi}_i = \tilde{k}) = 1\right) \cdot \mu_{hk}^{(t)} / \left(W \cdot \sigma_{hk}^{2(t)}\right) \end{aligned} \tag{B.8}$$

and

$$\begin{aligned}
\nu_{3\tilde{k}}^{(t)} &= \sum_{hi} I\left(\psi_h(\xi_{hi} = k \cap \tilde{\xi}_i = \tilde{k}) = 3\right) / \left(W \cdot \sigma_{hk}^{2(t)}\right) \\
\phi_{3\tilde{k}}^{(t)} &= \sum_{hi} I\left(\psi_h(\xi_{hi} = k \cap \tilde{\xi}_i = \tilde{k}) = 3\right) \cdot \mu_{hk}^{(t)} / \left(W \cdot \sigma_{hk}^{2(t)}\right). \qquad \text{(B.9)}
\end{aligned}$$

VITA

DAVID L. GOLD

Department of Statistics

Texas A&M University

3143 TAMU

College Station, TX 77843-3143

c/o Bani Mallick, Ph.D.

EDUCATION

2007 Ph.D. Statistics, Texas A&M University

2000 Master of Science, Statistics, Texas A&M University

1996 Bachelor of Arts, Economics, The University of Texas at Austin

RESEARCH INTERESTS

Methodology: Bayesian Hierarchical Modeling, Model Validation, Statistical Data Mining Applications: High-throughput Genomics and Clinical Diagnostics