Bayesian Meta-analysis Models for Heterogeneous Genomics Data

by

Lingling Zheng

Department of Computational Biology & Bioinformatics
Duke University

Date: _____

Approved:

_____
Sayan Mukherjee, Supervisor

_____
Joseph Lucas

_____
Elizabeth Hauser

_____
Jen-Tsan Chi

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Computational Biology & Bioinformatics
in the Graduate School of Duke University
2013

Abstract

Bayesian Meta-analysis Models for Heterogeneous Genomics Data

by

Lingling Zheng

Department of Computational Biology & Bioinformatics
Duke University

Date: _____
Approved:

_____
Sayan Mukherjee, Supervisor

_____
Joseph Lucas

_____
Elizabeth Hauser

_____
Jen-Tsan Chi

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Computational Biology & Bioinformatics
in the Graduate School of Duke University
2013

Abstract

The accumulation of high-throughput data from vast sources has drawn a lot attentions to develop methods for extracting meaningful information out of the massive data. More interesting questions arise from how to combine the disparate information, which goes beyond modeling sparsity and dimension reduction. This dissertation focuses on the innovations in the area of heterogeneous data integration.

Chapter 1 contextualizes this dissertation by introducing different aspects of meta-analysis and model frameworks for high-dimensional genomic data.

Chapter 2 introduces a novel technique, joint Bayesian sparse factor analysis model, to vertically integrate multi-dimensional genomic data from different platforms.

Chapter 3 extends the above model to a nonparametric Bayes formula. It directly infers number of factors from a model-based approach.

On the other hand, chapter 4 deals with horizontal integration of diverse gene expression data; the model infers pathway activities across various experimental conditions.

All the methods mentioned above are demonstrated in both simulation studies and real data applications in chapters 2-4.

Finally, chapter 5 summarizes the dissertation and discusses future directions.

Contents

List of Tables

List of Figures

Acknowledgements

First and foremost I would like to express my deepest gratitudes to my advisor Dr. Joseph Lucas. He has led me into the Bayesian and modeling world, which I find a lot interests in. I appreciate all the opportunities he has created for me to make my Ph.D. productive and stimulating. I am also thankful for the excellent example he has set, which constantly motivates me to become a successful biostatistician.

I am deeply grateful to my dissertation chair, Dr. Sayan Mukherjee. He has always been helpful and supportive for his students. His insightful comments and continual patience has guided me through the completion of this dissertation.

I also feel amazingly fortunate to have an exceptional doctoral committee. Dr. Jen-Tsan Chi has been a decent mentor. He always motivated me to develop a successful career, and taught me valuable life lessons. The research conversations with him were thought-provoking, and his practical advice has greatly helped me understand the problem from the biology perspective. I wish to thank Dr. Elizabeth Hauser, with whom I very much enjoyed to work with in my first rotation. I especially appreciate the time she devoted to write the recommendation letter, which helped me get started in my career.

I would like to thank my collaborators who helped me finish my dissertation. My collaboration with Minhua Chen has been a very good start for me to get into the statistical/machine learning field. Thanks to his introduction of Priyadip Ray, I found breakthroughs in my research, and our collaboration has led to some nice

results in statistical methodology. Ricardo Henao provided much extraordinarily perceptive advice for my work. His passion in methodology research and fantastic understanding of statistical theories greatly inspired me. I also appreciate the help from Xiao Yan, who has taught me interesting things about hematopoietic system. Working with him was a happy experience in my graduate study.

I gratefully acknowledge the funding sources that made my Ph.D. work possible, which include but are not limited to the Defense Advanced Research Projects Agency (DARPA-N66001-09-C-2082), the grant "R01 DK089705", and Biomedical Advanced Research and Development Authority (BARDA). I am also thankful for NSF, which kindly provided travel fellowship for my very first international conference talk. And for Broad Institute, I appreciate your generosity for sponsoring my trip to the sequence/omics data workshop.

My time at Duke was made enjoyable because of many friends that became a part of my life. To Gurkan, Yingbo, Fangpo, Jun, Sunil, Iris and many others: Thank you!

Lastly, I would like to thank my parents for their endless love and support from the other side of the globe. And for my loving, encouraging and patient boyfriend, Enliang Zheng, his faithful supports even during the tough time of my Ph.D. is more than appreciated.

1

Introduction

Most cellular activities can be organized as interacting regulatory modules: sets of genes with a common function co-regulated in response to internal and external stimuli (Segal et al., 2003). Examples include metabolic pathways and cell cycle gene modules (Moreno-Asso et al., 2013). Genes in the same module are coordinately activated or repressed under the same regulatory mechanism. This can be captured by genome-wide gene expression profiling. Microarray is a powerful tool to measure the expression level of tens of thousands of genes simultaneously, thus helps to capture gene regulatory patterns (a.k.a., co-expression patterns) for numerous individuals of various disease states. There have already been several publications on identification of gene expression signatures related to cancer specific phenotypes (Gui and Li, 2005; Garber et al., 2001) and cellular physiology of *S. cerevisiae* (Airoldi et al., 2009).

However, several challenges are not addressed using gene expression data alone:

i) How do genetic mutations (e.g., copy number variations and single nucleotide sequence changes) or epigenetic alterations (e.g., methylation and histone modifications) disrupt the regulatory network and manifest in disease, such as cancer? Fundamentally, this question is to inquire into the basic mechanism of cancer that

persistently drives tumor proliferation and metastasis, even though each individual tumor is highly heterogeneous.

ii) How to decipher the biological function responsible for the differential expression pattern of a regulatory module associated with a particular experimental trait? Meanwhile, how to obtain a systematic understanding of the module function among different cellular parts, genetic backgrounds, and environments (e.g., nutrients, radiation and cell micro-environments)?

A recent flood of genome-wide data generated by high-throughput technologies provides unprecedented opportunity to tackle the above problems (Pe'er and Hacohen, 2011). Therefore, there is a critical need for powerful statistical approaches that build models from diverse data types in a 'data integration' fashion. This chapter will provide more details regarding the above three aspects. It will mainly concentrate on using gene expression as an intermediary to build a cascade of events from DNA, to seek the root cause of cancer, and to connect genomic information from diverse biological conditions through modulated gene expression to phenotype.

The remainder of this chapter is organized as follows: Section 1.1 will elaborate on biological motivations for the three problems; In section 1.2, solutions to these questions will be provided with an overview of statistical models developed in this dissertation.

1.1 Biological motivation

1.1.1 Discovering driver mutations in cancer

Cancer is caused through a multistep process, in which a succession of genetic and epigenetic changes collectively influence the expression of multiple genes, leading to the alteration of key pathways and biological processes underlying malignant behaviors. Specifically, copy number variations (CNVs) change the dosage of key tumor-inducing and tumor-suppressing genes, thereby affecting mRNA transcription and neoplastic

cell proliferation. On the other hand, the mechanism of epigenetic alterations is more complicated. For example, DNA methylation patterns are globally disrupted in tumor cells. The cancer methylome is characterized by both global hypomethylation and region-specific hypermethylation at CpG islands. Hypomethylation may contribute to carcinogenesis via transcriptional activation of tumor-promoting genes (Wu et al., 2005), while hypermethylation at CpG islands is associated with silencing genes involved in growth regulation, cell cycle control, apoptosis and tumor suppression. It is even noted that hypermethylation is more likely prominent in transcriptional silencing and down-regulating pathways involved in drug resistance (chemoresistance) (Li et al., 2009). Therefore, genetic mutations and chromosomal aberrations are the central characteristics of tumor cells (Pe'er and Hacohen, 2011).

In recent years, the emergence of large-scale copy number assays and methylation platforms enables the possibility of tracing phenotypic differences back to their genetic/epigenetic source. However, only a few genetic mutations or epigenetic alterations provide a persistent fitness advantage across multiple tumors. Such a rare event could leave a 'genomic footprint' in the form of a gene expression signature (Akavia et al., 2010). Therefore, it becomes increasingly important to distinguish genetic/epigenetic changes that alter mRNA transcription, and thus promote cancer progression (driver mutation) from those with no selective advantage (passenger mutation) (Pe'er and Hacohen, 2011; Akavia et al., 2010).

We propose to integrate gene expression data of cancer patients with their multi-perspective genomic information by the development of two innovative statistical models (briefly overviewed in section 1.2.1) in *chapters 2 and 3*. Such a global analysis of genomic effects on the co-expression modules can have transformative value, which allows biologists to conduct refined experiments, decipher the underlying mechanism, and identify combinatory therapeutics for controlling cancer.

1.1.2 Pathway analysis via a compendium of expression profiles

The co-expression modules derived from gene expression profiles provide a good way to understand molecular mechanisms underlying cellular changes at a functional level. With the establishment of large knowledge bases, such as Gene Ontology, KEGG and BioCarta, it becomes straightforward to understand the cause of changes in gene expression. The general workflow is first to identify dysregulated pathways/gene sets that differentiate test from normal groups, such as disease subjects vs. healthy individuals, then to generate hypothesis that links the biological characteristic of the pathway/gene set to the experimental perturbation.

Over the past few decades, large compendia of gene expression data have been freely deposited in the public repositories, representing transcriptional responses to a vast number of perturbations including mutants, treatment with pharmaceutical compounds, etc. By comparing the expression profile caused by an uncharacterized perturbation to a large and diverse set of reference profiles, one can hypothesize the function of an uncharacterized module by borrowing information from known modules corresponding to other disturbances. Such an approach has great advantages over a conventional single assay, as it integrates measurements from each single cellular parameter to form a systematic landscape.

In *chapter 4*, we develop a statistical method (overviewed in section 1.2.1.) to integrate multiple gene expression experiments to infer pathway activities across diverse biological conditions. This method is useful to generate novel hypotheses about the causes and consequences of specific expression patterns associated with a particular phenotypic trait (Gower et al., 2011).

1.2 Statistical innovation

In order to fully decipher the biological knowledge contained in the tremendous amount of 'Omics' data, an essential approach is to combine multiple studies for *meta-analysis*. I adopt this term for the purpose of research synergy: identifying heterogeneity and homogeneity across multi-platforms and diverse gene expression profiles, developing biomarkers from training data sets and making predictions against the testing groups. For the first problem in 1.1.1, I define the combination of multiple sources of 'Omics' information from a given cohort of patients as *vertical integration*. Mathematically, the problem can also be viewed as stacking several matrices vertically, each representing a data set with same number of samples in columns, but different quantities of genes in rows. In the second scenario (1.1.2), I define combining gene expression profiles from diverse range of experiments as *horizontal integration*. Again, mathematically, this is equivalent to placing individual matrices parallel to each other to form a bigger dataset, where each one shares the same number of rows but has a different number of columns.

1.2.1 Integrative modeling with joint factor analysis approach

The initial chapters of this dissertation concentrate on integrative modeling. Genome-wide data is high dimensional and contains correlated information such as co-expression modules. One of the fundamental goals of genomic data analysis is to determine if the data arose from a mixture of several distinct populations. However, in high dimension, visually inspecting the difference between subpopulations is nearly impossible (Smith, 2005). Therefore, techniques, including factor analysis (FA), are developed to reduce the high dimensional data to much lower dimensions. The FA model has

the following expression:

$$\mathbf{X} = \mathbf{\Lambda F} + \boldsymbol{\epsilon} = \sum_{k=1}^{K} \mathbf{\Lambda}_k \mathbf{F}_k + \boldsymbol{\epsilon} \tag{1.1}$$

where $\mathbf{X}$ represents the genomic data with dimension $p \times 1$ and n samples, $n \ll p$; $\mathbf{\Lambda} = [\mathbf{\Lambda}_1, \mathbf{\Lambda}_2, \ldots, \mathbf{\Lambda}_K] \in \mathbb{R}^{p \times K}$ is the factor loading matrix with $K \ll p$, and $\mathbf{F} = [\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_K]^T \in \mathbb{R}^{K \times n}$ is the factor score matrix. $\boldsymbol{\epsilon}$ is the noise term and has a diagonal precision matrix, $diag(\boldsymbol{\phi})$. In a conventional FA model, in order to identify the correlation among genes, sparse constraints are further imposed by zeroing out many variables on the loading matrix, thus greatly reducing the number of parameters. This is necessary to perform robust inference on high-dimensional data (Karoui, 2008). Secondly, a standard normal prior is imposed on $\mathbf{F}$ to alleviate issues with identifiability of $\mathbf{F}$ and $\mathbf{\Lambda}$ due to scaling. Therefore, the marginal density function of $\mathbf{X}$ becomes $\mathcal{N}(\mathbf{X}; \mathbf{0}, \mathbf{\Lambda\Lambda}^T + diag(\boldsymbol{\phi}))$. It is clear that by modeling the sparseness and low-rank structure in the covariance matrix of data through $\mathbf{\Lambda}$, we can identify true signals in high dimensional genomic data.

Using this framework, each high-dimensional genomic dataset, e.g., gene expression, CNV or methylation, can be analyzed with an individual FA model. We take a step further by assuming some factors are shared by two or more datasets that explains the statistical correlations among them. *Chapter 2* introduces this approach by developing a two-FA model using two datasets as an example, aneuploidy and gene expression. We link individual factors from each dataset by sampling gene expression factors under a multivariate normal distribution centered on the CNV factors. This way allows us to impose the hypothesis that gene expression is directly affected by CNVs. It addresses the question of dimensionality discrepancy between two data modalities, and prevents the difference in data size from overwhelming the information available on associations between them. Meanwhile, there are fractions in

gene expression that cannot be explained by copy number variations, or the changes in CNVs do not necessarily result in expression dysregulation. In this sense, the model is flexible enough to incorporate unique factors specific to each data modality by imposing a diffuse standard normal prior on these factors. Additionally, each FA model has a unique factor loading matrix to account for the mapping between factors and observations. Sparseness is further imposed on factor loadings with a 'spike and slab' prior (Ishwaran and Rao, 2005). This chapter provides a general framework for *vertically integrating multi-platform genomic data*. By treating gene expression as a downstream event, it could decipher regulation mechanisms from different sources, such as epigenetic modifications (e.g., methylations) or post-transcriptional regulators (e.g., microRNAs).

The above model requires pre-specification of factor numbers by applying affinity propagation to one of the data sets. However, it is very challenging to check model assumptions in high dimensions. Besides, the common and unique factors need to be predefined as well. Without enforcing any constraints, factor labels suffer from arbitrary permutation. It becomes a long-standing label-switching problem (first proposed by Diebolt and Robert (1994)) that there is no unique answer for labeling the shared/unique factors.

To address this, non- or semi-parametric models are needed. *Chapter 3* extends the joint FA model to a nonparametric Bayesian formula. Based on the same assumption of factorizing the latent space into shared and data-specific components, the new model employs a beta-Bernoulli process (Griffiths and Ghahramani, 2005; Thibaux and Jordan, 2007; Paisley and Carin, 2009) to infer the dimensions of these latent spaces. Therefore, the factor matrix becomes the Hadamard product of both factor scores and a sparse binary matrix drawn from the beta-Bernoulli process. The latter also enables the possibility of discovering factors that are relevant to only a subset of the subjects. This is very useful in cancer research, since cancer is highly

heterogeneous even among patients with the same tumor subtype. The same regulatory mechanism cannot explain all variability among individuals. By doing so, we could assign zeros to a subset of row vectors in the factors, thus filtering out patients whose expression alterations are weakly related to their genetic/epigenetic mutations. A *vertical integration* of three different datasets, including gene expression, CNV and methylation from TCGA ovarian cancer patients, are explored with this model, followed by discussion of novel hypotheses.

Another complementary direction of meta-analysis is to *horizontally integrate diverse genomic data*, such as gene expression microarrays. By borrowing information from each data set, representing a different experimental perturbation, we are interested in identifying consistent co-expression patterns across diverse biological conditions. *Chapter 4* provides a novel statistical solution to this question based on the FA framework (1.1). We improve the model interpretability by correlating each gene in the microarray to a known pathway. Therefore, the sparseness becomes well defined by introducing a strongly informative prior on the loading matrix. Inspired by the application of a beta-Bernoulli process to uncover patient heterogeneity in chapter 3, we adopt a similar strategy to model the pathway-dataset membership using a three-layer beta-Bernoulli hierarchical distribution. In this way, the method can identify pathways specific to a particular experimental trait or consistently differentially expressed under a variety of interventions. Application of this model to a radiation study reveals novel insights into the molecular basis of time- and dose-dependent response to ionizing radiation in mice peripheral blood. It provides a broadly applicable approach to generate biological hypotheses in a gene expression data-driven and pathway-centric manner.

1.3 Summary of contributions

Through this dissertation, biological and methodological aspects of the high-dimensional meta-analysis problems are discussed along with the development of several innovative statistical models. Core contributions of each chapter are highlighted as follows: chapter 2 introduces a novel statistical model, joint Bayesian factor analysis model, which vertically integrates multi-platform genomic data to uncover key mutations in cancer. Chapter 3 discusses an extension of this model to a nonparametric Bayesian formula. Application to analyzing multi-dimensional genomic data for ovarian cancer is also included. Using a similar joint factor analysis framework, chapter 4, however, talks about a horizontal integration approach for pathway analysis using diverse gene expression data. The successful application to a mice radiation study demonstrates the utility of this approach. Finally in chapter 5, I conclude this work with the discussion of some extended questions and follow-up goals.

2

Joint Bayesian Factor Analysis–A Vertical Integration Approach to Model Multi-platform Genomics Data

2.1 Introduction

Human cancers are heterogeneous due to combined effects of genetic instability and selection, where the accumulation of the most advantageous set of genetic aberrations results in the expansion of cancer cells (Pinkel and Albertson, 2005). There are many different types of instability that occur during tumor development, such as point mutation, alteration of microsatellite sequences, chromosome rearrangements, DNA dosage aberrations and epigenetic changes such as methylation. These abnormalities acting alone or in combination alter the expression levels of mRNA molecules. However, the genetic history of tumor progression is difficult to decipher. Because it is only a sufficiently protumorigenic aberration or obligate products of a crucial alteration that results in tumor development (Pinkel and Albertson, 2005).

Genomic DNA copy number variations (CNVs), kilobase- or megabase-sized duplications and deletions, are frequent in solid tumors. It has been shown that CNVs

are useful diagnostic markers for cancer prediction and prognosis (Kiechle et al., 2001; Lockwood et al., 2005). Therefore, studying the genomic causes and their association with phenotypic alterations is emergent in cancer biology. The underlying mechanism of CNV related genomic instability amongst tumors includes defects in maintenance/manipulation of genome stability, telomere erosion, chromosome breakage, cell cycle defects and failures in DNA repairs (Albertson, 2003). Consequential copy number aberrations of the above mentioned malfunctions will further change the dosage of key tumor-inducing and tumor-suppressing genes, which thereby affect DNA replication, DNA damage/repair, mitosis, centrosome, telomere, mRNA transcription and proliferation of neoplastic cells. In addition, microenvironmental stresses play a role in exerting strong selective pressure on cancer cells with amplification/deletion of particular regions of the chromosome (Lucas et al., 2010). Recently, high-throughput technologies have mapped genome-wide DNA copy number variations at high resolution, and discovered multiple new genes in cancer. However, there is enormous diversity in each individual's tumor, which harbors only a few driver mutations (copy number alterations playing a critical role in tumor development). In addition, CNV regions are particularly large containing many genes, most of which are indistinguishable from the passenger mutations (copy number segments affecting widespread chromosomal instability in many advanced human tumors) (Akavia et al., 2010). Thus analysis based on CNV data alone will leave the functional importance and physiological impact of genetic alteration ineluctable on the tumor. Gene expression has been readily available for profiling many tumors, therefore, how to incorporate it with CNV data to identify key drivers becomes an important problem to uncover cancer mechanism.

This chapter is laid out as follows: Section 2.2 covers a variety of CNV data topics, starting with a range of different CNV measurement techniques, which includes a brief discussion of the data format. Practical examples are used to show collecting,

generating and assessing data, plus several ways to manipulate data for normalization. In the end, different computational approaches are introduced for analyzing CNV data. Section 2.3 focuses on a novel algorithm for integrating CNV with mRNA expression data, which can be potentially extended to incorporate multiple genomic data. Basic concepts of Bayesian factor analysis are briefly mentioned. Case studies then provide detailed description for this particular approach. Section 2.4 provides a brief wrap-up of the main ideas in the chapter. It illustrates the advantage of our statistical models on studying cancer genomics, and discusses the significance of the approach for clinical application.

2.2 Copy number analysis

2.2.1 Copy number analyses techniques

Comparative genome hybridization (CGH) is a recently developed technology and profiles genome-wide DNA copy number variations at high resolution. It has been popular for molecular classification of different tumor types, diagnosis of tumor progression, and identification of potential therapeutic targets (Jönsson et al., 2010; McKay et al., 2011). The use of CGH array offers many advantages over traditional karyotype or FISH (fluorescence *in situ* hybridization). It can detect microduplications/deletions throughout the genome in a single experiment. A review of different CGH array techniques is provided as follows:

•*BAC Array*

The CGH array using BAC (bacterial artificial chromosome) clones has been widely used. The spotted genomic sequences are inserted BACs: two DNA samples from either subject tissue (target sample) or control tissue (reference sample) are labeled with different fluorescent dyes–for example, with the test labeled in green and reference in red. The mixture is hybridized to a CGH array slide containing hundreds or thousands of defined DNA probes. The probes targeting regions of the

chromosome that are amplified turn predominantly green. Conversely, if a region is deleted in the test sample, the corresponding probes become red. However, given the resolution limitation on the order of 1Mb and array size of 2.4K to ∼30K unique elements, the BAC array data is relatively low density.

•*cDNA/oligonucleotide Array*

cDNA and oligonucleotide arrays are designed to detect complementary DNA 'targets' derived from experiments or clinics. It allows greater flexibility to produce customized arrays, and reduces the cost for each study. But the shorter probes spotted on these new arrays are less robust than large segmented BACs, because they contain a large number of genes that are not of interest to the researchers. However, they do provide higher resolution in the order of 50-100kb, where oligonucleotide array is a particular case.

•*Tiling Array*

Tiling arrays are available now for finer resolution of specific CNV regions. These arrays are designed to cover the entire genome or contiguous regions within the genome. The number of elements on the array ranges from 10K to over 6M. This relatively high resolution technique allows the detection of micro-amplifications and deletions.

•*SNP Array*

SNP (single nucleotide polymorphism) arrays are a high-density oligonucleotide-based array that can be used to identify both loss of heterozygosity (LOH) and CNVs. LOH is the loss of one allele of a gene, which can lead to functional loss of normal tumor suppressor genes, particularly if the other copy of the gene is inactive. LOH is quite common in malignancies. Therefore, utilization of SNP arrays to detect LOH provides great potential for cancer diagnosis.

•*Array CGH*

Array comparative genomic hybridization (array CGH, or aCGH) is a high-

resolution technique for genome-wide DNA copy number variation profiling. This method allows identification of recurrent chromosome changes with microamplifications and deletions, and detects copy number variations on the order of 5-10kb DNA sequences. In the rest of this chapter, we will use the CNV data generated from the Agilent Human Genome CGH microarray 244A.

2.2.2 Array CGH data

The CNV data is obtained from The Cancer Genome Atlas (TCGA) project. TCGA is a joint effort of the National Cancer Institute and the National Human Genome Research Institute (NHGRI) to understand genomic alterations in human cancer. It aims to study the molecular mechanisms of cancer in order to improve diagnosis, treatment and prevention. Since the importance of DNA copy number variations has been demonstrated in many tumors, TCGA performs high-resolution CNV profiling in a large-scale study, using diverse tumor tissues and across different institutes. In this section, we will show an example from the TCGA project.

Sample collection

Biospecimens were collected from newly diagnosed patients with ovarian serous cystadenocarcinoma (histologically consistent with ovarian serous adenocarcinoma confirmed by pathologists), who had not received any prior treatment, including chemotherapy or radiotherapy. Technical details about sample collection and quality control are described in (TCGA, 2011). Raw copy number data was generated at two centers, Brigham and Women's Hospital of Harvard Medical School and Dana Farber Cancer Institute, using the Agilent Human Genome Comparative Genome Hybridization 244A platform.

Data process

After the array CGH is constructed and tumor DNA samples hybridized to the platform, several steps need to be completed for detecting regions of copy number gains or losses: image scanning, image analysis (including gridding, spot recognition, segmentation and quantification, and low-intensified feature removal or mark), background noise subtraction, spot intensity ratio determination, log-transformation of ratios, signal normalization and quality control on the measured values. For Agilent 244K array, there are specific details on the data generation (TCGA, 2008). First of all, the raw signal is obtained by scanning images using Agilent Feature Extraction Software (v9.5.11), followed by image analysis procedures mentioned above.

Background correction: The background corrected intensity ratios for both channels are calculated by subtraction of median background signal values (median pixel intensities in the predefined background area surrounding the spot) of each channel from the median signal values (median pixel intensities computed over the spot area) of each probe in the corresponding channel. Since there are multiple copies of probes on an array, the final background corrected values are computed by taking the median across the duplicated probes. The $\log_2$ ratios of the above results are then estimated based on the background corrected values of sample channel over that of the reference channel.

Normalization of logarithmic ratio: The normalization procedure involves the application of LOWESS (locally weighted regression and scatterplot smoothing) algorithm on $\log_2$ ratio data. This method assumes that the majority of probe $\log_2$ ratios do not change, and are independent of background corrected intensities of the probes. To develop the LOWESS model, a 21-probe window is applied for smoothing process after sorting the chromosome positions. It corrects the $\log_2$ ratio data so that the corresponding central tendency after normalization lies along zeros, assuming an

equal number of up- and down- regulated features in any given intensity range. In addition, the artifact of the difference in the probe GC content on $\log_2$ ratios is considered for correction, in which case, the probe GC%, regional GC % (GC% of 20KB of genome sequence containing the probe sequence) and $\log_2$ ratio are used in the LOWESS model.

Quality control: There are several criterions taken into account for quality assurance at various stages. 1) Probes that are flagged (marking spots of poor quality and low intensity) or saturated by the Agilent feature extraction software are eliminated; 2) Screening of the array image is conducted to exclude probes whose median signal values are lower than that of the background intensity; 3) Arrays with over 5% probes flagged out or being faint are considered as low quality; 4) The square root of the mean sum squares of variance in $\log_2$ ratio data between consecutive probes are calculated for quality assessment. Arrays with the value over 0.3 are considered as low quality.

The final result after these processes forms a data set containing 227614 probes with normalized $\log_2$ ratio values for every sample. The logarithmic ratios are computed as $\log_2(\mathrm{x}) - \log_2(2)$, where x is the copy number inferred by the chip. Thus, ratios should be 0 for double loss, $\frac{1}{2}$ for a single loss, 1 for the normal situation, $\frac{3}{2}$ for a single gain, and $\frac{n}{2}$ for n copies. TCGA provides an Array Design Format file with annotation data, including information on chromosomal location and gene symbol for each probe.

Algorithms for CNVs detection

The main biomedical question for studying CNVs and downstream research is to accurately identify genomic/chromosomal regions that show significant amplification or deletion in DNA copy number. Satisfactorily solving this problem requires a method that reflects the underlying biology and key features of the technological

platform. The array CGH data has particular characteristics: The status of DNA copy number remains stable in the contiguous loci, and the copy number of a probe is a good predictor for that of the neighboring ones, whereas for probes located far apart, it provides less information to predict the likely state of its neighboring probes (Rueda and Uriarte, 2007). However, widely used array CGH platforms, such as cDNA/oligonucleotide arrays, do not have equally spaced probes, making them less informative based on consecutive probes. Furthermore, the identification of disease causal genes sometimes requires examining the amplitude of CNVs, especially when high-resolution technologies are available, it can be valuable to distinguish between moderate copy number gains and large copy number amplification.

A number of well-known methods have been developed to carry out automatic identification of copy number gains/loss, and correlate that with diseases. These approaches are designed to estimate the significance level and location of CNVs. Models differ in distribution assumption and incorporation of penalty terms for parameter estimation. Subsequently, smoothing algorithms were derived for denoising and estimating the spatial dependence, such as wavelets (Hsu et al., 2005) and lowess methods (Beheshti et al., 2003; Cleveland, 1979). Later on, a binary segmentation approach, called circular binary segmentation (CBS) (Olshen et al., 2004), was proposed that allows segments in the aCGH data in each chromosome, and computes the within-segment means. CBS recursively estimates the maximum likelihood ratio statistics to detect the narrowed segment aberrations. A more complicated likelihood function was used with weights chosen in a completely data adaptive fashion (Adaptive weights smoothing procedure, AWS) (Hup et al., 2004). A different kind of modeling approach involves the hidden Markov model (HMM) (Fridlyand et al., 2004), which assigns hidden states with certain transition probabilities to underlying copy numbers. Thus, it adequately takes advantage of the physical dependence information of the nearby fragments. However, questions arise on how to appropriately

select the number of hidden states. The sticky hidden Markov model with a Dirichlet distribution (sticky DD-HMM) (Du et al., 2010) was then developed to infer the number of states from data, while also imposing state persistence. Alternatively, the reversible jump aCGH (RJaCGH) (Rueda and Uriarte, 2007) was introduced to fit the model with varying number of hidden states, and allow for transdimensional moves between these models. It also incorporates interprobe distance.

2.3 Joint analysis of copy number variation and gene expression

2.3.1 Overview

With the increasing availability of concurrently generating multiple different types of high throughput data on single samples, there is a lot of interest to jointly analyze this information and refine the generation of relevant biological hypotheses. This will lead to a greater, more integrated understanding of cellular mechanism, and will allow the identification of genomic regulators as well as suggest potentially synergistic drug targets for those regulators, which will lead to potential combination therapies for the treatment of human cancer. A number of approaches have demonstrated an ability to select specific genes from joint analysis and test specific hypotheses regarding the regulation of cellular responses, which is a tremendous advantage over the pathway analyses that can be obtained from gene expression or CNVs alone.

Recently, there are publications that highlight the impact of combining other types of DNA modification and gene expression. Parsons et al. (2008) have identified a number of potential driver mutations in Glioblastoma through an analysis of mutation, copy number variation and gene expression. Their approach is designed around the use of currently available methods for the analysis of individual data types to create a compressed set of features which are then used independently in predictive models. They utilize tree models, however the compressed features are independent variables that can, in principle, be used in any type of predictive model.

The approach does make use of correlation within each type of data, but not across different data types.

A similar approach to the integration of disparate types of data is outlined in (Lanckriet et al., 2004), but in this case features are compressed through the use of kernel functions. These must be predefined for each data type, but once that is done all of the different data types are mapped to the same vector space allowing joint analysis. The approach is particularly suited to the use of support vector machines, rather than tree models, for the generation of models from all of the different data types. The approach is remarkably general in that almost any type of data may be incorporated, and in the paper they include compelling examples of the integration of expression and protein sequence data. It, however, does suffer from the same flaws as Parsons et al. (2008) in that there is no provision for dealing with correlation across data types.

Another approach to integrative analysis is through the use of data from different assays to filter lists of genes sequentially. Garraway et al. (2005) describe such an approach, in the context of the identification of MITF as a genomic determinant in malignant melanoma. The algorithm first identifies genomic regions that show copy number variation in the condition of interest, and then searches for genes that are significantly over or under expressed in samples that have duplications or deletions in that region. This is a very powerful approach in cases where there are few genes that pass the filtering criteria and where the relationship between gene expression and CNV is direct. Through our own experimentation, we find that there are often many genes that pass both filtering criteria. Additionally, the approach is dependent on the order in which the data types are used to perform the filtering. This is because the filtering criterion on the second data set is determined by the behavior observed on the first.

The version of integrative genomic analysis that is most similar to our own pro-

posal is CONEXIC, detailed in Akavia et al. (2010). CONEXIC is based on gene modules, which was initially developed for the analysis of gene expression data in isolation. Gene modules consist of groups of genes that are coexpressed, and these are embedded as leaves in a binary tree structure where the nodes are populated by putative gene expression regulators. In its original incarnation, the approach was intended to identify important regulators of groups of genes in the context of experimental interventions. As such, expression is assumed to be constant within any particular experimental group. Also, the original approach depends on a list of putative regulators, which can be tricky to generate. With CONEXIC, the identification of lists of potential regulators is generated from regions of the genome that demonstrate consistent copy number variation, and the gene module algorithm is largely retained. Fundamental to a binary tree model is the assumption that the expression pattern of a leaf, conditional on the expression pattern of its parent node, is independent of all other elements in the tree. This is a shortcoming of the CONEXIC approach. It is quite reasonable to expect that there are many ways that a cell is able to control the expression of a particular gene, including CNV, methylation, inactivation of promoters, and RNA interference. Multiple different regulators may have combined effect to ultimately regulate gene expression. Because each node of the tree contains only one putative regulator, the model assumes that only one regulator is responsible for the observed expression pattern of a module.

2.3.2 Bayesian factor analysis

Bayesian factor analysis is a dimension reduction method to decompose variability among observations into a lower number of unobserved, uncorrelated factors. It has been widely applied in microarray analysis (Carvalho et al., 2008b; Lucas et al., 2009), where the data usually comes with a much higher dimension than the number of observed samples. Therefore, it is desirable to select important genes that should

bear some biological meanings. Recent developments in Bayesian multivariate modeling has enabled the utility of sparsity induced structure in genomic studies (Lucas et al., 2006). Such a sparse factor model implies that only those genes with non-zero loadings on those factors are relevant, and higher values indicate more significant gene-factor relationship.

2.3.3 Sparse regression model of Bayesian factor analysis

Our statistical framework utilizes high-dimensional sparse factor model, and is extended to incorporate gene expression, CNVs and other high-throughput genomic data. The underlying hypothesis is that the gene signatures of expression variation can be represented by the estimated factors. Furthermore, given the potential contribution of chromosomal aneuploidy and CNVs to the altered mRNA expression of relevant genes during oncogenesis, we could use the factor model to test for the association between gene expression signatures and CNVs. The model assumes that the input data are from the same patient cohort. Suppose the data structure is given as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ with dimension $n \times p_x$, where n denotes the sample size, p_x the number of genes, and $\mathbf{x}_i$ the fluorescence level from probes of gene expression measurements. The CNV data is represented by $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$ with similar strucutre. Therefore, the linear regression model for sample i can be expressed as

$$\mathbf{x}_i = \mathbf{B}_h \mathbf{h}_i + \mathbf{B}\mathbf{F}_i + \boldsymbol{\epsilon}_i \tag{2.1}$$

$$\mathbf{y}_i = \mathbf{A}_h \mathbf{h}_i + \mathbf{A}\mathbf{G}_i + \boldsymbol{\zeta}_i \tag{2.2}$$

with the following components:

- $\mathbf{B}$ is the $p_x \times k$ factor loadings matrix for sample $\mathbf{x}_i$, with elements $\beta_{g,j}$ for $g = 1, \ldots, p_x$ and $j = 1, \ldots, k$.
- $\mathbf{F}_i = [\mathbf{f}_i^C; \mathbf{f}_i^{(r)}]^{\mathrm{T}}$. $\mathbf{f}_i$ is a k-dimension vector of factor scores, where $\mathbf{f}_i^{(r)}$, the r-th factor for sample i, are specific to data $\mathbf{x}_i$, and $\mathbf{f}_i^C$ consists of the factors ***common*** between both data.

• $\mathbf{B}_h$ is the $p_x \times r$ regression matrix for dataset $\mathbf{x}_i$, with elements $b_{g,j}$ for $g = 1, \ldots, p_x$ and $j = 1, \ldots, r$.

• $\mathbf{h}_i = [h_{1,i}, \ldots, h_{q,i}]^{\mathrm{T}}$ is the q design factors of sample i.

• $\boldsymbol{\epsilon}_i = [\epsilon_{1,i}, \ldots, \epsilon_{p_x,i}]^{\mathrm{T}}$ is the idiosyncratic noise vector with dimension p_x.

The priors for each parameters are defined as follows:

$$\beta_{g,j} \sim (1-\rho_j)\delta_0(\beta_{g,j}) + \rho_j \mathcal{N}(\beta_{g,j}; 0, \tau_j) \tag{2.3}$$

$$\rho_j \sim Beta(\rho_j; s_0, l_0); \tau_j \sim Gamma({\tau_j}^{-1}; \frac{a_\tau}{2}, \frac{b_\tau}{2}) \tag{2.4}$$

$$b_{g,j} \sim (1-\pi_j)\delta_0(b_{g,j}) + \pi_j \mathcal{N}(b_{g,j}; \mu_{0,j}, {\sigma_0}^2) \tag{2.5}$$

$$\pi_j \sim Beta(\pi_j; t_0, v_0) \tag{2.6}$$

$$\mathbf{f}_i^{(r)} \sim \mathcal{N}(\mathbf{f}_i^{(r)}; \mathbf{0}, \mathbf{I}) \qquad \mathbf{f}_i^{C} \sim \mathcal{N}(\mathbf{f}_i^{C}; \mathbf{g}_i^{C}, \Sigma) \tag{2.7}$$

$$\boldsymbol{\epsilon}_i^{(r)} \sim \mathcal{N}(\boldsymbol{\epsilon}_i^{(r)}; \mathbf{0}, \boldsymbol{\Phi}); \boldsymbol{\Phi} = \mathrm{diag}(\phi_1, \ldots, \phi_{p_x}); \phi_g \sim Gamma(\phi_g; \frac{a_{\phi_x}}{2}, \frac{b_{\phi_x}}{2}) \tag{2.8}$$

The parameters and prior structures are similar for copy number data $\mathbf{y}_i$.

Prior Choices

• $\beta_{g,j}$: The regression coefficient. Here we consider the long-standing problem of variable selection in a multivariate linear regression model. That is, in gene expression analysis the number of gene features is huge (usually larger than 20,000) compared with the number of samples available. A direct way is to use regression model on the high-dimensional genomic data and impose sparseness on the coefficients. In this way, most of the coefficients will be shrunk towards zero. Bayesian spike and slab approaches (Mitchell and Beauchamp, 1988; George and McCulloch, 1993; Bühlmann and Hothorn, 2005) have been proposed to address the variable selection problem. As indicated in 2.3, it sets up a two-component mixture distribution with the spike part centered at zero and the slab part distributed diffusely without informed prior knowledge.

• ρ_j: This parameter controls the prior probability of a coefficient being non-zero. We assume coefficients that are promising have posterior latent variables $\hat{\rho}_j = 1$ (the slab). The opposite occurs when $\hat{\rho}_j = 0$ with a delta function $\delta_0(\cdot)$ indicating the point-mass at zero (the spike). Here we use beta priors, defining the probability ρ_j distributed on the interval (0,1). The hyperparameters s_0 and l_0 determine the domain of the beta distribution. Small values of ρ_j reflect high prior skepticism about the coefficients, while large ρ_j means the knowledge of more theoretical importance of the variables and more skeptical about the sampling of the data.

• τ_j: the variance for the slab part of the mixture prior for $\beta_{g,j}$. This gamma distribution is the conjugate prior for the precision of the normal distribution $\mathcal{N}(\beta_{g,j}; 0, \tau_j)$. In addition, it allows the Markov chain to identify and adjust the appropriate sample space for updating coefficients. Different combinations of ρ_j and τ_j prior choices are usually required to obtain desirable mixing and shrinkage in $\beta_{g,j}$.

• $\mathbf{f}_i$: Unknown latent factors for sample i. For factors unique for each data, we use a diffuse, conjugate prior distribution such that $f_{j,i} \sim \mathcal{N}(0, 1)$, in order to alleviate issues with identifiability of $\mathbf{f}_i$ and $\boldsymbol{\beta}$ due to scaling. On the other hand, since high-throughput data can vary in size by orders of magnitude, e.g. CGH data is approximately ten times larger than gene expression. Thus one data set may dominate the factor model given a large size discrepancy. Therefore, rather than utilizing the uninformative prior, we link individual factors from each data using $\mathbf{f}_i^C \sim \mathcal{N}(\mathbf{g}_i^C, \Sigma)$ based on the hypothesis that gene expression is directly influenced by CNVs. This will prevent difference in data size from overwhelming the information available on associations between them. In addition, the systematic error between two data sets will be considered by estimation of the covariance matrix Σ.

Updated Distributions

• $p(\beta_{g,j}|-)$:

For factor j, let $x^*_{g,j} = x_{g,j} - \sum_{j=1}^{r} b_{g,j} h_{j,i} - \sum_{l \neq j}^{k} \beta_{g,l} f_{l,i}$, so that $x^*_{g,j} \sim \mathcal{N}(\beta_{g,j} f_{j,i}, \phi_g)$. In order to be mathematically identifiable for $\mathbf{B}$, we assume the regression coefficients a lower triangular matrix with positive diagonal elements (Carvalho, 2006). This gives the following posterior updates where $g \neq j$:

$$
\begin{aligned}
p(\beta_{g,j}|-) &\propto \prod_{i=1}^{n} p(x^*_{g,j}|\beta_{g,j} f_{j,i}, \phi_g) p(\beta_{g,j}) \\
&= \prod_{i=1}^{n} \mathcal{N}(x^*_{g,j}; \beta_{g,j} f_{j,i}, \phi_g)((1-\rho_j)\delta_0(\beta_{g,j}) + \rho_j \mathcal{N}(\beta_{g,j}; 0, \tau_j)) \\
&= (1-\hat{\rho}_j)\delta_0(\beta_{g,j}) + \hat{\rho}_j \mathcal{N}(\beta_{g,j}; \mu_{g,j}, \Omega_{g,j})
\end{aligned}
$$

where $\Omega_{g,j} = (\tau_j^{-1} + \sum_{j=1}^{k} f^2_{j,i}/\phi_g)^{-1}, \mu_{g,j} = \Omega_{g,j}(\sum_{i=1}^{n} x^*_{g,i} f_{j,i})\phi_g^{-1}$ and $\beta_{g,j} \neq 0$ with probability

$$
\hat{\rho}_j = \frac{\rho_j}{\rho_j + (1-\rho_j)\frac{\mathcal{N}(0;0,\tau_j)}{\mathcal{N}(0;\mu_{g,j},\Omega_{g,j})}}
$$

For the constrained diagonal elements of $\mathbf{B}$, the posterior conditional distribution is given as

$$
p(\beta_{j,j}|-) \sim \mathcal{N}(\mu_{j,j}, \Omega_{j,j})\mathbf{I}(\beta_{j,j} > 0)
$$

with similar forms of $\mu_{j,j}$ and $\Omega_{j,j}$.

- $p(\rho_j|-)$:

$$
\begin{aligned}
p(\rho_j|-) &\propto \prod_{j=1}^{k} p(\beta_{g,j}|\rho_j) p(\rho_j) = (1-\rho_j)^{p_x - j - S_j} \rho_j^{S_j} Beta(\rho_j; s_0, l_0) \\
&\sim Beta(s_0 + S_j, l_0 + p_x - j - S_j)
\end{aligned}
$$

with $S_j = \sum_{g=j}^{p_x} \mathbf{I}(\beta_{g,j} \neq 0)$.

$$
\begin{aligned}
p(\tau_j|-) &\propto \prod_{g=1}^{p_x} p(\beta_{g,j}|\rho_j, \tau_j) p(\tau_j) = \prod_{g=1}^{p_x} \mathcal{N}(\beta_{g,j}; 0, \tau_j) Ga({\tau_j}^{-1}; \frac{a_\tau}{2}, \frac{b_\tau}{2}) \\
&\sim InvGamma(\tau_j; \frac{a_\tau + \omega_j}{2}, \frac{b_\tau + \sum_{g=1}^{p_x} \beta^2_{g,j}}{2})
\end{aligned}
$$

with $\omega_j = \sum_{g=j}^{p_x} \mathbf{I}(\beta_{g,j} \neq 0)$.

- $p(\mathbf{f}_i|-), p(\mathbf{g}_i|-)$:

Let $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_n]$. The posterior distribution of $\mathbf{F}$ can be updated as:

$$p(\mathbf{F}|-) \propto p(\mathbf{X}|\mathbf{F}, \mathbf{B}, \mathbf{\Phi})p(\mathbf{F}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\mathbf{f}_i, \mathbf{B}, \mathbf{\Phi})p(\mathbf{f}_i)$$

$$= \prod_{i=1}^{n} \mathcal{N}(\mathbf{x}_i - \mathbf{B}_h\mathbf{H}_i; \mathbf{B}\mathbf{f}_i, \mathbf{\Phi})\mathcal{N}(\mathbf{f}_i; \mathbf{g}_i, \Sigma)$$

$$\propto \prod_{i=1}^{n} \mathcal{N}(\mathbf{f}_i; \mathbf{E}1_i, \mathbf{V}1_i)$$

where $\mathbf{V}1_i = (\Sigma^{-1} + \mathbf{B}'\mathbf{\Phi}^{-1}\mathbf{B})^{-1}$, $\mathbf{E}1_i = \mathbf{V}1_i(\mathbf{B}'\mathbf{\Phi}^{-1}(\mathbf{x}_i - \mathbf{B}_h\mathbf{H_i}) + \mathbf{G}_i\Sigma^{-1})$.

Similarly $p(\mathbf{g}_i|-)$ takes the form

$$p(\mathbf{G}|-) \propto \prod_{i=1}^{n} \mathcal{N}(\mathbf{g}_i; \mathbf{E}2_i, \mathbf{V}2_i)$$

where $\mathbf{V}2_i = (\mathbf{I} + \mathbf{A}'\mathbf{\Psi}^{-1}\mathbf{A})^{-1}$, $\mathbf{E}2_i = \mathbf{V}2_i(\mathbf{A}'\mathbf{\Psi}^{-1}(\mathbf{y}_i - \mathbf{A}_h\mathbf{H_i}))$. $\mathbf{\Psi}$ is the covariance matrix of $\mathbf{y}_i$.

- $p(\phi_g|-)$:

$$p(\phi_g|-) \propto \prod_{i=1}^{n} p(x_{g,i}|\beta_g f_i, \phi_g)p(\phi_g)$$

$$= \prod_{i=1}^{n} \mathcal{N}(x_{g,i} - \sum_{j=1}^{r} b_{g,j}h_{j,i}; \beta_g f_i, \phi_g)Ga({\phi_g}^{-1}; \frac{a_{\phi_x}}{2}, \frac{b_{\phi_x}}{2})$$

$$\sim InvGamma(\phi_g; \frac{a_{\phi_x} + n}{2}, \frac{b_{\phi_x} + \sum_{i=1}^{n}(x_{g,i} - b_g h_i - \beta_g f_i)^2}{2})$$

2.3.4 Example: joint analysis of ovarian cancer gene expression and CNVs

We applied our joint factor model on ovarian cancer gene expression and CNV data from the TCGA project. This study is aimed to detect correlations between them,

which will lead to the identification of pivotal genomic determinants of cancer phenotypes. We used data from 74 ovarian cancer individuals and 1 disease-free patient. In order to capture genes with differential expression patterns and their association with the CNVs in the narrowed chromosomal regions, we established a filtering criteria: 1) select Affymetrix HT_HG-U133A probes with sample mean above 8, and standard deviation above 0.6; take out probes without matched gene symbols. It results in a gene expression data set downsized from 22277 to 921 probes; A more relaxed thresholding will generate a larger dataset, and we will discuss about it in chapter 3. 2) apply the basic Bayesian factor model 2.1, i.e., the one that only analyzes one data set, and generate signature expression factors; 3) remove CNV segments (Agilent Human Genome CGH 244A probes) not showing significant correlation (p-value < 0.01 after Bonferroni correction) with the gene expression factors. It reduced the CNV data dimension from 227613 to 7278. Therefore, we fitted our joint factor model 2.1 and 2.2 to the shrunk data.

We obtained 11 factors in the two data sets, i.e., $\mathbf{F}_{11*75}$ and $\mathbf{G}_{11*75}$, and selected the most strongly associated pair using Pearson correlation. It turns out that the largest factor loadings in the corresponding CNV factor come mostly from the long arm of chromosome 8 (figure 1A), that the factor correlates well with the paired gene expression factor (figure 1B), and that the gene expression factor correlates with individual SNP observations in the long arm of chromosome 8 (figure 1C). Based on these results, we further examined the genes loaded on this correlated CGH factor and gene expression factor. By ranking the squared factor loadings, we selected the top 16 Affymetrix probe sets (Table 3.1) and 178 CGH probe sets, because the variance in these probes are best explained by the corresponding factors compared with all other data. Pearson correlations between the values of mRNA expression levels and copy number variations were calculated on these heavily loaded genes. We noted that the copy number gains of EBAG9 (CGH probe position: 8q23.2, size 60

bp; mean copy number 2.63 (1-6)) and MTDH (CGH probe position: 8q22.1, size 60 bp; mean copy number 2.38 (1-6)) significantly accompanies their overexpression of mRNAs in the corresponding regions, where correlation coefficients indicate a good linearity between CNVs and gene expression with r = 0.758 for EBAG9 and r = 0.806 for MTDH. Interestingly, in the same factor, 3 CGH loci with duplicated DNAs show significant correlation with MTDH overexpression (r>0.8, p-val<0.01) and are located 0.2M upstream, 5M and 12M downstream of MTDH CGH locus, respectively; and 11 CGH loci are identified with copy number gain and 3Mb upstream of EBAG9 CGH clone (r>0.75, p-val<0.01). These findings may provide evidence for distant regulatory of transcription elements or interactions within a potential gene network.

The product of EBAG9 has been identified as an estrogen receptor binding site associated antigen 9 identical to RCAS1 (Nakashima et al., 1999). Overexpression of EBAG9/RCAS1 inhibits growth of tumor-stimulated host immune cells and induces their apoptosis (Nakashima et al., 1999). Furthermore, it has been reported that RCAS1 is expressed with high frequency in ovarian and lung cancers (Akahira et al., 2004; Iwasaki et al., 2000), and the copy numbers of the region increase in breast cancer (Rennstam et al., 2003). These lines of evidence, together with the results obtained above, imply that overexpression of EBAG9 in ovarian serous cystadenocarcinoma may be triggered by increased gene copy number, which is likely to play an important role in the immune escape of tumor cells and causing cancer progression.

In addition, MTDH, also known as AEG1, is an oncogene cooperating with Ha-ras as well as functioning as a downstream target gene of Ha-ras and may perform a central role in Ha-ras-mediated carcinogenesis (Lee et al., 2007). Overexpression of this gene has been reported in various cancers including breast, brain, prostate, melanoma and glioblastoma multiforme (Emdad et al., 2007; Kikuno et al., 2007). In particular, it has been revealed that MTDH overexpression is associated with 8q22 chromosomal gain in breast cancer, and has been considered as an important

therapeutic target for enhancing chemotherapy efficacy and reducing metastasis risk (Hu et al., 2009). Therefore, we believe that, our results along with the above findings suggest the copy number gain activated MTDH overexpression is a potential indicator in epithelial ovarian cancer.

Validations on the above hypotheses regarding critical genes in cancer progression and their regulation mechanisms can be carried out in several directions. A number of databases can be used to validate these hypotheses. For instance, GATHER and GOrilla are good resources to annotate gene functions; Tumorscape helps interpret copy number variations; DAVID Bioinformatics provides pathway analysis for genes identified by the model. In addition, experimental validation can be performed to quantitatively justify that the activation/inactivation of identified genes are caused by copy number variations. Moreover, we could identify drug susceptibilities of these candidates by searching against reference information from DrugBank (http://www.drugbank.ca), then using these results for experimental validation. The general approach is to grow cell lines in the presence of a particular treatment, whose genomic drivers are disrupted by the introduction of RNA inference and transfection with viral plasmids. A similar strategy can also been applied to predict potential therapies by the identification of new drug targets. Therefore, these will lead to a greater understanding of cancer progression, and allow the identification of combined therapies for individual tumors.

Tumor segmental aneuploidy association with gene expression factors has been demonstrated in a previous study (Lucas et al., 2010) that it makes significant contributions to variation in gene signature of breast cancer under the stress of lactic acidosis or hypoxia. We are interested to test if this is consistent in other tumor tissues, which will provide potential treatment choices for different cancers. We used a similar approach (Lucas et al., 2010) by projecting the breast expression factors into TCGA ovarian and glioblastoma gene expression data and identified correlated

CNVs under the same interventions of lactic acidosis/hypoxia. The ability of projecting the factor model into other data sets allows the possibility of comparing new experimental data to different genomic information, such as CNVs from aCGH. The underlying assumption is that genes showing shared expression patterns in tumors of different origins can be represented by the same loadings matrix. Therefore, in order to estimate the factor scores for the new data, this translates into a well known problem of inverse regression $\mathbf{F}_y = (\mathbf{I}_k + \mathbf{B}'\mathbf{\Phi}^{-1}\mathbf{B})^{-1}\mathbf{B}'\mathbf{\Phi}^{-1}\mathbf{Y}$, where $\mathbf{B}$ is the loadings matrix and $\mathbf{\Phi}$ the diagonal matrix containing the gene by gene variance estimators in the original data, $\mathbf{Y}$ the new set of expression data and $\mathbf{F}_y$ the factor scores on the new data set. With this approach, we estimated factor scores for the TCGA data and calculated their correlations with CNVs. In our analysis, about half of the breast expression factors are also associated with copy number variations in ovarian cancer and that about a quarter are associated with CNVs in glioblastoma. For example, the CNV activated expression pattern in breast cancer (not shown) is also discovered in both ovarian cancer and glioblastoma within the same region (figure 2A and 2B). Therefore, it is likely that similar CNVs might be selected under the same pressure of hypoxia/ lactic acidosis in difference cancers.

2.4 Conclusions

This chapter has built upon a basic understanding of a layout on the correlation between copy number variations and gene expression to deepen knowledge of key concepts and methods. By introducing and comparing a diverse range of techniques for measuring CNVs, we provide the scope of localizing cancer related genes using different platforms. By describing an appreciation of the use of several statistical methods to assist the positioning of CNV regions, we are aimed to better identify cancer driver mutations within the copy number gain/loss regions. Moreover, we have also included examples from TCGA project to show the unique features of

CNV data.

The key challenge of finding candidate drivers is to distinguish it from passenger genes, which are physically located close to the driver mutations and whose variations are not causal to convey growth advantage on cancer cells. In our analysis, we focus on genes with cis-regulated CNVs, and postulate that cancer driver mutation is associated with the expression of a group of genes, and it is likely to localize in DNA amplified or deleted regions in tumors. This is because DNA dosage variations may result in functional changes of affected genes, and thus cause expression change of downstream genes. We have proposed a generic framework to jointly analyze disparate data sets, which is extendable to incorporate diverse information such as proteomics data. This will allow for more robust analysis of the relationship between mRNA expression and protein abundance. Our results not only identify candidate genes whose mRNA expression is statistically significantly correlated with their CNVs, but also successfully recover the region where similar gene expression pattern is triggered by the same genomic program across tumors of different organ systems. This approach is able to estimate the probability of each gene regulated by genomic sources and the relative importance of each source. Additionally, two genes, EBAG9 and MTDH, suggest that abnormal abundance in their DNA copy numbers may contribute to proliferation in ovarian serous cystadenocarcinoma. For these two predicted drivers, we also find many CNVs in the same region but poorly correlated with their gene expression, thus consider them no apparent effect in cancer. Copy number variation is only one of many ways that gene expression can be altered. We believe that a number of complementary approaches are needed to validate possibly driving alterations, as illustrated in the previous section. Therefore, we envision that our model is used as screening guidance to assist the identification of potential cancer drivers with possibly therapeutic importance.

Our work presents a framework toward a broad understanding of the genomic

determinants of cancer. With this approach, we anticipate being able to generate testable biological hypothesis regarding the regulation of cellular responses, which is a tremendous advantage over any single data analyses that can be obtained from gene expression or CNVs alone. This will lead to a greater, more integrated understanding of cellular mechanism, and will allow the identification of genomic regulators as well as enhancement of anticancer drug specificity targeting those regulators. This is key to the discovery of potential combination therapies for the treatment of human cancer. Moreover, genomic patterns related to therapeutic response and clinical outcomes can be identified as biomarkers, which will improve early cancer detection, prognosis and outcome prediction as well as treatment selection. All in all, this will create a comprehensive picture of heterogeneity in tumor genomes, and offer a valuable starting point for new therapeutic approaches.

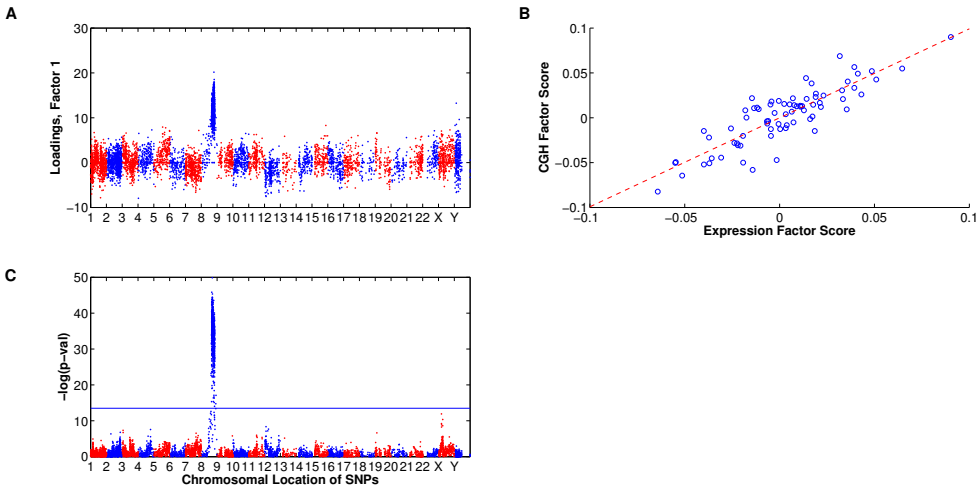


FIGURE 2.1: Factor analytic relationship between CNV and gene expression. Panel A shows the factor loadings from the first factor of the joint factor model fit to CNV data. Panel B shows a scatterplot of significant correlation between gene expression factor and the CNV factor, of which it is linked to. Panel C shows the significance of correlation between the expression factor and each individual SNP from the high-density CGH array. The y-axis shows the -log(p-value) of the Pearson correlation between CNVs and gene expression factor. The horizontal line shows the threshold of p-value less than 0.01 after Bonferroni correction for multiple testing.

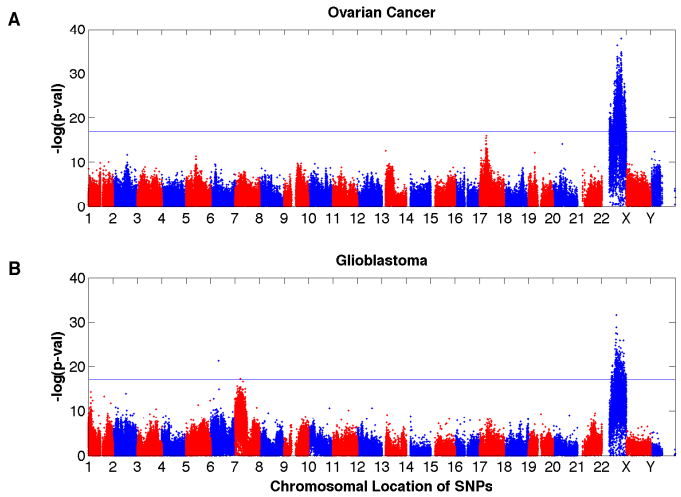


FIGURE 2.2: Panel A and B show the the association between gene expression factor and CNVs across tumors of different origins. Each scatter plot indicates the evidence of association between the same factor that was learned on breast cancer data and copy number changes of different tumor tissues. Plot A shows correlation between the factor, projected onto ovarian cancer expression data, and ovarian CGH data. Plot B shows the same for Glioblastoma. Each point corresponds to one of the SNPs measured in the high-dimensional CGH array. The y-axis shows the -log(p-value) of the Pearson correlation between CNVs and gene expression factor. The horizontal line shows the threshold of p-value less than 0.01 after Bonferroni correction for multiple testing.

Table 2.1: Genes on chromosome 8 showing significantly differential expression in ovarian cancer. The list is ranked by the squared factor loadings.

Gene symbol	Gene
MTDH	LYRIC/3D3 (UID: 92140)
EBAG9	estrogen receptor binding site associated, antigen, 9 (UID:9166)
YWHAZ	tyrosine 3-monooxygenase (UID:7534)
LAPTM4B	lysosomal protein transmembrane 4 beta (UID:55353)
ESRP1	epithelial splicing regulatory protein 1 (UID:54845)
NBN	nibrin (UID:9048)
RAD21	RAD21 homolog (S. pombe) (UID:5885)
RNF139	ring finger protein 139 (UID:11236)
ZNF706	HSPC038 protein (UID:51123)
AZIN1	antizyme inhibitor 1 (UID:51582)
DERL1	Der1-like domain family, member 1 (UID:79139)
ENY2	enhancer of yellow 2 homolog (Drosophila) (UID:56943)
EXT1	exostoses (multiple) 1 (UID:2131)
CTSB	cathepsin B (UID:1508)
DECR1	2,4-dienoyl CoA reductase 1, mitochondrial (UID:1666)
PTDSS1	phosphatidylserine synthase 1 (UID:9791)

3

Nonparametric Joint Bayesian Factor Analysis–A Vertical Integration Approach to Model Multi-platform Genomics Data

A nonparametric Bayesian factor model is proposed for integrating multiple disparate, but statistically related datasets. The approach is based on factorizing the latent space (feature space) into a shared component and a data specific component with the dimensionality of these components (spaces) inferred via a beta-Bernoulli process. The proposed approach is demonstrated by jointly analyzing multiple types of genomic data, including gene expressions, copy number variations and methylation for ovarian cancer patients, and show that the proposed model can potentially uncover key drivers related to cancer.

3.1 Introduction

An important research problem in statistical signal processing and machine learning is the integration/fusion of multiple disparate, but statistically related datasets. For example, in genomic signal processing, integration of DNA copy number variation

and gene expression may help identify key drivers in cancer mechanism. Though the range of potential applications is immense, the increase in data dimensionality, data heterogeneity and the presence of noise often makes such data fusion problems extremely challenging.

A key assumption employed when modeling such high-dimensional data is that the intrinsic dimension of the data is much lower than the observed data dimension, i.e., the data lie in or are close to a low-dimensional subspace. For modeling multiple disparate datasets, approaches often rely on the assumption that the data are different manifestations of a single shared low-dimensional latent space (feature space). The problem then lies in identifying this low-dimensional shared feature space and the data specific mappings from this shared space to the observed data. Classical data analysis techniques for multiple datasets, such as canonical correlation analysis (CCA) (Hotelling, 1936; Borga, 1998; Hardoon et al., 2004), compute a low-dimensional shared linear embedding of a set of variables, such that the correlations among the variables is maximized in the embedded space. Probabilistic approaches to CCA have been proposed in (Bach and Jordan, 2005; Wang, 2007; Rai and Daume, 2009). For joint analysis of multiple data sets, Bach and Jordan (2005); Wang (2007); Rai and Daume (2009) assume the existence of underlying shared latent variables and conditional independence of the data given the latent variables. However the assumption of a single shared latent space may be limiting, and a more flexible approach is to factorize the latent space into a component that is shared among all datasets and a component that is specific to each. Such models are more likely to capture the shared features among all datasets while still preserving the idiosyncratic features unique to each.

Bayesian and semi-Bayesian latent variable models have developed to factorize the latent space into a shared and data-specific part (Archambeau and Bach, 2008; Klami and Kaski, 2008). However, in these approaches the number of latent factors

are chosen *a priori*. Alternatively, one may consider multiple factor models, each with a different number of factors, and perform model selection based on information criteria such as AIC (Akaike, 1987) or BIC (Schwarz, 1978). However, as it is often challenging to check modeling assumptions in high-dimensions, a nonparametric or semiparametric model is desirable. In this chapter we propose a nonparametric Bayesian factor analysis approach for integrating multiple heterogeneous datasets, with the number of factors *inferred* from the available data. Our proposed approach is based on factoring the latent space into shared and data-specific components, employing a beta-Bernoulli process (Griffiths and Ghahramani, 2005; Thibaux and Jordan, 2007; Paisley and Carin, 2009) to infer the dimension of these latent spaces.

We demonstrate the proposed approach on the joint analysis of genomic data for ovarian cancer patients, including three different datasets: gene expressions levels, copy number variations and DNA methylation levels data. We demonstrate that the joint analysis of gene expressions/copy number variations and gene expressions/DNA methylation levels can potentially identify genomic and epigenomic regulators influencing cancer pathophysiology outcomes.

A preliminary version of the model developed in this chapter was presented in (Ray and Carin, 2011). However, this chapter extends substantially (Ray and Carin, 2011) and includes the analysis of a new dataset (heterogeneous genomic data) and also provides detailed theoretical analysis as well as stronger empirical justification of the model. The remainder of the chapter is organized as follows: In Section 3.2 we present the proposed hierarchical Bayesian model for jointly analyzing heterogeneous data and in Section 3.6 we demonstrate the performance of the joint factor model on the analysis of heterogeneous genomic data.

3.2 Joint Bayesian factor analysis-a nonparametric model

Let $\left\{\boldsymbol{X}^{(r)}\right\}_{r=1,R}$ represent data from R different modalities, where $\boldsymbol{X}^{(r)} = (\boldsymbol{x}_1^{(r)}, \ldots, \boldsymbol{x}_M^{(r)}) \in \mathbb{R}^{N_r \times M}$. In sparse factor modeling, learning a single shared matrix of factor loadings for different signal classes has been proposed in (Mairal et al., 2008). However for heterogeneous data such as that considered here, learning a shared set of factor loadings is more difficult.

The joint factor model may be represented as

$$\boldsymbol{X}^{(r)} = \boldsymbol{D}^{(r)} \left(\boldsymbol{W}^{(c)} + \boldsymbol{W}^{(r)}\right) + \boldsymbol{E}^{(r)} \tag{3.1}$$

The matrix $\boldsymbol{D}^{(r)} = (\boldsymbol{d}_1^{(r)}, \ldots, \boldsymbol{d}_K^{(r)}) \in \mathbb{R}^{N_r \times K}$ consists of the factor loadings specific to data modality r, factor scores $\boldsymbol{W}^{(r)} = (\boldsymbol{w}_1^{(r)}, \ldots, \boldsymbol{w}_M^{(r)}) \in \mathbb{R}^{K \times M}$ are *specific* to data from modality r, $\boldsymbol{W}^{(c)} = (\boldsymbol{w}_1^{(c)}, \ldots, \boldsymbol{w}_M^{(c)}) \in \mathbb{R}^{K \times M}$ consists of the factor scores *common* among all modalities, and $\boldsymbol{E}^{(r)} = (\boldsymbol{\epsilon}_1^{(r)}, \ldots, \boldsymbol{\epsilon}_M^{(r)}) \in \mathbb{R}^{N_r \times M}$ consists of the noise/residual specific to data of modality r.

We wish to impose the condition that any $\boldsymbol{x}_i^{(r)}$ is a sparse linear combination of the factor loadings. Hence, the factor scores are represented as,

$$\boldsymbol{w}_i^{(r)} = \boldsymbol{s}_i^{(r)} \odot \boldsymbol{b}_i^{(r)} \quad \text{and} \quad \boldsymbol{w}_i^{(c)} = \boldsymbol{s}_i^{(c)} \odot \boldsymbol{b}_i^{(c)} \tag{3.2}$$

where $\boldsymbol{s}_i^{(r)} \in \mathbb{R}^K$, $\boldsymbol{s}_i^{(c)} \in \mathbb{R}^K$, $\boldsymbol{b}_i^{(r)} \in \{0,1\}^K$, $\boldsymbol{b}_i^{(c)} \in \{0,1\}^K$ and $\odot$ represents the Hadamard product (elementwise vector product).

The sparse binary vectors $\boldsymbol{b}_i^{(r)}$ are drawn from the following beta-Bernoulli process (Griffiths and Ghahramani, 2005; Thibaux and Jordan, 2007; Paisley and Carin,

2009)

$$\boldsymbol{b}_i^{(r)} \sim \prod_{k=1}^{K} \text{Bernoulli}(\pi_k) \tag{3.3}$$

$$\boldsymbol{\pi} \sim \prod_{k=1}^{K} \text{Beta}(c\alpha, c(1-\alpha)) \tag{3.4}$$

with π_k representing the kth component of $\boldsymbol{\pi}$ and $\alpha \in (0,1)$. In practice K is finite, and the above equation represents a finite approximation to the beta-Bernoulli process, where the number of non-zero components of each $\boldsymbol{b}_i^{(r)}$ is a random variable drawn from Binomial (K, α). If α is set to $\frac{\rho}{K}$, in the limit $K \to \infty$ this reduces to the number of non-zero components in $\boldsymbol{b}_i^{(r)}$ being drawn from Poisson(ρ); this corresponds to the Indian buffet process (IBP) (Griffiths and Ghahramani, 2005; Thibaux and Jordan, 2007; Paisley and Carin, 2009). We may therefore explicitly impose a prior belief on the number of non-zero components in $\boldsymbol{w}_i^{(r)}$. The shared binary vectors $\boldsymbol{b}_i^{(c)}$ are modeled similarly as $\boldsymbol{b}_i^{(r)}$. The noise or residual in (3.1) is modeled as

$$\boldsymbol{\epsilon}_i^{(r)} \sim \mathcal{N}(0, {\gamma_\epsilon^{(r)}}^{-1} \boldsymbol{I}_{N_r}), \quad \gamma_\epsilon^{(r)} \sim \text{Gamma}(a_0, b_0) \tag{3.5}$$

where $\boldsymbol{I}_{N_r}$ represents the $N_r \times N_r$ identity matrix.

The construction in (3.1) imposes the belief that there are underlying (low-dimensional) features represented by the factor scores that may be shared across modalities, via $\boldsymbol{W}^{(c)}$; however, each modality has a unique mapping from these low-dimensional factor scores to the high-dimensional data, reflected by $\boldsymbol{D}^{(r)}$. Further, each modality may also have idiosyncratic low-dimensional features, characterized by $\boldsymbol{W}^{(r)}$. The common and idiosyncratic features are learned jointly, via the simultaneous analysis of all modalities. A unique feature of the above construction is that it allows *complete* sharing of some low-dimensional features across different data

modalities as well as *partial* sharing, i.e., a shared feature may be slightly perturbed via $\boldsymbol{W}^{(r)}$ and shared across different modalities.

3.3 Imposing structure on factor loadings

3.3.1 Simple construction

In the absence of covariates, the factor loadings may be drawn i.i.d. from a Gaussian distribution (for ease of notation, we henceforth drop the modality index r, unless referring to multiple data modalities simultaneously),

$$\boldsymbol{d}_k \sim \mathcal{N}(0, \gamma_s^{-1}\boldsymbol{I}_N), \quad \gamma_s \sim \text{Gamma}(a_5, b_5) \tag{3.6}$$

3.3.2 Imposing sparsity

In many biological applications, it is desirable that the factor loading matrix is sparse (Carvalho et al., 2008a). To impose sparsity on the factor loadings, we employ a Student-t sparseness-promoting prior (Tipping, 2001). In this construction, d_{jk}, the jth component of $\boldsymbol{d}_k$, is drawn

$$d_{jk} \sim \mathcal{N}(0, \tau_{jk}^{-1}) \tag{3.7}$$

$$\tau_{jk} \sim \text{Gamma}(a_1, b_1) \tag{3.8}$$

However, there are multiple ways one may desire to impose sparsity, such as using the spike-slab prior (Ishwaran and Rao, 2005; Carvalho et al., 2008a; Chen et al., 2011). This consists of a discrete-continuous mixture of a point mass at zero, referred to as the 'spike', and any other distribution, such as the Gaussian distribution, known as the 'slab'. A hierarchical beta-Bernoulli construction of the spike-slab prior for imposing sparsity on the factor loadings is provided in (Chen et al., 2011). We found that the spike-slab prior works as well as the model presented above; however, for

the sake of brevity, we include only the results for the Student-t sparseness prior in this chapter.

3.4 Imposing structure on factor scores

In the simplest scenario, the factor scores may be drawn i.i.d. from a Gaussian distribution as,

$$\boldsymbol{s}_i^{(r)} \sim \mathcal{N}(0, \gamma_r^{-1}\boldsymbol{I}_K) \qquad \boldsymbol{s}_i^{(c)} \sim \mathcal{N}(0, \gamma_c^{-1}\boldsymbol{I}_K) \tag{3.9}$$

We impose broad gamma prior on γ_r and γ_c: $\gamma_r \sim \text{Gamma}(a_2, b_2)$ and $\gamma_c \sim \text{Gamma}$ (a_3, b_3).

3.5 MCMC inference

The conditional posterior distribution of all the model parameters for the joint factor model may be derived analytically. We use a Gibbs sampler to draw samples from the posterior distribution of the model parameters. For the factor analysis results on multiple genomic data presented in Sections 3.6.1 and 3.6.2, the number of Gibbs burn-in samples is set to 3000 and the number of collection samples is set to 1000. Broad gamma hyperpriors are chosen for the variance terms with $a_0 = b_0 = a_2 = b_2 = a_3 = b_3 = 10^{-5}$. The results are relatively insensitive to these settings and various other settings such as $a_0 = b_0 = a_2 = b_2 = a_3 = b_3 = 10^{-3}$ or $a_0 = b_0 = a_2 = b_2 = a_3 = b_3 = 10^{-6}$ yielded very similar results. The shrinkage parameters on the factor loadings are set at $a_1 = 10^{-3}$ and $b_1 = 10^{-6}$ (for Gene-copy number analysis results) and $a_1 = 1$ and $b_1 = 10^{-2}$ (for Gene-Methylation analysis results).

Since much correlation is encoded in the priors, the mixing of the MCMC sampler was also carefully examined. The sampler was run extensively for different number of burn-in and collection samples. It was also run multiple times in parallel with

different initial values. The results of these experiments were found to be consistent and repeatable across such runs.

3.6 Joint analysis of multi-platform genomic data

There are numerous publications on combining different types of DNA modifications with gene expression. Perhaps the most natural of these are the brute force methods such as expression quantitative trait loci (eQTL) analysis (Kendziorski et al., 2006). Joint analysis of single nucleotide polymorphism (SNP) data with gene expression data by eQTL involves testing every gene-SNP pair for association with a t-test, then corrects for multiple hypothesis testing. CNAmet (Louhimo and Hautaniemi, 2011) defines a similar approach to relate gene expression changes with either copy number change or DNA methylation. Other approaches use well established models for each of the individual data types, then combine the results into a statistic that addresses the problem of interest. The approach of Jeong et al. (2010) is an example of this for the identification of genes that are regulated by DNA methylation. A shortcomming of all of these approaches is that they do not reduce the dimension of the individual data sets through an accounting of their respective correlation structures.

In (Lanckriet et al., 2004) the authors used kernel functions predefined for each data type, and mapped to the same vector space, which allows joint analysis in the common range of the kernels. Copy number and expression in cancer (CONEXIC) (Akavia et al., 2010) has been proposed as a Bayesian scoring function that measures how well a set of candidate gene regulators correlate with the expression of gene *modules* (groups of genes that are correlated with each other). Another approach (Lucas et al., 2010) utilizes a sparse factor model to model the correlation structure of the gene expression data, but uses post-hoc hypothesis tests to draw connections between gene expression and copy number data. These approaches do allow for effective dimension reduction, but don't use correlation structure in one data set to

inform estimations of correlation in the others.

The most direct approach to jointly modeling the correlation structure of heterogeneous genomic data is to require the factor matrix to be shared, as in Shen et al. (2009). Their model does not contain a data-type-specific factor structure equivalent to $\boldsymbol{W}^{(r)}$ in our model, and is therefore somewhat less flexible. In addition, they utilize standard normal distributions on the elements of the factor matrix, eliminating the possibility of discovering factors that are relevant for only a subset of the subjects.

Data description

The data in this study includes ovarian cancer gene expression, copy number variation (CNV) and methylation data collected from the Cancer Genome Atlas (TCGA) project (http://cancergenome.nih.gov/). We aim to integrate gene expression/CNVs and gene expression/methylation from 74 ovarian cancer patients. For computing purposes, we downsized the original massive data into smaller sets. Independent gene-by-gene filtering (based on criteria such as overall mean and overall variance) is typically employed to reduce data dimension as well as increase the number of discoveries in high-throughput experiments (Bourgon et al., 2010; Gentleman et al., 2005; Talloen et al., 2007). In our analysis, a filtering criteria was established for the gene expression data to eliminate probes with sample mean below 6, or standard deviation below 0.4, which resulted in a gene expression data set downsized from 22277 to 5976. Comparative genomic hybridization (CGH) data was filtered to remove Agilent Human Genome CGH 244A probes containing missing values. This set was further filtered by keeping only one in 50 probes, leaving 4443 probes. Methylation data (Illumina Infinium human methylation 27K bead assay) was filtered to retain only higher variance samples (resulting in 4722 probes) and was inverse-probit transformed to lie on the real line.

3.6.1 Analysis of gene expression and copy number variation data

We applied the joint Bayesian factor model to gene expression and CGH in order to identify factors that are representative of correlated changes in gene expression and DNA copy number variations. We set the upper bound on the number of factors as $K = 60$, and obtained 1 specific to gene expression, 4 unique to CNVs and 19 shared between both modalities (Figure 3.1).

Figure 3.2 shows the correlation structure of the probe sets (gene expression) and CGH clones (CNVs) that are included in joint factor number 41 (the factor numbering is arbitrary, and changes between collection samples, with these results are illustrative; in these and related results we depict the maximum likelihood collection sample). As expected, correlation between the factor genes for those patients who were included in this factor is higher than for those not included.

It is well known that some variations in cancer gene expression are caused by gene dosage changes due to CNVs. In addition, because of the mechanism by which CNV occurs, it tends to happen in contiguous regions. Of the 20 CNV factors identified, one is a nearly perfect representation of batch effects in the data and the remaining 19 display copy number amplification/deletion in specific chromosomal regions. Most of these show similar gene expression changes in the same region. We demonstrate this behavior in Figure 3.3, which shows that the largest factor loadings from both CNV and gene expression for factor 18 are clustered around the same region of chromosome 8.

We identified highly associated copy number variations in the chromosomal arm 8q12.3-8q24.13 (factor 18), which is a known region for frequent high-level amplication associated with disease progression in human cancers (Frank et al., 2007; Pils et al., 2005). The rediscovery of genes in this region also validates our approach. For example, E2F5 (8q21.2, Unique ID: 1875), an important gene in the regulation of the

cell cycle, is known to be overexpressed in ovarian epithelial cancer (Kothandaraman et al., 2010). Over-expressed genes, MTDH (8q22.1, Unique ID: 92140) and EBAG9 (8q23, Unique ID: 9166), have been recognized in a variety of cancers including ovarian and breast cancers (Akahira et al., 2004; Rennstam et al., 2003; Emdad et al., 2007). Another gene in this region whose expression level is known to be important in cancer biology is WWP1 (8q21, Unique ID: 11059). This recapitulation of some of the well known features of aneuploidy in cancer suggests that our joint model is appropriately capturing correlation structure between gene expression and CGH data.

As described above, many factors we obtained are associated with individual chromosomal locations, as demonstrated in Figure 3.3. However, there is also a subset of factors (1, 14, 32, 41, 45, 57) which are representative of multiple regions. Figure 3.4 shows that the largest factor loadings in CNV/gene expression for factor 41 come from both chromosome 6 and 17. This is the explanation of the checkerboard regions of positive and negative correlation in Figure 3.2 as well. The copy number variations from the top ranked CGH probes in the two locations are highly negatively correlated, with copy number gain in chromosome 6 and loss in the other. There are a number of possible mechanistic explanations for this feature. For example, it is possible that wholesale duplication of one region is lethal to the cells without shutting down the apoptosis pathway. Such a shut down might be accomplished by deletion of other regions. Previous approaches to the joint analysis of gene expression and CNV through the use of factor models, such as Lucas et al. (2010), have failed to find these relationships.

The proposed joint factor model provides the flexibility of discovering factors that are relevant only for a subset of the subjects. It is interesting to note that a similar model which enforces that all subjects are included in the inferred factors, performed poorly compared to the proposed model and discovered much fewer factors which

captured correlated changes in gene expressions and copy number variations.

3.6.2 Analysis of gene expression and DNA methylation data

For computational purpose, we selected probes with highest variances across samples and obtained 1000 probes for both gene expression and methylation. 18 common factors were thus inferred between the two data sets. Unlike CNVs, methylation does not typically occur in contiguous regions, therefore it is not surprising that no regional peaks were detected. Methylation acts as an epigenetic regulator and silences tumor suppressor genes by changing chromosomal structures. We detected a gene, SPON1(11p15.2, Unique ID: 10418), which appears to be predominantly regulated by methylation of its CpG site (Figure 3.5). Elevated expression of this gene relative to normal tissue is a known hallmark of ovarian cancer (Pyle-Chenault et al., 2005), however, the mechanism of this overexpression was previously unknown. SPON1 encodes VSGP/F-spondin protein promoting proliferation in vascular smooth cell during ovarian folliculogenesis, which has been identified as a potential diagnostic marker or therapeutic target for ovarian carcinoma (Pyle-Chenault et al., 2005; Miyamoto et al., 2001).

In contrast to the almost single gene precision of factor 5, factor 24 shows strong correlation between methylation and gene expression in many different loci across the entire genome (Figure 3.6). The list of CpG sites heavily loaded on this factor are displayed in Table 3.1. Pathway analysis on these candidate genes reveals that many are involved in DNA binding and regulation of transcription. The correlation of methylation levels at all of these sites combined with their correlated gene expression levels suggests that they are all the targets of a single methlylation program, however, the existence of coordinated methylation enzymes that target these locations is unconfirmed.

We implemented the joint factor model for analysis of multiple genomic data in

non-optimized Matlab on a quad core PC with 2.2 GHz CPU and 4 GB ram. The average time per iteration of the Gibbs sampler for the results in Section 3.6.1 is 72 seconds and for the results in Section 3.6.2 is 55 seconds.

3.7 Discussion

3.7.1 Batch effects

TCGA project collected tumor samples in different institutions and at different times, and thus the data can be vulnerable to systematic noise such as batch effects. In the nonparametric joint Bayesian factor analysis model, we applied the same shrinkage prior on all the factor loadings, and thus inferred some factors that represent batch effects rather than genetic variances. Without explicitly modeling the confounding effects, it may cause difficulty in interpretation. Since the batch information is available from TCGA, we can add design matrix to model the sample-wise batch effects. Chapter 2 implemented this procedure using the parametric version of the joint FA model. We found the coefficients of the corresponding design vectors tend to be *dense*, as opposed to the *sparse* loadings. The latter captures the genetic effects well. Therefore, we may add the same design matrix to the nonparametric joint FA model, and that will distinguish batch effects from genetic variances.

In addition, even after the batch effects removal procedure, it is possible that some factors still constitute biological or experimental confounders, while the remaining factors that represent genetically driven variants. In such case, we wish to separately model the confounding components as dense factors and others as sparse factors by imposing different levels of shrinkage. The proportion of factors being confounders or signals can be inferred by imposing a beta-Bernoulli prior on a binary indicator matrix, where confounding factors are inferred with assignment 1 indicating the corresponding loadings have a flat gamma prior, or vice versa. This allows fully automated learning from the data, which eases the interpretation.

3.7.2 Comparison between the parametric and nonparametric joint FA model

In chapters 2 and 3, We proposed two joint FA models, respectively, to vertically integrate different genomics data and infer driver mutations in cancer. The hypotheses between two methods are similar that they assume the data is a linear combination of latent factors plus idiosyncratic noise, where the latent factor space can be further decomposed into a shared component representing the correlation among different data modalities as well as a data-specific component. However, they differ in many ways.

The parametric model links heterogeneous datasets by sampling the factor scores of one data modality around the other, assuming that they have similar values. It is flexible enough to incorporate large/small sample-wise variances in the sampling distribution. This addresses the issue of large dimensionality discrepancy among different array platforms.The nonparametric model constrains the common factors to be exactly the same for different data modalities. It simplifies model assumption, which is easier for computation, especially with more than two data modalities.

The parametric model takes into account batch effects, which are omitted by the nonparametric model. Therefore, in chapter 3, some factors reflect such effect. Section 3.7.1 proposes possible solutions to address this problem.

The main advantage of the nonparametric model is to infer number of factors, shared and unique components automatically from the data. While for parametric model, we need to specify these parameters in advance, which becomes hard to verify especially in high dimension.

In summary, both methods will work for modeling only two datasets. When number of data modalities becomes larger, it seems more straightforward to use the nonparametric approach and infer one set of common factors based on multiple data modalities. On the other hand, the parametric method requires modeling a pair of

two datasets at a time.

3.8 Conclusions

A nonparametric joint factor analysis method is introduced for modeling multiple disparate but statistically related data. The proposed approach was demonstrated on the joint analysis of heterogeneous genomic data related to ovarian cancer. The proposed model uncovered key drivers of cancer, some of which have been previously reported in literature as well as some new genomic causes of cancer (potentially).

In this chapter we have focussed on integrating multiple heterogeneous but statistically correlated datasets, via a joint factor analysis approach where the latent space is factorized into a shared component and data specific components. Moreover, data specific linear mappings from the latent space to the observation spaces where obtained via joint analysis of all data modalities. However, for certain applications, the assumption that the data lie in or close to a low-dimensional subspace is restrictive and a better assumption is that the data lie on a manifold. In the future we wish to relax the linearity assumption of our joint factor model via a mixture of factor analyzers (MFA) approach.

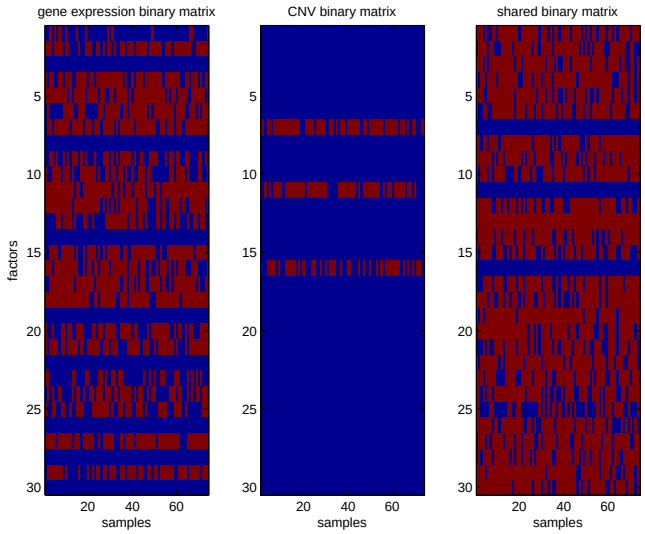


FIGURE 3.1: The estimated feature selection matrices unique to specific data $\boldsymbol{B}^{(r)}$ and common between both modalities $\boldsymbol{B}_c$. From left to right, the heat maps display sparse binary matrix of gene expression, CNVs and the one shared between these two, respectively. The y-axis shows the indicator of each factor, and x-axis represents the 74 subjects. The inferred factors and samples selected by the model are assigned as 1 (red), otherwise 0 (blue).

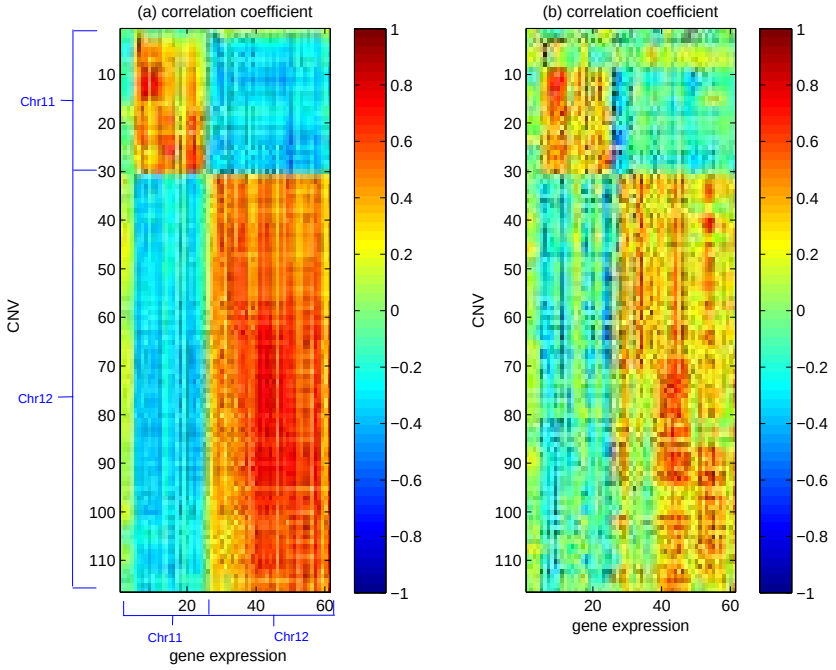


FIGURE 3.2: Correlation structure between gene expression and CNVs of top loaded genes from factor 23. The figure displays correlation coefficients between the two data. Panel (a) and (b) shows the correlation results from patients selected and dropped out by the model, respectively. It is clearly shown that CNVs from chromosome 11 and genes from chromosome 12 has a reverse correlation pattern, or vice versa.

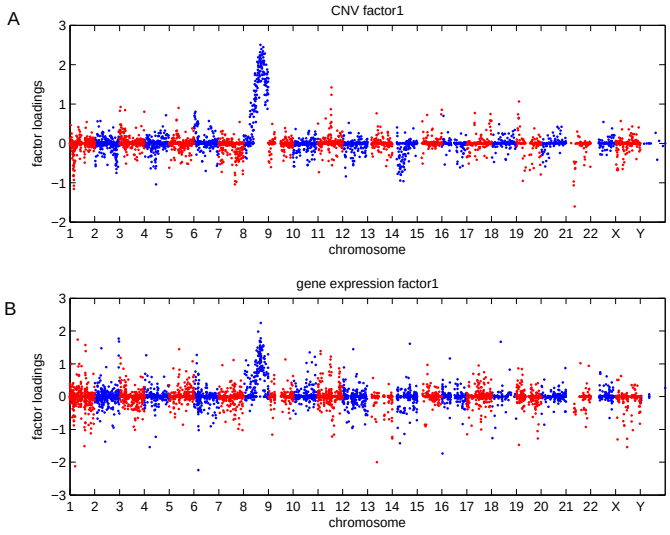


FIGURE 3.3: Factor analytic relationship between CNV and gene expression. The figures show the factor loadings from the first factor of the joint factor model fit to CNV (Panel A) and gene expression data (Panel B), respectively.

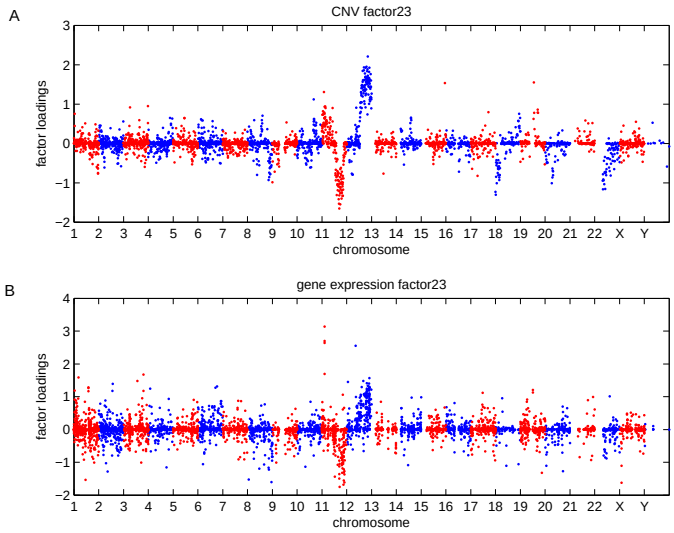


FIGURE 3.4: Dual peaks shown in the loadings of factor 23 of the joint factor model fit to CNV (Panel A) and gene expression (Panel B) data.

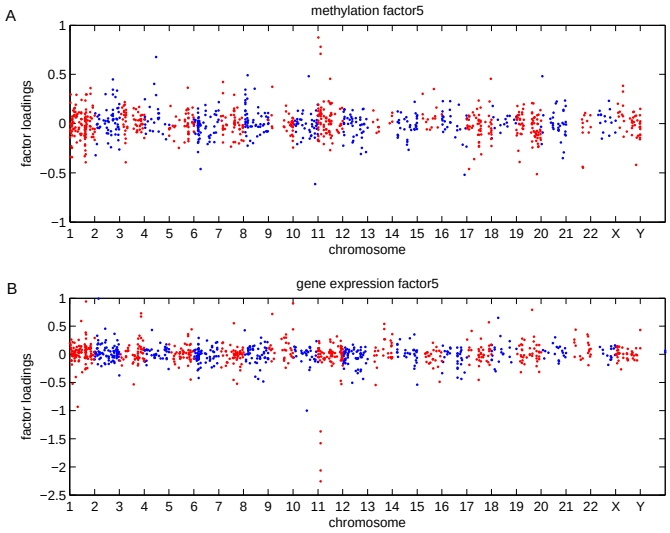


FIGURE 3.5: SPON1 gene identified in the loadings peak from factor 5 of the joint factor model fit to DNA methylation(Panel A) and gene expression(Panel B) data.

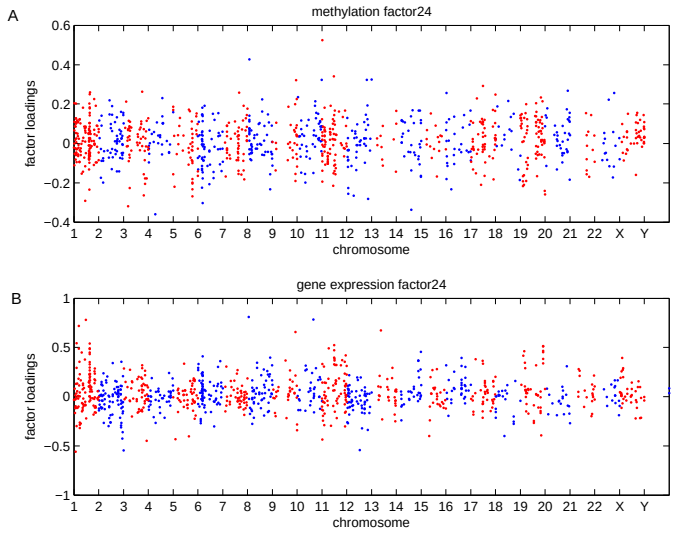


FIGURE 3.6: Loadings from factor 24 with strong correlations between methylation(Panel A) and gene expression(Panel B) at many different loci.

Table 3.1: Genes from factor 24 showing significantly differential methylation pattern in their CpG sites. The list is generated from candidates displayed in Figure 3.6A.

HOXA13	SOCS2	SLC38A4	SPAG6	EPHX3	BST2	PITX2	FERD3L
GPR133	FCRL3	F10	BCAN	ALX4	CDIPT	CPT1C	BCAP31
SOX1	ZNF385D	IGF2AS	ADCY4	EOMES	GATA4	CABP5	PAX7
FLRT1	LEP	TRPA1	HOXD10	PLEKHB1	GPR142	STK19	EVX1
SLC4A11	ZDHHC11	ZNF750	NXN	AJAP1	VSX2	TRH	FOXI1
RAC3	RENBP	MYO3A	GATA4	GRIK3	CARD14	APCDD1L	CA3
PCDHAC1	BCAP31	SNCAIP	CYP4F22	FCN1	SSNA1	GBP4	CASQ1
ARHGAP4	KLHL6	CEACAM3	CEBPG	ABCB4	LYZL4	TRHDE	CDX2
SCML1	PTHLH	KLF11	SLC22A18	DENND2D	C2orf43	PI3	ESX1
CBLN4	MAGEB6	AIM2	ZDHHC8P	HEPACAM2	A2BP1	TERC	C3

4

Joint Bayesian Factor Analysis–A Horizontal Integration Approach to Model Diverse Gene Expression Data for Pathway Analysis

Pathway analysis has become a central approach to understanding the underlying biology of differentially expressed genes. As large amounts of microarray data have been accumulated in public repositories, flexible methodologies are needed to extend the analysis of simple case-control studies in order to place them in context with the vast quantities of available and highly heterogeneous data sets. To address this challenge, we have developed a two-level model, consisting of 1) a joint Bayesian factor model that integrates multiple microarray experiments and ties each factor to a predefined pathway 2) a point mass mixture distribution that infers which factors are relevant/irrelevant to each dataset. Our method can identify pathways specific to a particular experimental trait or concurrently induced/repressed under a variety of interventions. In this chapter we describe the model in depth and provide examples of its utility in simulations as well as real data from a study of radiation exposure. Our analysis of the radiation study leads to novel insights into the molecular basis of

time- and dose- dependent response to ionizing radiation in mice peripheral blood. This broadly applicable model provides a starting point for generating specific and testable hypotheses in a pathway-centric manner.

4.1 Introduction

High-throughput technologies have enabled comprehensive monitoring of biological systems. They provide an entry point to uncovering molecular mechanisms by genome-wide searches for cellular changes in response to any experimental perturbation. However, the generation of large profiles leads to difficulties with interpretation. Researchers have developed knowledge bases (Vastrik et al., 2007; Kanehisa and Goto, 2000) to help tackle this hurdle. These include databases of biological processes or functional groups that explain gene-pathway memberships as well as interactions among genes. By mapping genes in a list to pathways in a database one can decipher molecular mechanisms that link the biological characteristic of the predefined gene set to experimental traits. This approach helps scientists to identify pathways associated with disease which can lead to hypotheses regarding pathway-specific biomarkers and drug targets.

There are a number of approaches to identifying statistically significant associations between gene lists or gene expression profiles and pathways. Given a set of genes and a predefined pathway, we say that the gene set is *enriched* for the pathway if the overlap between the two sets of genes is significantly higher than expected by chance (Shamir, 2010). The particular statistics used for estimating enrichment vary by package and include Fisher exact probability, z-score, Chi-square, hypergeometric distribution and binomial distribution (Huang et al., 2009a). In recent years, a number of publicly available high-throughput functional annotation tools have been proposed using this type of approach.

DAVID (The Database for Annotation, Visualization and Integrated Discovery)

(Dennis et al., 2003) adopted the EASE score (Hosack et al., 2003). It utilizes the Fisher exact test to assess whether the overlap between a gene list and pathway is significant. A jackknifing procedure is applied to assess the stability of the statistic. Interpretation from DAVID relies on a scoring system of negative logarithmic transformed p-values which leads to suggestions for possible relevant annotations.

Another type of enrichment method takes into account the gene-phenotype correlations. GSEA (Gene Set Enrichment Analysis) (Subramanian et al., 2005) ranks the genes based on the significance of differential expression between two experimental groups (e.g. disease versus control), then uses a weighted Kolmogorov-Smirnov(K-S)-like-statistic to compute a gene-set-wise enrichment score (ES). It is designed to evaluate enrichment only in two classes, which imposes a fundamental challenge if a large cohort of data is queried with continuous or multiple phenotypic categories. In order to analyze large transcriptional profiles such as those available in GEO (Edgar et al., 2002) simultaneously, we need a different approach to assess pathway variabilities across a heterogeneous population with diverse phenotypic traits.

Gower et al. (2011) developed openSESAME (Search of Expression Signature Across Many Experiments); This algorithm first calculates a signature association (SA) score for each experimental sample with user defined "up-" and "down-" regulated gene sets. SA summarizes the expression signature of a particular sample as induction, repression, or not changed under a Wilcoxon rank-sum statistic. It then identifies enriched datasets that display co-expression patterns that are similar to the query signature using either K-S or Fisher's exact test. Without the prior knowledge of phenotypic traits, openSESAME searches for coherent differential expression across numerous biological states. However, since the algorithm is designed to query one signature at a time, it is difficult to extend for a simultaneous query across a wide variety of gene signatures.

None of the above mentioned methods take gene correlations into account, which

might result in an increased volume of false positive rates (Tamayo et al., 2012). GSVA (Gene Set Variation Analysis) (Hanzelmann et al., 2013) was thus proposed. This approach uses a Gaussian kernel to estimate the cumulative density function of normalized gene expression data. It follows methods similar to GSEA, which calculates sample-wise enrichment score using maximum deviation of K-S-like-statistic, or normalized ES to account for concordantly over-/under-expressed genes in a pathway. It does not perform pathway-wise enrichment analysis. Instead it requires an additional step of hypothesis testing between experimental and control groups, the evaluation of which eventually depends on the choice of significance threshold. The kernel function does not account for variances among different data sets, therefore, without modification it cannot be applied to integrate multiple microarray experiments.

Our approach to this problem is to extend a factor model-based approach because of its practical application in genomic studies (Zheng and Lucas, 2012; Lucas et al., 2010, 2009, 2006). The factor analysis framework allows dimension reduction and estimation of correlation structure in the context of high-dimensional gene expression data. One objection to this approach is that it generates results that are difficult to interpret. In order to address this, we associate each factor to a predefined pathway or gene set through the use of strongly informative priors. The model is capable of incorporating a large collection of different experimental data sets, which leverages the ever-growing body of publically available transcriptional profiles. In this chapter, we perform pathway selection which simplifies the decision-making procedure of determining which pathway/gene set is most represented under a variety of experimental perturbations. The remainder of the chapter is organized as follows: we present the proposed model in the Methods section, and provide examples of its utility using both synthetic data and a study of radiation exposure. Comparison with existing pathway analysis methods is discussed afterwards, followed by conclusions.

4.2 Methods

Let $g \in 1 \cdots G$ be an index over genes, $d \in 1 \cdots D$ be the index over data sets and $j \in 1 \cdots N_d$ be an index over samples in data set d. We assume a latent factor model for the expression of a particular gene, data set and sample, $x_{g,d,j}$. The model includes a gene and data set specific mean expression, $\mu_{g,d}$ and idiosyncratic noise $\epsilon_{g,d,j}$. We assume that there are K factors with $p \times K$ dimensional loadings matrix β and latent factor expression levels for each factor, data set and sample, $f_{k,d,j}$.

$$x_{g,d,j} = \mu_{g,d} + \sum_{k=1}^{K} \beta_{g,k} f_{k,d,j} + \epsilon_{g,d,j} \tag{4.1}$$

Because we are working with experimental interventions in the background of relatively homogeneous biology, we expect that some factors will be relevant to some data sets and not others and we expect data set specific residuals. With this in mind, we use point mass mixture prior distributions for F that allow factors to be zero for some data sets and non-zero for others. We utilize the notation $\mathbf{0}$ and $\mathbf{1}$ to denote vectors of zeros and ones respectively. The length of the vectors will be implied by the data set to which the variable is associated.

$$\begin{aligned}
f_{k,d,j} &= \gamma_{k,d,j} y_{k,d,j} \\
\gamma_{k,d,\cdot} &\sim (1-\pi_k)\delta_{\mathbf{0}} + \pi_k \delta_{\mathbf{1}} \\
\pi_k &\sim (1-\rho) Beta(\alpha_0, \kappa_0) + \rho Beta(\kappa_0, \alpha_0) \\
\rho &\sim Beta(e_0, l_0) \\
\epsilon_{g,d,j} &\sim N(0, \theta_{g,d})
\end{aligned}$$

In cases where a factor is identified by the model as non-zero, we assume the typical prior distribution for elements of the factor matrix.

$$y_{k,d,j} | \gamma_{k,d,\cdot} = 1 \sim N(0,1)$$

Where variance 1 is assumed in order to address non-identifiability issues between the scale of β and F.

There are a number of published approaches to latent factor modeling that impose constraints on the loadings matrix in order to identify the model. Examples include principal components, non-negative factorization (Schmidt et al., 2009) and sparse factorization (Bhattacharya and Dunson, 2011). We are, however particularly interested in relating our results to known gene pathways such as those published in the KEGG database. As such, we propose a strongly informative prior distribution on the factor loadings matrix that relates each factor to a known pathway through the identification of which genes have non-zero loadings. We presume that the kth pathway in our database consists of a list of genes that belong in the pathway. Let $z_{g,k} \in \{0, 1\}$ indicate whether gene g is in pathway k. We assume $\beta_{g,k} = a_{g,k} z_{g,k}$, which forces the loading for a gene to be zero if the gene is not in the predefined pathway. While this restricts model flexibility significantly, it offers a resolution to one of the loudest criticisms of factor models for gene expression data by providing a clear interpretation to the meaning of each factor. To complete the model specification, we assume that $a_{g,k} \sim N(m_{g,k}, \phi_{g,k})$.

4.3 Experiments: Synthetic data

In this section, we perform a variety of experiments to evaluate the effectiveness and accuracy of our proposed model. We generate synthetic data from the model under a few different circumstances.

Parameters common to all synthetic data generation

We consider the data is of size $\{G, D, K\} = \{9986, 5, 219\}$. This is chosen to approximately match the size of the radiation exposure data we will analyze later. We set a series of 5 datasets, the sample sizes of which are 10, 20, 30, 70, and 250. We use

an external pathway database, Molecular Signatures Database (MSigDB) (Liberzon et al., 2011) from which we take 219 mouse specific pathways, the union of which contains 9986 genes that are measured in our system. We use a range of data set and pathway sizes in order to evaluate how well our approach can recover known parameters under different conditions. The model parameters are drawn from the following distributions.

$$\begin{aligned}
\mu_{g,d} &\sim N(8,2) \\
a_{g,k} &\sim N(0,1) \\
\gamma_{k,d,\cdot} &\sim Bernoulli(\tau_0) \\
y_{k,d,j}|\gamma_{k,d,j} = 1 &\sim N(0,1) \\
\theta_{g,d}^{-1} &\sim Gamma(1.1, 0.02)
\end{aligned}$$

We create a sparsity matrix z from the pathway database and set $\boldsymbol{\beta} = \mathbf{a} \oplus \mathbf{z}$ where $\oplus$ denotes element-wise multiplication. We draw $x_{g,d,j}$ using equation 4.1. Hyperparameters are set as $\{\alpha_0, \kappa_0, e_0, l_0, u_0, n_0, c_0, d_0, a_0, b_0, m_0, s_0\} = \{1, 10, 10, 90, 0, 1, 1.1, 1, 1.1, 0.01, 8, 2\}$. We run the Gibbs sampler for 500 iterations and use the first 250 as burin-in. This process is repeated 10 times.

Ranges of sparsity in factor utilization, τ_0

In the first scenario, we want to explore different sparsity options of the feature selection matrix $\boldsymbol{\gamma}$, and examine which setting results in a more accurate estimation. The sparsity is determined by τ_0, the range of which is set in Table 4.1. Results from the above simulation show that the model correctly learns factors used to generate the data and infers the mixing combination well. We use ROC curve to compare the estimated binary pathway selection matrix $\boldsymbol{\gamma}$ with its ground truth. Figure 4.1(a) shows that given the condition where most pathways are represented in the datasets ($\tau_0 = 0.7$), our model can recover the pathway-dataset association accurately with

averaged AUC>90% for experiments of different sizes. As shown in Table 4.1, results obtained from 70 samples and 250 samples are quite similar (mean AUC=0.99189 and 0.99137, respectively), but the smallest dataset (sized at 10 samples) yields a slightly lower accuracy (mean AUC=0.96741). As the feature selection matrix $\boldsymbol{\gamma}$ gets more sparse ($\tau_0 = 0.3$), the impact of sample size becomes more obvious. As shown in Figure 4.1(b), our model is still capable of recovering most of factors, but we find (i) the mean AUC for each dataset drops, (ii) there is a slight decrease in statistical power and (iii) there is an increased type-I error rate (Figure 4.2 and Table 4.1). We also compare the model estimates with the ground truth for every other parameter. The probabilities of true values falling into 95% credible intervals of their corresponding estimates are $Pr(\boldsymbol{\mu} \in [\hat{\boldsymbol{\mu}}_{\frac{\alpha}{2}}, \hat{\boldsymbol{\mu}}_{1-\frac{\alpha}{2}}]) = 0.92, Pr(\boldsymbol{\theta} \in [\hat{\boldsymbol{\theta}}_{\frac{\alpha}{2}}, \hat{\boldsymbol{\theta}}_{1-\frac{\alpha}{2}}]) = 0.92$ and $Pr(\boldsymbol{a} \in [\hat{\boldsymbol{a}}_{\frac{\alpha}{2}}, \hat{\boldsymbol{a}}_{1-\frac{\alpha}{2}}]) = 0.94$, where $\alpha = 0.05$. We normalized $\hat{\boldsymbol{a}}$ before comparing with its ground truth, because the scale is subject to change in loadings, and the real value is sampled from a standard normal distribution. The comparison results are similar for both simulations with different sparsity settings.

Impure pathways

Another scenario is to simulate partial pathways/gene sets. Since not all the genes in a pathway will be turned on/off simultaneously, some genes are rubbish thus not contributing to the overall signal. In light of this, we randomly selected 40% from the original 219 pathways to which we assign a varying number of noise genes. We create noise genes by setting the loading, $a_{g,k} \sim N(a_{g,k}; 0, 0.01)$, thus ensuring that the expression of the corresponding gene consists almost entirely of idiosyncratic noise. The proportion of garbage genes for any particular pathway ranges from 10% to 90% in 10% increments. We use $\tau_0 = 0.7$ with all other settings remaining as above. Classification result on $\boldsymbol{\gamma}$ is quite accurate with averaged AUC above 90% for data sets larger than or equal to 20 samples, if the ratio of noise genes is between

10% and 70%. As shown in Figure 4.3(a), the larger the sample size is, the more accurate the classification becomes. While the smallest data (sized at 10) performs less accurately (averaged AUC below 90%). The AUC curve fluctuates a little and drops below 0.9 if 80-90% of genes in the pathways are useless. Figure 4.3(b) and 4.3(c) further examines these results through comparisons of statistical power and type-I error rates, respectively. For those 'garbage' pathways correctly identified as relevant to the datasets, their loadings from point estimate successfully shrink the 'rubbish' genes towards zero.

4.4 Experiment: Ionizing radiation

4.4.1 Data collection

A total of 355 C57BL/6 mice peripheral blood gene expression profiles were measured in 3 different batches using the Affymetrix mouse 430A 2.0 microarray. The experiment is designed to assess blood gene expression changes after exposure to ionizing radiation between 0 and 1050 cGy. Samples were collected at 6, 24, 48, 72, 120 and 168hrs after a single dose exposure. In order to place resulting gene expression changes in context, we also incorporate 7 additional murine datasets from GEO (Parkitna et al., 2010; Wang et al., 2010; Jaffe et al., 2010; Beier et al., 2011; De Zoeten et al., 2011; Habermehl et al., 2011). These samples all came from the same mouse strain and microarray platform but were treated under different conditions. A detailed overview of each dataset can be found in Table 4.2. We use pathways from the knowledgebase that are pertinent to radiation and blood cells from GSEA MSigDB. A total of 222 pathways, including C2 (curated) and C3 (motif) gene sets, were selected. These contain 10571 probes belonging to the mice Affy array, as shown in Figure 4.4.

4.4.2 Data analysis

We conduct 10000 MCMC iterations with 6000 burn-in samples. The hyperparameters are set to $\{\alpha_0, \kappa_0, e_0, l_0, u_0, n_0, c_0, d_0, a_0, b_0, m_0, s_0\} = \{1, 10, 10, 90, 0, 1, 1.1, 0.01, 1.1, 1, 8, 2\}$. Pathway selection is inferred by computing the posterior odds in favor of $\gamma_{k,d,\cdot} = 1$. The model selects factors using posterior probability threshold greater than 0.5. This results in 75 factors shared across the 3 radiation batches, of which 15 are also found to be associated with gene expression changes in some subset of the GEO sets (Figure 4.5).

Clustering of pathway responsive patterns induced by radiation

Because of the high degree of co-expression between many genes in this experiment as well as the experiments downloaded from GEO, we find that there are numerous cases in which groups of pathways share a common expression pattern across samples. A tree structure (Henao et al., 2013) was used to organize the identified pathways into groups based on this shared expression. In order to examine pathway activity among different cell lines, we display only the clustering results from 26 gene sets of such features (Figure 4.6). Four groups of pathways are thus obtained.

Cluster 1. Figure 4.7 shows a generalized pattern of repression in a dose-dependent manner. Specifically, expression increases from 6 to 168 hrs at high dose levels (600 $\sim$ 1050 cGy), while for low-dose exposure (200 $\sim$ 450 cGy) it increments to baseline at 72 hrs but decreases afterwards, and it shows faster recovery rate at a lower dose 100 cGy, and reaches baseline at 24 hrs.

Pathways in this group include gene sets down regulated in fibroblast cell lines after high dose UV-C exposure (Gentile et al., 2003) (factor 64), and those up-regulated in common lymphoid progenitor (CLP) cells (Han et al., 2012) (factor 98). CLP cells can be further differentiated into B cells, the process of which was reported to be affected by radiation in bone marrow (Han et al., 2012).

Cluster 2. Figure 4.8 shows the behavior of genes in pathways from this cluster. Expression is slightly repressed at 6 hrs and maintaining an induction pattern from 48 to 168 hrs with high dose radiation treatment, in contrast, low dose exposure suppresses gene expression from 120 to 168 hrs.

Pathways in this group are associated with regulation of cell death (Smirnov et al., 2012) (factor 190), those specific to mast cells (Nakajima et al., 2001) (factor 130) and hematopoietic stem cells (HSC) (Georgantas et al., 2004) (factor 70). Proliferation of mast cells has been shown to be affected by ultraviolet or infrared radiation (Kim et al., 2009); It is possible that ionizing radiation triggers a similar mechanism. HSC are considered to be somewhat resistant to radiation, and may be important for regeneration of the hematopoietic system after radiation damage. The pathway reported in (Sesto et al., 2002) (factor 185) whose expression changed in human primary keratinocytes by UVB irradiation is also found to be activated in our experiment. Examples of this group are RRAS (gene related to RAS viral (r-ras) oncogene homolog) and GADD45A (gene involved in growth arrest and induced in response to DNA damage).

Cluster 3. Figure 4.9 shows that, for genes in this cluster, radiation induces up-regulation from 6 to 120 hrs with a maximum at 48 or 72 hrs, which then return to basal levels at 168 hrs except that high doses further repress the expression. One element of this cluster is intrathymic T progenitor (ITTP) (Lee et al., 2004) (factor 116). It is the earliest progenitor; genes in this pathway show a different expression pattern after radiation stimulation, compared to the ones associated with more matured T thymocytes (cluster 4).

Cluster 4. Figure 4.10 shows that this group of pathways are slightly up-regulated at 6 hrs and keep a repression pattern from 48 to 168 hrs after high level radiation exposure, whereas low dose exposure induces up-regulation from 120 to 168 hrs. Pathways in this group include epigenomic biomarkers in acute lymphoblastic leukemia

(Taylor et al., 2007) (factor 195) and acute promyelocytic leukemia (Nouzova et al., 2004) (factor 138), suggesting novel targets for radiotherapy monitoring. Gene sets enriched in both double polar (DP) (factor 113) and single positive 4 (SP4) thymocytes (factor 119) show this pattern of expression changes in our data. Many DP genes are found to be involved in cell cycle progression or proliferation and are related to thymocyte differentiation into T/B/NK cells. SP4 enriched genes are known to be important for T cell functions, such as inhibition of T cell apoptosis, regulation of T cell homeostasis and T cell differentiation (Lee et al., 2004). Given this evidence as well as our discovery of depressed gene expression activities in DP and SP4 thymocytes, we theorize that radiation, especially at high dose, may disturb cell cycle and affect cell renewal and differentiation.

Identification of radiation-induced pathways related to glucocorticoid receptor ablation experiment in GEO

Glucocorticoid (GC) is a type of steroid hormone, involved in lung maturation and used as treatment for cancers and cardiovascular diseases. GC causes its effect through binding to the Glucocorticoid receptor (GR). In GSE30143, murine GR-ablated mesenchymal cells were investigated to recapitulate GR activities. The gene expression profile was obtained for mutants and controls at both embryonic day 16.5 (E16.5) and day 18.5 (E18.5), a progression period during murine lung development (Habermehl et al., 2011). Our model identifies several factors associated with both this data set and the radiation sets. Examples include Biocarta Acute Myocardial Infarction (AMI) pathway (factor 10), gene set targeted by ATM (Ataxia Telangiectasia Mutated gene) regulation (Rashi-Elkeles et al., 2005) (factor 149) and genes expressing a constitutively active form of STAT5 in HSC (Schuringa et al., 2004) (factor 178). These pathways are significantly differentially regulated in controls from E16.5 $\sim$ E18.5 with a P value <0.0001 (t test), which is consistent with the

original experiment.

The transition phase from E16.5 to E18.5 coincides with increased pulmonary GR expression (Habermehl et al., 2011), which is reflected in our results that factor 10 and 178 are up-regulated in E18.5 while down-regulated in E16.5 (Figure 4.11(a) and 4.11(b)). GR over-expression has an aggravation effect during rats' myocardial infarction (Mihailidou et al., 2009). The disease risk greatly increases while exposed with irradiation, the hazard of which is believed to potentiate the negative effects on pathogenesis of AMI (Karpov et al., 2012). Because genes in this pathway react to both radiation exposure and GR-ablation, we hypothesize that the increased risk of AMI post radiation exposure could be related to GR activation. We also identify from factor 149 that a gene encoding serum/glucocorticoid regulated kinase, SGK1, is activated at E18.5 while repressed at E16.5 (Figure 4.11(c)). In contrast, the Human papillomaviruses (HPV) positive cervical cancer gene signatures (Pyeon et al., 2007) (factor 146) is associated with GR in a reversed expression pattern (Figure 4.11(d)). GR is often used for cervical cancer treatment as an inhibitor of radiotherapy-induced apoptosis (Buxant et al., 2009).

4.4.3 Comparison with existing approaches

To further examine the performance of our model, we compare the pathway selection results obtained from our model with the results generated using GSEA and GSVA. These are two gene-set-enrichment-score-based methods evaluating (1) pathway activities between two biological states if the phenotypic information is available (GSEA), or (2) evaluating pathway variances across samples without the use of a phenotype (GSVA) (Subramanian et al., 2005; Hanzelmann et al., 2013). We also compare to DAVID, an online bioinformatics resource providing functional annotation of large lists of genes derived from different genomic studies (Dennis et al., 2003). As in our approach, all these algorithms use public knowledgebases of *a priori*

defined gene sets or pathways.

Since none of the three methods provide integrated analysis, we use only one radiation batch (244 samples) for comparison. We keep all the data as original (455 samples) for our Bayesian joint factor model. We utilize a pathway knowledgebase constructed using KEGG (Kyoto Encyclopedia of Genes and Genomes), because all these analyses have their own gene signature resources, but KEGG is used by all. We filter the KEGG mouse pathways to include gene sets containing 10 to 500 genes, and eliminate KEGG genes which do not have an Affy probe ID. A total of 211 KEGG pathways are used with 8035 probes belonging to mice Affymetrix 430A 2.0 array. We next use these probes to initiate DAVID queries, and create zero versus non-zero radiation dosimetry as binary phenotypic label for GSEA and GSVA.

Figure 4.12 exhibits numbers of pathways identified by each method. Our model selects 26 pathways based on posterior estimation, GSEA finds 21 gene sets enriched at FDR q-value<0.25, while DAVID identifies 103 significantly enriched gene sets with Fisher Exact p-value<0.05. We set an adjusted p-value<0.05 for GSVA enrichment scores, which leads to 141 pathways differentially activated between non-irradiation and irradiated groups. Since the experiment was designed to measure recovery from radiation exposure using peripheral blood samples, we expect to detect hematopoiesis related signaling pathways. Some of these pathways may be dysregulated due to radiation damage, and others activated as a recovery response. Among the 26 pathways our model identifies, 5 represent pathways known to be associated with radiation exposure.

Examples include MAPK (Mitogen-Activated-Protein-Kinase) signaling pathway, which was shown to be activated by ionizing radiation or other cellular stresses, and in turn regulates cell cycle progression, apoptosis induction and differentiation (Chung and Kondo, 2011). This is consistent with what DAVID finds (ranked in the 2nd place, Bonferroni adjusted p-val=4.49E-17), and is undiscovered by GSEA and

GSVA.

Another well-known example, the Wnt signaling pathway, is also identified by our model as well as top selected by DAVID (ranked in the 4nd place, Bonferroni adjusted p-val=2.96E-09) but not by GSEA or GSVA. Wnt pathway is involved in the maintenance of normal HSC functions, with wnt protein localized in the blood cells regulating proliferation, differentiation and survival of HSC (Wilusz and Majka, 2008). Our model also identifies two important genes in this pathway based on their loadings, WNT2 and GSK3B, respectively. The former is a negative regulator for hematopoiesis, and the latter involved in the maintenance of HSC function (Huang et al., 2009b; Wilusz and Majka, 2008). Radiation exposure is likely to ablate rapidly cycling cells and induce loss of blood cells. In response to such injury, HSC activate intracellular signals to initiate proliferation that ultimately leads to hematopoiesis. During HSC regeneration, Wnt pathway is activated to enhance HSC regrowth. It has been shown that modulating Wnt pathway may be an effective therapeutic strategy to accelerate recovery after injury (Congdon et al., 2008).

A third example, the hedgehog signaling pathway, is activated to induce cycling and expansion of primitive HSC during acute regeneration. This contributes to the maintenance of the integrity of the system (Trowbridge et al., 2006). It is again recovered by our model as well as ranked 22nd in DAVID (Bonferroni adjusted p-val=1.19E-03), but not shown in GSEA and GSVA.

We uncover up-regulated PPAR (peroxisome proliferator-activated receptor) pathway and a cell death related pathway-natural killer cell mediated by cytotoxicity-with p-val at 6.36E-05 and 1.20E-06, respectively, based on comparison of factor scores between irradiated and base-line samples. They are less significant in DAVID (PPAR: Bonferroni adjusted p-val=0.441; natural killer cell mediated by cytotoxicity: Bonferroni adjusted p-val=0.560), and GSEA and GSVA failed to identify either. The factor model finds three unique pathways that are not detected by other methods,

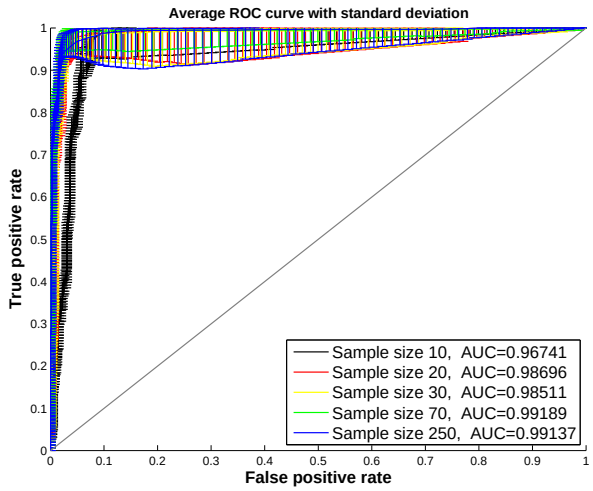
including protein processing in endoplasmic reticulum, protein digestion and absorption and drug metabolism, which suggests a potential function during hematopoietic repair after radiation stimulation.

Two moderately significant pathways (p-val<0.05) identified by DAVID and GSVA are not covered by factor model or GSEA (shown in Figure 4.13).The two examples are (1) Jak/STAT signaling pathway which is involved in suppression of HSC survival, induction of apoptosis and inhibition of leukemic growth, and (2) VEGF (vascular endothelia growth factor), which participates in the formation of blood cells and controls HSC survival and repopulation (McCubrey et al., 2008; Gerber and Ferrara, 2003). GSEA and GSVA both uncover several DNA repair related pathways, such as mismatch repair, nucleotide excision repair, homologous recombination, non-homologous end-joining and notch signaling pathway. Ionizing radiation does induce DNA damage, leading to activation of the DNA repair pathways. However, only two genes in these pathways show differential expression greater than two-fold between non-irradiated and irradiated samples, as indicated in Figure 4.14. These are CTBP2 (fold-change: 2.7327, p-val: 1.93E-14) and APH1A (fold-change: 2.0179, p-val: 6.41E-15), both in the NOTCH signaling pathway. This pathway is believed to exert multiple important functions in the hematopoietic system, ranging from supporting the first definitive hematopoiesis during fetal life, all the way to maintaining the HSC in the adult bone marrow and regulating differentiation of matured hematopoietic cells in the immune system (Sandy et al., 2012).

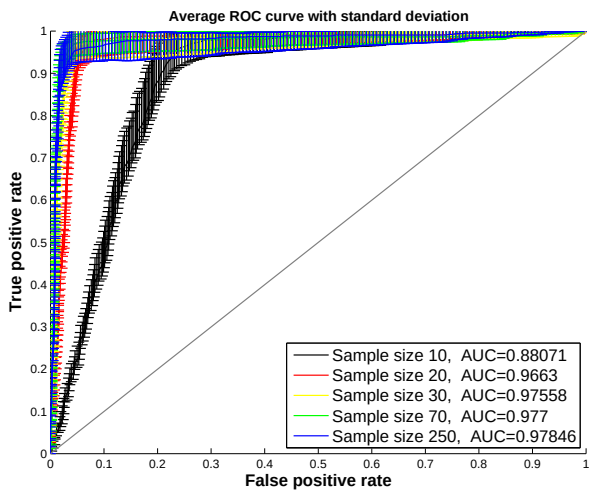
There are only three common pathways uncovered by all methods, including DNA replication, Spliceosome and Ribosome. These pathways are involved in DNA duplications, post-transcription splicing and translations, respectively, which are related to cell proliferations, and thus aid hematopoietic system recovery after radiation damage.

4.5 Conclusions

We have presented a hierarchical model that can identify similar pathway-level gene co-expression patterns across diverse experimental states, and uncover pathway induction/repression for each dataset. This method provides an unsupervised framework for pathway selection and data integration without the use of explicit phenotypic comparisons. Our approach allows the incorporation of *a priori* information on gene-pathway association, which enhances interpretability for each factor. Compared with existing gene-set annotation approaches, our model is based on observed co-regulation across potentially many data sets; Based on comparison studies our approach is somewhat synergistic with other published approaches. Posterior parameters obtained from the model as a summary of the overall pathway activities can be applied for classification, survival regression and clustering. Finally, the radiation example demonstrates a successful application of our model. We expect that the joint Bayesian factor model will be a broadly applicable approach to leveraging the growing body of available transcriptional profiles and generate hypotheses in a gene expression data-driven, pathway centric manner.

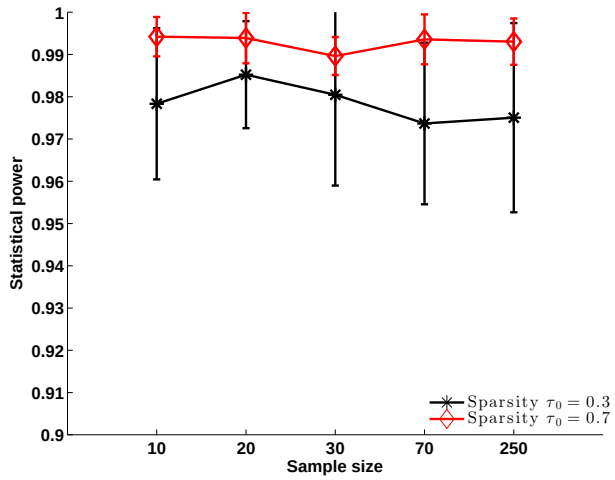


(a)

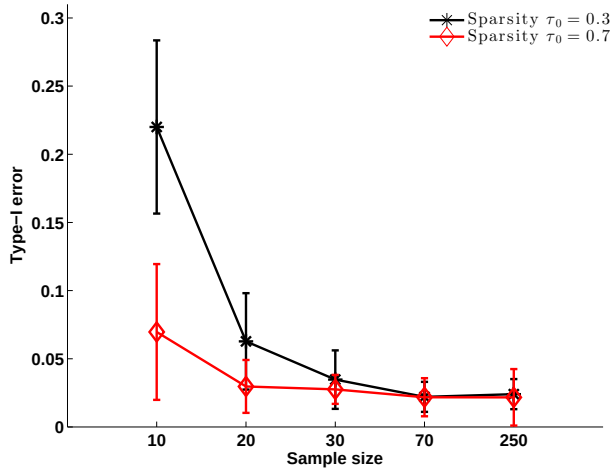


(b)

FIGURE 4.1: Simulation results: comparison of estimated feature selection matrix $\boldsymbol{\gamma}$ with ground truth. The two plots are averaged ROC curves with standard deviations (error bars) obtained from 10 simulations. AUC here indicates the averaged area under curve. Each color represents different data set with varied sample size. Figure 4.1(a) is the result obtained with sparsity $\tau_0 = 0.7$, while figure 4.1(b) is the one with $\tau_0 = 0.3$.

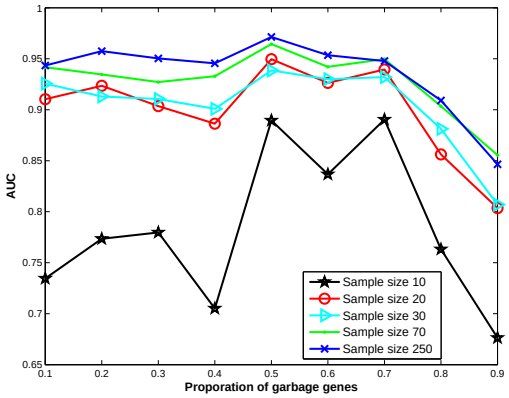


(a)

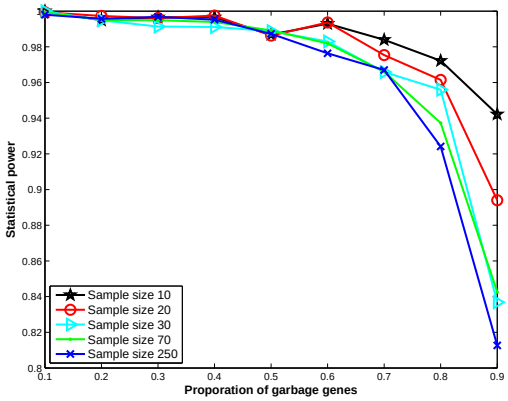


(b)

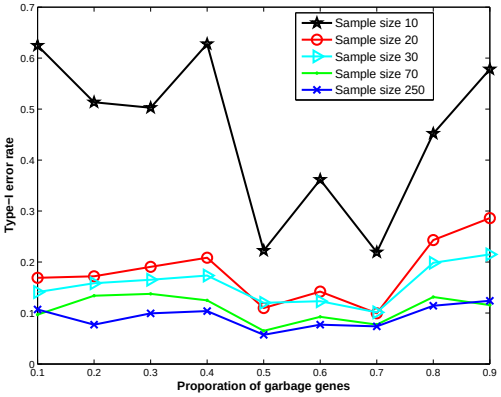
FIGURE 4.2: Simulation results: the statistical power and type-I error rate obtained from different sparsity settings in the feature selection matrix $\boldsymbol{\gamma}$. The averaged results of 10 simulations with standard deviation are plotted across different sample sizes on the x axis. Y axis displays either the statistical power (4.2(a)) or the type-I error rate (4.2(b)). Black star symbol indicates results obtained from sparsity of feature selection matrix set as 0.3, while red diamond shows the one with a less sparse setting with $\tau_0 = 0.7$.



(a)



(b)



(c)

FIGURE 4.3: Simulation results: comparing estimated feature selection matrix $\boldsymbol{\gamma}$ with ground truth. X-axis displays different percentages of garbage genes synthesized in the pathways. Y-axis displays averaged area under curve (4.3(a)) from 10 simulations, mean statistical power (4.3(b)) and mean type-I error rate (4.3(c)), respectively. Each colored line represents a dataset with different sample sizes marked by distinct symbols.

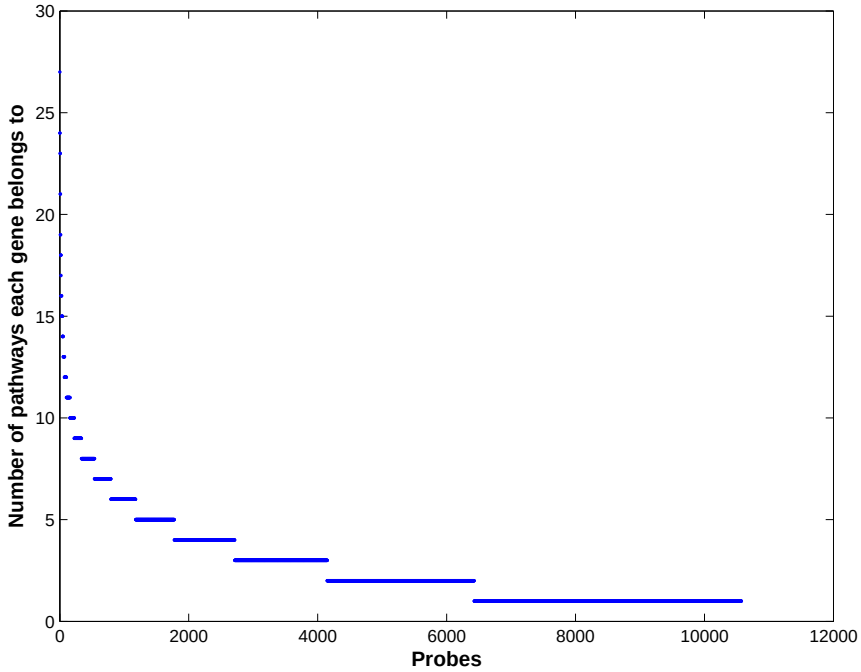


FIGURE 4.4: Plot of overlaps among pathways. 39.2% genes are unique to specific pathways, 49.7% are shared by less than or equal to five pathways and only 11.1% are shared among more than five pathways with the maximum common to twenty-seven different pathways. The y-axis shows number of pathways shared by each Affy probe, and x-axis is the indicator of probes sorted by the number of overlaps in decreasing order.

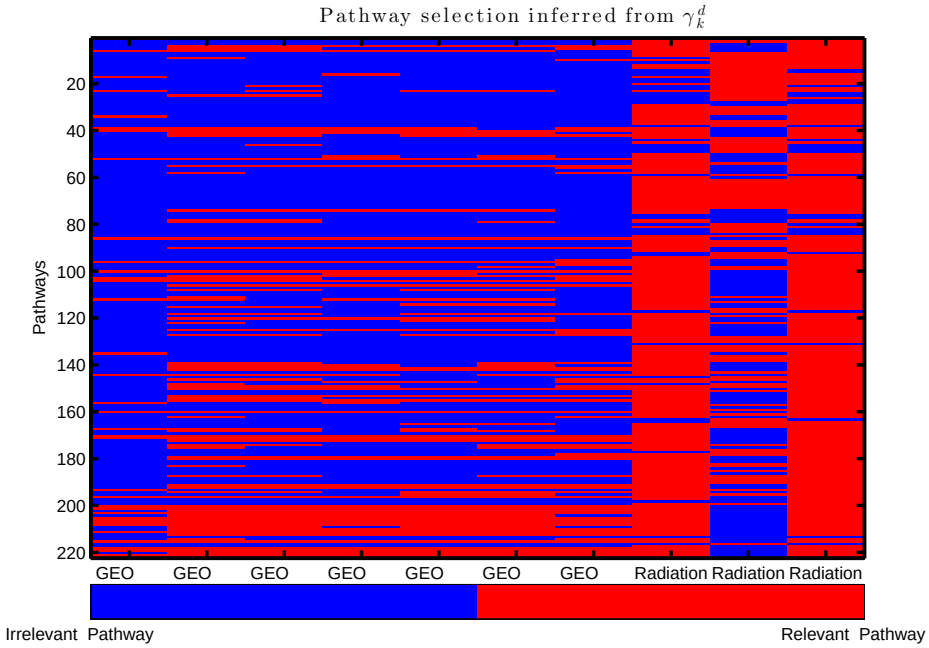


FIGURE 4.5: Heat map of the estimated feature selection matrix. The sparse binary matrix shows factors unique to specific data set and common among them. The y-axis represents indicator of each factor, i.e., pathways in our case. The x-axis represents the 10 datasets with the same order from Table 4.2. The inferred factors are assigned as 1(red), otherwise 0(blue).

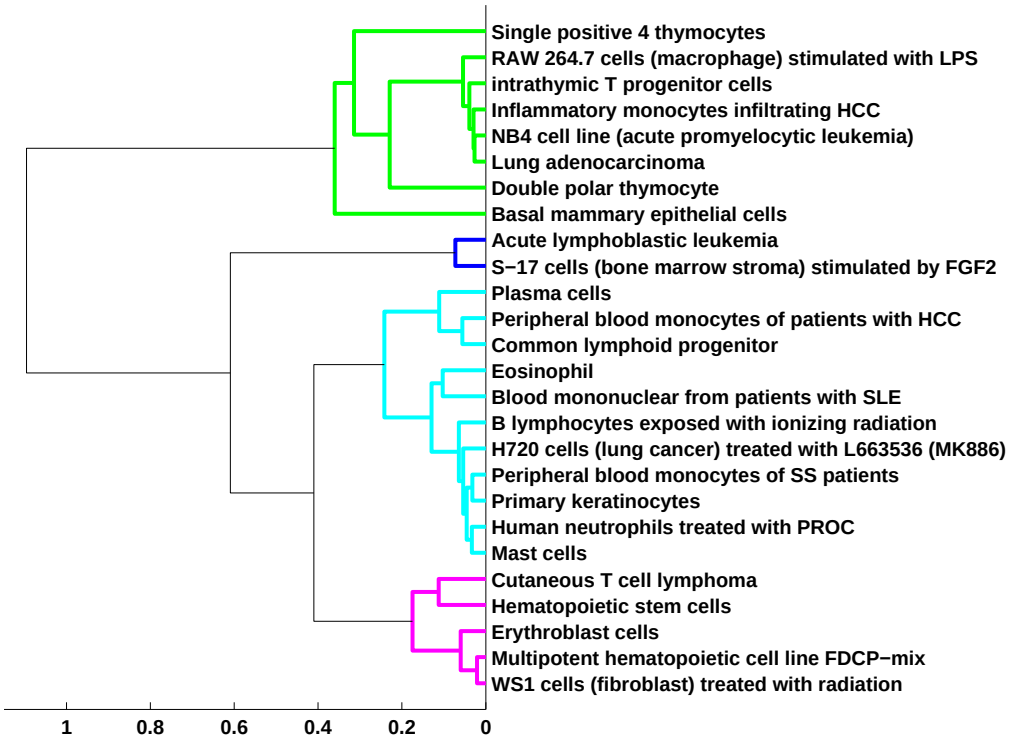


FIGURE 4.6: Dendrogram of irradiation-induced pathways. 26 gene sets that changed their expression are grouped into four clusters based on their factor scores. Each cluster is labeled with a different color. From bottom to top, magenta represents cluster 1 with five sets, cyan is cluster 2 containing eleven sets, blue for cluster 3 with two sets and cluster 4 (green) includes eight leaves. Gene set names are marked on the y-axis, each described as a cell type that different genetic or chemical perturbations triggers a distinct expression. The x-axis represents the height of each U-shaped line, which is the distance between two data points being connected. A smaller value means a closer link.

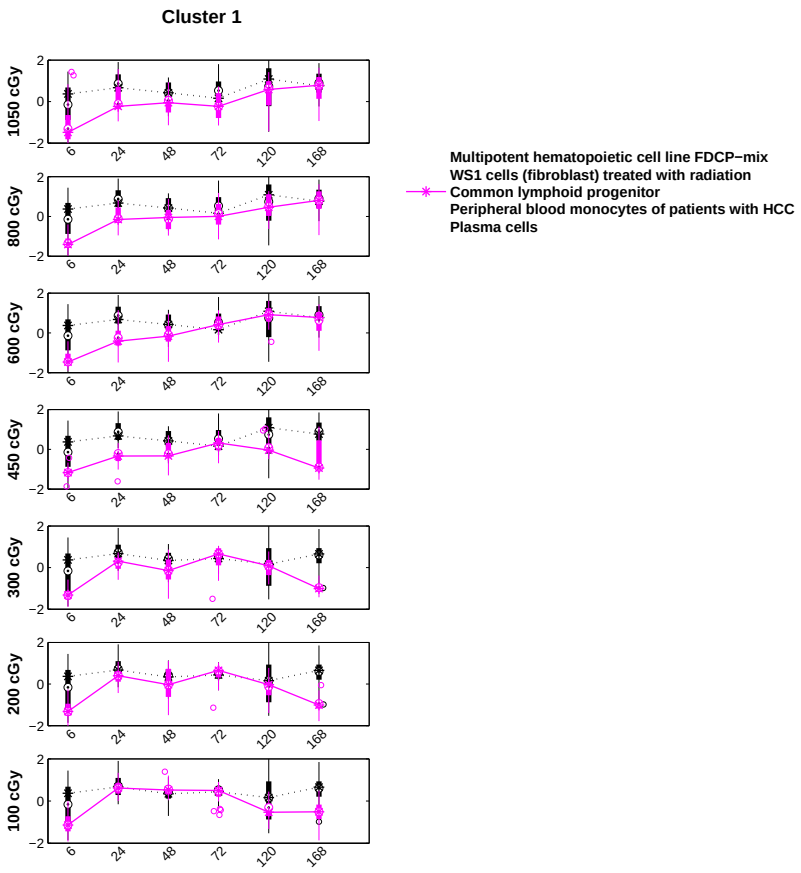


FIGURE 4.7: Cluster 1: pathways responsive to irradiation. Each sub figure is a box plot of five factors across 6 time points with solid line representing radiation treatment and dashed line non-irradiation. Different radiation dose levels are displayed from top to bottom in a decreasing order.

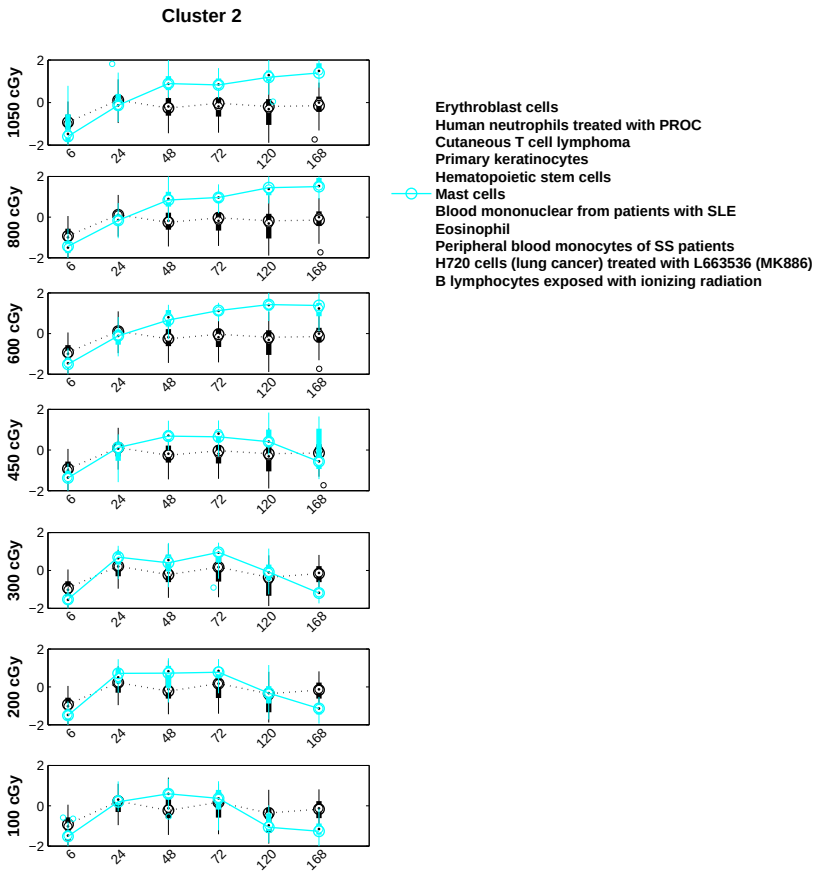


FIGURE 4.8: Cluster 2: pathways responsive to irradiation. Each sub figure is a box plot of eleven factors across 6 time points with solid line representing radiation treatment and dashed line non-irradiation. Different radiation dose levels are displayed from top to bottom in a decreasing order.

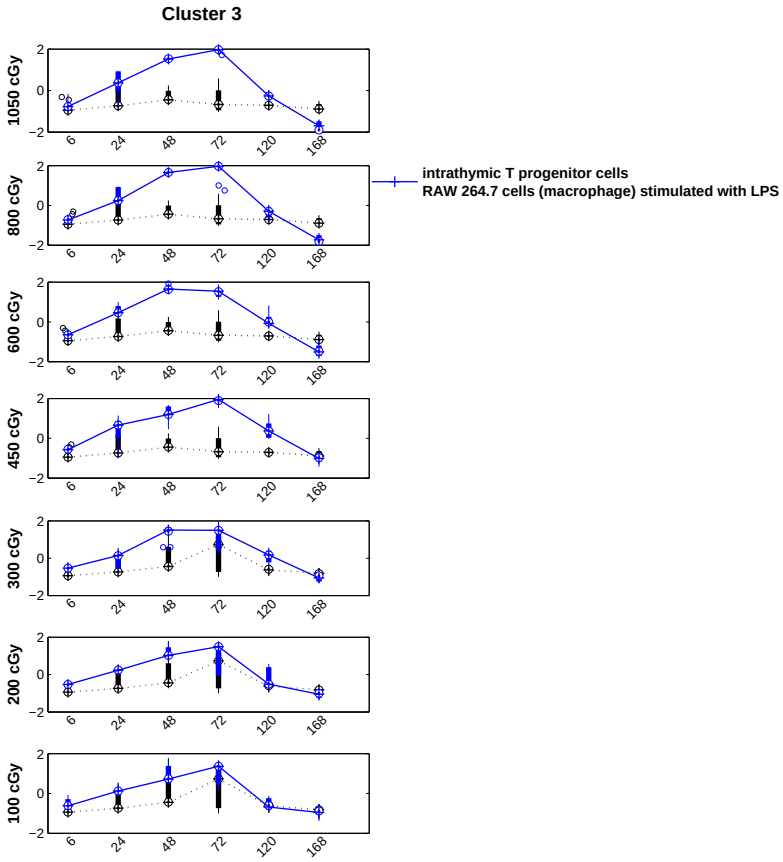


FIGURE 4.9: Cluster 3: pathways responsive to irradiation. Each sub figure is a box plot of two factors across 6 time points with solid line representing radiation treatment and dashed line non-irradiation. Different radiation dose levels are displayed from top to bottom in a decreasing order.

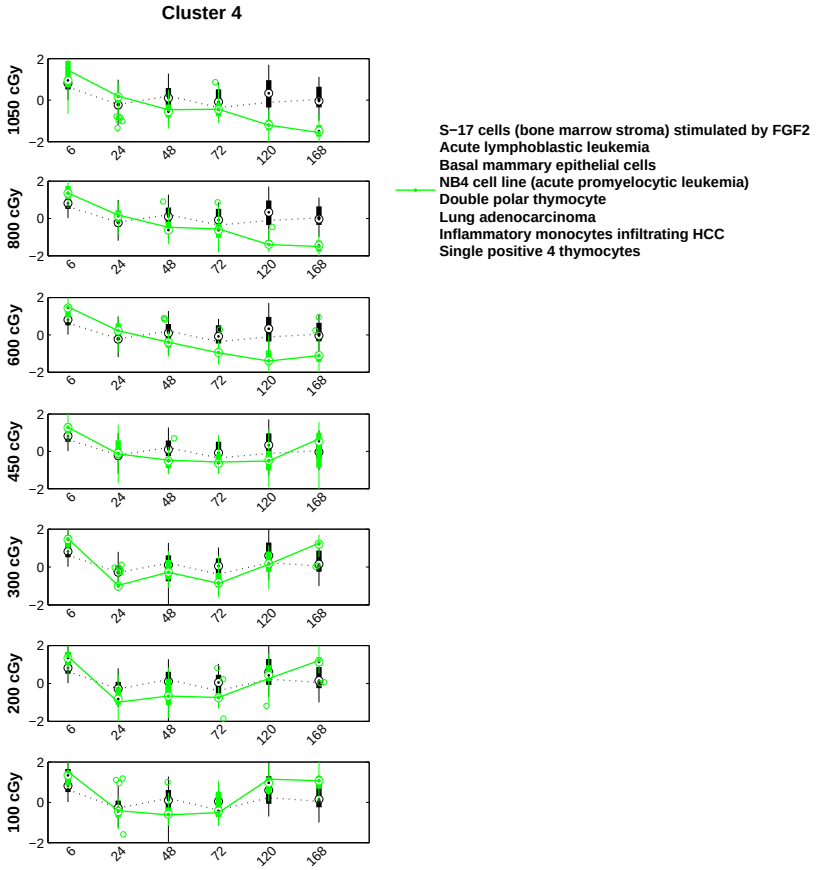


FIGURE 4.10: Cluster 4: pathways responsive to irradiation. Each sub figure is a box plot of eight factors across 6 time points with solid line representing radiation treatment and dashed line non-irradiation. Different radiation dose levels are displayed from top to bottom in a decreasing order.

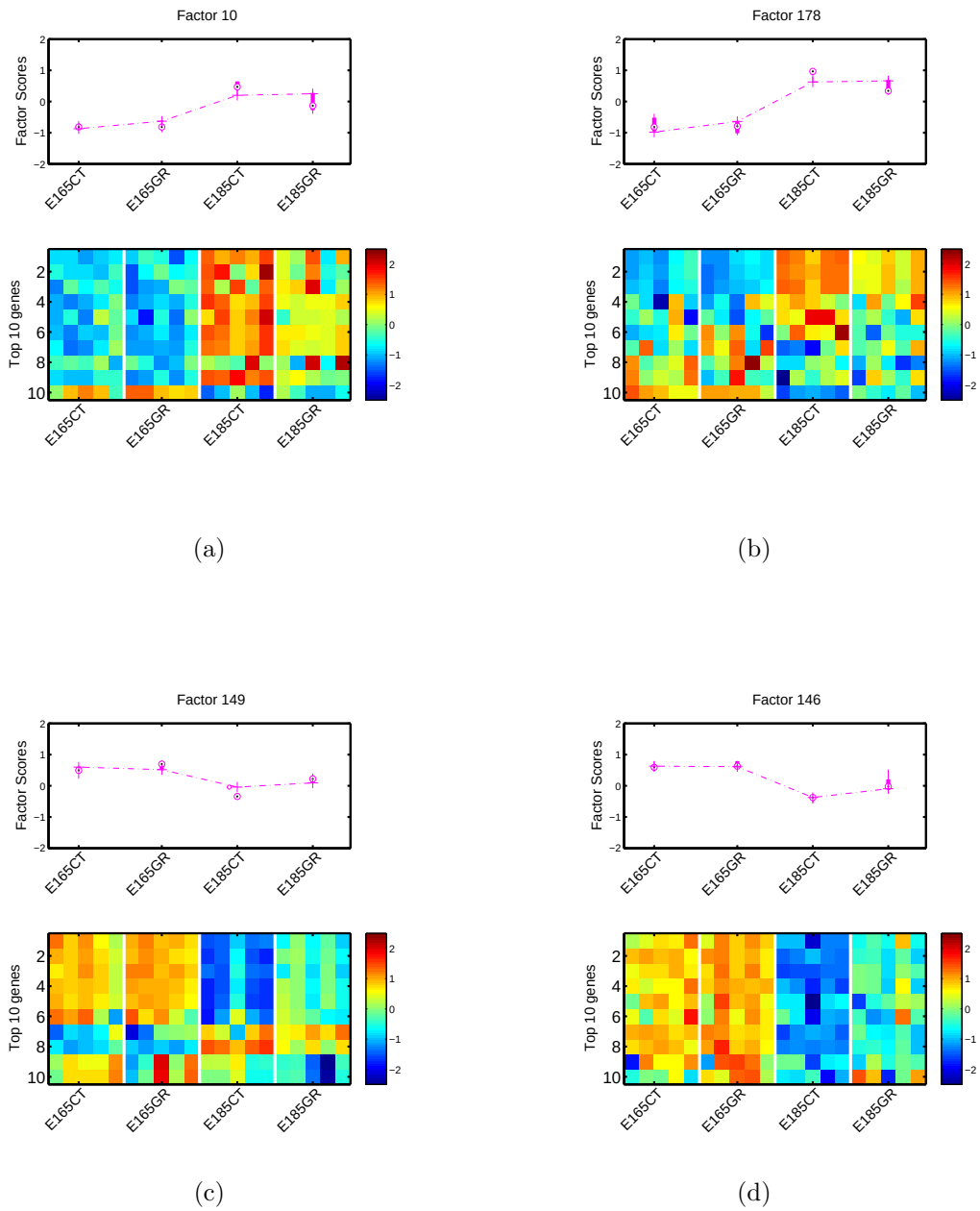


FIGURE 4.11: Gene expression patterns of four pathways summarized by factor scores. Four pathways or factors are identified to be associated with GEO series GSE30143. The top box plot of each subfigure represents the overall expression pattern. Each heat map shows the top 10 genes from every factors ranked by the signal-to-noise ratio. The P value (t test) of the difference between E16.5 and E18.5 animals is 4.76e-6, 9.58e-6, 4.42e-5 and 1.32e-9, from (a) to (d) respectively. CT indicates control groups and GR indicates mutant animals.

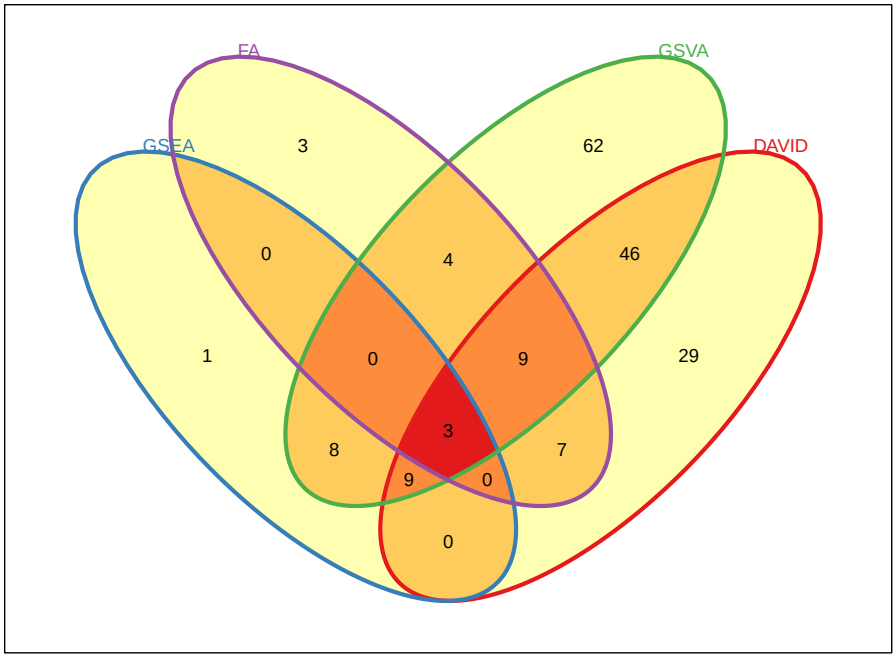
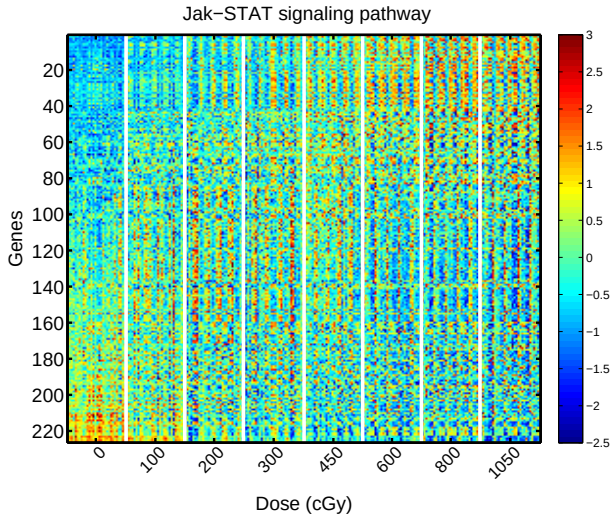
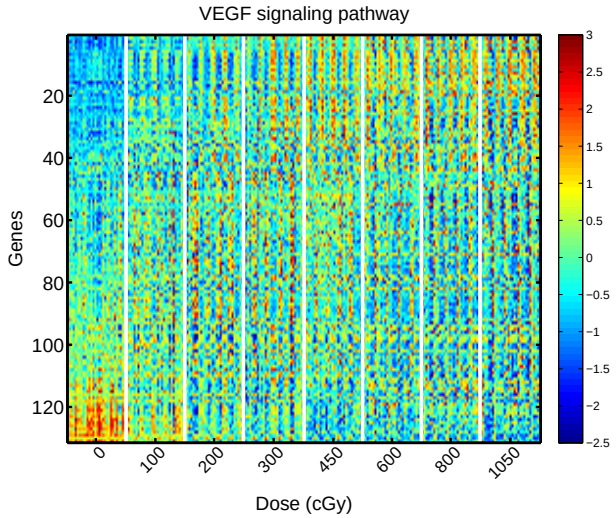


FIGURE 4.12: Comparison of four different methods is illustrated in a Venn diagram. Each colored ellipse represents one particular approach, i.e., FA (colored as plum) is our factor analysis model, GSEA (blue) is gene set enrichment analysis, GSVA (green) is gene set variation analysis and DAVID (red) is the web-based DAVID bioinformatics tool. Numbers within ellipse are number of pathways identified by each method and their overlaps.

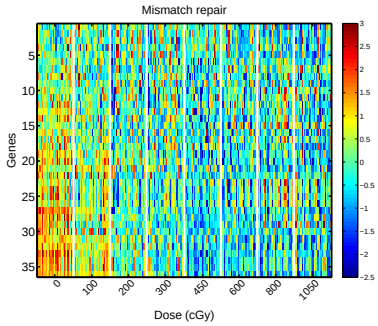


(a)

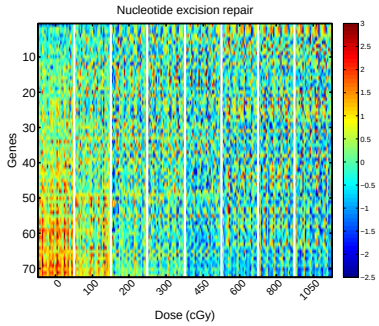


(b)

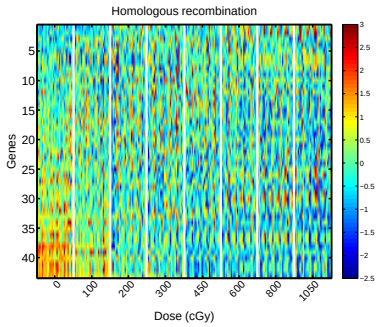
FIGURE 4.13: Pathways commonly selected by DAVID and GSVA. Y-axis displays gene probes in each pathway, sorted by fold-changes between non-irradiated (0 cGy) and irradiated samples (>0 cGy). Pathway (a) shows the 226 gene probes in Jak-STAT signaling pathway, and (b) contains 131 probes. Pixels in the image represents standardized expression intensity with positive values indicating over-expression and vise versa.



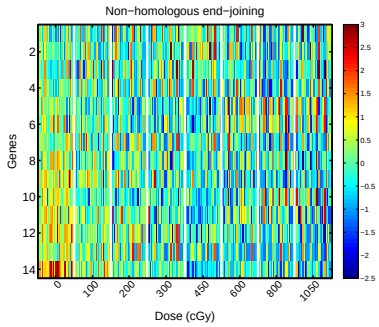
(a)



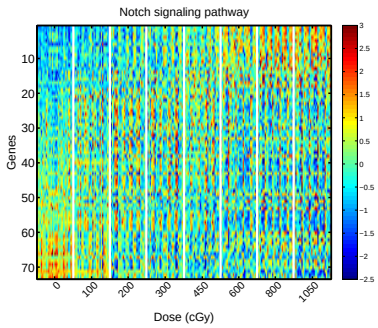
(b)



(c)



(d)



(e)

FIGURE 4.14: Five DNA repair pathways commonly selected by GSEA and GSVA. Y-axis displays numbers of gene probes in each pathway, sorted by fold-changes between non-irradiated (0 cGy) and irradiated samples (>0 cGy). Pixels in the image represents standardized expression intensity with positive values indicating over-expression and vise versa.

Table 4.1: Simulation results from different sparsity settings (τ_0) in feature selection matrix $\boldsymbol{\gamma}$. 1 indicates results obtained from $\tau_0 = 0.7$, 2 represents those using $\tau_0 = 0.3$, and Δ is the difference between the above two. $AUC^{\Delta} = -\frac{AUC^1 - AUC^2}{AUC^1}, Power^{\Delta} = -\frac{Power^1 - Power^2}{Power^1}, Type-I-error-rate^{\Delta} = \frac{Type-I-error-rate^2 - Type-I-error-rate^1}{Type-I-error-rate^2}$. A positive Δ value means an increased percentage when τ_0 diminishes from 0.7 to 0.3, while a negative value indicates a decreased percentage. AUC, power and type-I error rate are averaged values calculated after 10 simulations.

	$\tau_0 = 0.7$			$\tau_0 = 0.3$					
size	AUC^1	Power^1	Type-I-error-rate^1	AUC^2	Power^2	Type-I-error-rate^2	AUC$^{\Delta}$	Power$^{\Delta}$	Type-I-error-rate$^{\Delta}$
10	0.96741	0.99420	0.06970	0.88071	0.97830	0.22000	-8.96190%	-1.59650%	68.34289%
20	0.98696	0.99390	0.02970	0.96630	0.98520	0.06270	-2.09270%	-0.87350%	52.66600%
30	0.98511	0.98960	0.02750	0.97558	0.98050	0.03470	-0.96770%	-0.92520%	20.66460%
70	0.99189	0.99360	0.02180	0.97700	0.97370	0.02200	-1.50050%	-2.00360%	1.20310%
250	0.99137	0.99300	0.02170	0.97846	0.97510	0.02400	-1.30210%	-1.81140%	9.70020%

Table 4.2: Summary of data sets used in our analysis. Each set is illustrated with a GEO accession number if it's a GEO experiment or batch ID if it's our data, number of samples and a brief explanation of the experiment.

Dataset	Sample Size	Experimental Samples
GSE10870	12	striatum of mice brain with Srf deletion and WT treated with cocaine or salin
GSE18483	20	murine leukemias originated from hematopoietic stem (HSC) or committed progenitors cells
GSE20377	15	leukemia stem cells, HSC
GSE23566	18	mouse aorta treated with vehicle or aldosterone
GSE26425	9	T-regulatory cells(Treg) of Sirt1 KO and WT
GSE27896	6	Treg of HDAC6 KO and WT
GSE30143	20	lungs of mice with mesenchyme-specific GR ablation between 16 and 18 day of embryonal development
Batch 2202	35	mice peripheral blood treated with irradiation
Batch 2665	242	mice peripheral blood treated with irradiation
Batch 2666	78	mice peripheral blood treated with irradiation

5

Concluding Remarks and Future Directions

5.1 Summary

This dissertation presents novel methodologies using regulatory modules derived from genome-wide gene expression profiles as a media, which thereby integrate different types of genomic data and provide alternative ways of hypothesis generation. The major contribution is the development of several innovative Bayesian statistical models and computational methods for the co-expression-module-centric problems arising in pathway annotation, cancer, and other diseases. Application of these models is demonstrated in a series of examples and studies, including deciphering the root cause of ovarian cancer from genetic and epigenetic perspectives, and identifying pathway responsive patterns in mice hematopoietic system after ionizing radiation exposure. The work combines advanced statistical models and computational techniques with successful applications in various biological systems. Although developed and discussed in the context of mostly cancer research, these statistical models are broadly applicable to any other biological studies based on genome-wide gene expression data and extendable to model other high-dimensional 'Omics' data.

The meta-analysis models developed in this dissertation are used to identify co-regulatory modules in different circumstances: *i)* Vertical integrative analysis combines multi-omics data of the same patient cohort to investigate driver genes and regulatory networks, even if these driver mutations exist only in a subgroup of samples. *ii)* Horizontal integration approach combines multiple same-type genomic data (e.g., gene expression) and pathway knowledge base to identify enriched gene sets under different biological states with increased statistical power. These models can be applied separately to address the above three problems, or used together to achieve 'personalized medicine'. Considering the example of TCGA, we first perform vertical integration to identify driver mutations underlying a specific cancer subtype. We then provide functional annotations for the above genes using the horizontal approach. This method also enables simultaneous comparison with other patients' gene expression data from the same cancer subcategory, which helps to validate the co-regulatory modules in additional samples. It is also possible to combine data from other cancer types to discover the same origins of disease among different tumor tissues.

5.2 Future directions

The accumulation of 'Omics' data provides unique opportunities to uncover subtle distinctions in complex disease phenotypes. While we mainly focus on gene expression, CNV and methylation, there are other types of data (e.g., exon and protein expressions, and somatic mutations in TCGA) that has not been covered yet. Therefore, the next step in this area of research is to continue to incorporate additional sources of 'Omics' information into our models, where further adjustment of these methods is needed.

5.2.1 Jointly modeling mRNA and microRNA expression

Chapters 2 and 3 uncover underlying mechanisms of cancer by tracing changes in mRNA expression back to its genetic/epigenetic cause. We identified several novel genes and regions that might be driver mutations in ovarian cancer. Similarly, another upstream event, microRNA (miRNA), acting as a post-transcriptional regulator for mRNA, can also be incorporated into the same model.

MiRNAs are a class of endogenous, small and highly conserved noncoding RNAs that regulate gene expression by binding to complementary sequences on target mRNA, and influencing mRNA through translational repression or target degradation (Tseng et al., 2011). MiRNAs are believed to play an important role in the development of various cellular processes, such as cancer (Qin, 2008). Dysregulated miRNA expression is characterized in many different cancer types, including carcinomas of the breast, ovary and lung (Eder and Scherr, 2005; Zhang et al., 2006; Creighton et al., 2010). As large-scale expression profiling of miRNA becomes available, we could integrate miRNA with mRNA, which is likely to provide an additional clue towards unveiling the mechanisms of tumorogenesis.

In order to identify miRNA-mRNA comodules, we need to take into account the following aspects: *i)* Anti-correlation between the two data modalities; *ii)* Grouped effect: individual miRNAs have only limited impact on their targets and multiple different miRNAs often work in groups to drastically reduce mRNA transcriptions (Malumbres, 2012); *iii)* Individual effect: each miRNA could target hundreds of genes (Le and Bar-Joseph, 2011); *iv)* Current knowledge of sequence/structure information (Kertesz et al., 2007; Enright et al., 2003; John et al., 2004; Lewis et al., 2005; Coronnello and Benos, 2013) could be utilized to predict miRNA targets, via the prior probability model.

Based on the above information, we could transform the FA model in either

chapter 2 or 3 to incorporate miRNA expression data. In order to capture the negative association between miRNA and mRNA, we place more constraints on the model by introducing positively truncated normal priors on loadings and forcing negative signs on the shared components of miRNA factors. Spike-and-slab prior will be used on miRNA loadings to introduce sparseness, while a different prior can be placed on the probability parameter $\rho_{g,j}$ that controls loadings being non-zero. $\rho_{g,j}$ is forced to be a bernoulli random variable with $\rho_{g,j} \sim Bernoulli(r_{g,j})$; $r_{g,j}$ is the success rate, indicating the probability of assigning g-th miRNA to the j-th factor (i.e., non-zero loading) is $r_{g,j}$. Gene-miRNA association scores obtained from sequence/structure information are used in a logistic function to sample $r_{g,j}$ with $P(r_{g,j}|\mu, s_{g,j}) = \frac{exp(\mu+\tau s_{g,j})}{1+exp(\mu+\tau s_{g,j})}$, where μ and τ are unknown mean and coefficient, and $s_{g,j}$ represents the association scores (Stingo et al., 2010).

5.2.2 FA model for data generated from sequencing-based technology

During the past decades, microarrays have been the most important and widely used approach to characterize the molecular basis of phenotypic variation in biology. But a recent emergence of high-throughput sequencing technology (next-generation sequencing) has provided a powerful alternative. It is also used in transcriptional profiling by sequencing cDNA (RNA-seq), and hence generating millions of short reads. These reads are further mapped to a target region of the reference genome, and the number of reads linearly reflects the abundance of the target transcript (Mortazavi et al., 2008). RNA-seq offers several advantages over microarray-based profiling, including the ability of detecting and quantifying unknown transcripts and isoforms, and providing information on alternative splicing and genetic variations.

The RNA-seq data is observed as over-dispersed and repeated counts, which is very different from expression data. The analysis of RNA-seq data is, however, very challenging. For example, the read coverage may not be uniformly distributed

along the genome because of the variations in nucleotide composition. In addition, cDNA library sizes or sequencing depths vary among different samples, therefore, it may not be appropriate to directly compare the read counts between samples. Besides, highly expressed genes tend to contribute to a large part of the sequenced reads, which thereby represses the counts of all other genes (Wesolowski et al., 2013). Traditional methods for measuring differential expression in microarray data are not immediately transferable to analyze RNA-seq data. Several methods have been developed to tackle these problems (Li and Tibshirani, 2013; Tarazona et al., 2011; Leng et al., 2012; Hardcastle and Kelly, 2010; Auer and Doerge, 2011; Di et al., 2011; Robinson et al., 2010; Anders and Huber, 2010). There are also growing interests in adjusting our FA model for it, which can be added into the joint FA framework and complete an integrated analysis.

The straightforward way seems to use the Poisson distribution and a log link, and factorize the link function afterwards. However, this single-parameter distribution constrains the variance to be equal to the mean, which may not be well-suited to model the high variability in biological replicates. Placing a gamma prior on the mean of the Poisson distribution produces a negative-binomial distribution (a.k.a, gamma-Poisson), which is a better way to address the over-dispersion problem by introducing two parameters, i.e., number of successes and the success probability. It seems reasonable to factorize count matrix using this gamma-Poisson type of prior. Zhou et al. (2012) proposed a Poisson Factor Analysis approach (PFA), and extended it to a beta-gamma-gamma-Poisson hierarchical structure ($\beta\gamma\Gamma$-PFA) that assigns Dirichlet prior to factor loadings and gamma prior to factor scores. This model is easily extendable to reflect non-negative matrix factorization (Lee and Seung, 2000) and latent Dirichlet allocation (Blei et al., 2003), which are popular approaches to model discrete data. Therefore, adjusting the $\beta\gamma\Gamma$-PFA prior for RNA-seq data is another future direction for biological exploration.

5.2.3 Software

Models from this dissertation are implemented in MATLAB. The joint Bayesian factor analysis model for horizontal integration from chapter 4 is available in `http://people.duke.edu/~lz35/Home_files/jfa.m`

Appendix A

Appendix for Chapter 3

A.1 Posterior computation

An MCMC algorithm for posterior inference of the joint Bayesian factor model proposed in the paper is provided below:

Sample $\boldsymbol{d}_k^{(r)}$

$$p(\boldsymbol{d}_k^{(r)}|-) \sim \prod_{i=1}^{M} \mathcal{N}\left(\boldsymbol{x}_i^{(r)}; \boldsymbol{D}^{(r)}(\boldsymbol{s}_i^{(c)} \odot \boldsymbol{b}_i^{(c)} + \boldsymbol{s}_i^{(r)} \odot \boldsymbol{b}_i^{(r)}), {\gamma_\epsilon^{(r)}}^{-1} \boldsymbol{I}_{N_r}\right) \mathcal{N}\left(\boldsymbol{d}_k^{(r)}; 0, \boldsymbol{\Sigma}_k\right) \tag{A.1}$$

In this and the notation below, $p(\boldsymbol{d}_k^{(r)}|-)$ is the probability of $\boldsymbol{d}_k^{(r)}$ conditioned on all other parameters being fixed to the last value in the sequence of Gibbs update equations.

It can be shown that $\boldsymbol{d}_k^{(r)}$ is drawn from a normal distribution

$$p(\boldsymbol{d}_k^{(r)}|-) \sim \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{d}_k^{(r)}}, \boldsymbol{\Sigma}_{\boldsymbol{d}_k^{(r)}}\right) \tag{A.2}$$

where

$$\boldsymbol{\Sigma}_{\boldsymbol{d}_k^{(r)}} = \left(\boldsymbol{\Sigma}_k^{-1} + \gamma_\epsilon \sum_{i=1}^{M} \left(\boldsymbol{s}_{ki}^{(c)} \odot \boldsymbol{b}_{ki}^{(c)} + \boldsymbol{s}_{ki}^{(r)} \odot \boldsymbol{b}_{ki}^{(r)} \right)^2 \boldsymbol{I}_{N_r} \right)^{-1} \tag{A.3}$$

$$\boldsymbol{\mu}_{\boldsymbol{d}_k^{(r)}} = \gamma_\epsilon \boldsymbol{\Sigma}_{\boldsymbol{d}_k} \sum_{i=1}^{M} (\boldsymbol{s}_i^{(c)} \odot \boldsymbol{b}_i^{(c)} + \boldsymbol{s}_i^{(r)} \odot \boldsymbol{b}_i^{(r)}) \hat{\boldsymbol{x}}_i^{-k,(r)} \tag{A.4}$$

where

$$\hat{\boldsymbol{x}}_i^{-k,(r)} = \boldsymbol{x}_i^{(r)} - \boldsymbol{D}^{(r)}(\boldsymbol{s}_i^{(c)} \odot \boldsymbol{b}_i^{(c)} + \boldsymbol{s}_i^{(r)} \odot \boldsymbol{b}_i^{(r)}) + \boldsymbol{d}_k^{(r)}(\boldsymbol{s}_{ki}^{(c)} \odot \boldsymbol{b}_{ki}^{(c)} + \boldsymbol{s}_{ki}^{(r)} \odot \boldsymbol{b}_{ki}^{(r)}) \tag{A.5}$$

Note, $\boldsymbol{\Sigma}_k = \gamma_s^{-1} \boldsymbol{I}_{N_r}$.

Sample $\boldsymbol{b}_k^{(c)}, \boldsymbol{b}_k^{(r)}$

$$p(b_{ik}^{(c)}|-) \sim \prod_{r=1}^{R} \mathcal{N}\left(\boldsymbol{x}_i^{(r)}; \boldsymbol{D}^{(r)}(\boldsymbol{s}_i^{(c)} \odot \boldsymbol{b}_i^{(c)} + \boldsymbol{s}_i^{(r)} \odot \boldsymbol{b}_i^{(r)}), \gamma_\epsilon^{(r)^{-1}} \boldsymbol{I}_{N_r} \right) \text{Bernoulli}(b_{ik}^{(c)}; \pi_k) \tag{A.6}$$

The posterior probability that $b_{ik}^{(c)} = 1$ is proportional to

$$p_1 = \pi_k \prod_{r=1}^{R} \exp\left(-\frac{\gamma_\epsilon^{(r)}}{2} \left(s_{ik}^{(c)^2} \boldsymbol{d}_k^{(r)^{\mathrm{T}}} \boldsymbol{d}_k^{(r)} - 2 s_{ik}^{(c)} \boldsymbol{d}_k^{(r)^{\mathrm{T}}} \boldsymbol{x}_i^{-k,(r)} \right) \right) \tag{A.7}$$

The posterior probability that $b_{ik}^{(c)} = 0$ is proportional to

$$p_0 = 1 - \pi_k \tag{A.8}$$

Hence, $b_{ik}^{(c)}$ may be drawn from a Bernoulli distribution

$$b_{ik}^{(c)} \sim \text{Bernoulli}\frac{p_1}{p_1 + p_0} \tag{A.9}$$

Similarly, $b_{ik}^{(r)}$ may be drawn from a Bernoulli distribution

$$b_{ik}^{(r)} \sim \text{Bernoulli}\frac{p_1'}{p_1' + p_0'} \tag{A.10}$$

where

$$p_1' = \pi_k \exp\left(-\frac{\gamma_\epsilon^{(r)}}{2}\left(s_{ik}^{(r)^2}\boldsymbol{d}_k^{(r)^{\mathrm{T}}}\boldsymbol{d}_k^{(r)} - 2s_{ik}^{(r)}\boldsymbol{d}_k^{(r)^{\mathrm{T}}}\boldsymbol{x}_i^{-k,(r)}\right)\right) \tag{A.11}$$

$$p_0' = 1 - \pi_k \tag{A.12}$$

Sample $\boldsymbol{s}_{(k)}^{(c)}, \boldsymbol{s}_{(k)}^{(r)}$

In this and the notation below, $\boldsymbol{x}_{(i)}^{(r)}$ represent the ith column of row of matrix $\boldsymbol{X}^{(r)}$.

$$p(\boldsymbol{s}_{(k)}^{(c)}|-) \sim \prod_{r=1}^{R}\prod_{i=1}^{N_r}\mathcal{N}\left(\boldsymbol{x}_{(i)}^{(r)}; \left(\boldsymbol{S}^{(c)^{\mathrm{T}}} \odot \boldsymbol{B}^{(c)^{\mathrm{T}}} + \boldsymbol{S}^{(r)^{\mathrm{T}}} \odot \boldsymbol{B}^{(r)^{\mathrm{T}}}\right)\boldsymbol{d}_{(i)}^{(r)}, \gamma_\epsilon^{(r)^{-1}}\boldsymbol{I}_M\right)\mathcal{N}\left(\boldsymbol{s}_{(k)}^{(c)}; 0, \boldsymbol{\Sigma}_k'\right) \tag{A.13}$$

Note, $\boldsymbol{s}_{(k)}^{(c)}$ represents the kth column of row of matrix $\boldsymbol{S}^{(c)}$.

It can be shown that $\boldsymbol{s}_{(k)}^{(c)}$ is drawn from a normal distribution

$$p(\boldsymbol{s}_{(k)}^{(c)}|-) \sim \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{s}_k}, \boldsymbol{\Sigma}_{\boldsymbol{s}_k}\right) \tag{A.14}$$

where

$$\boldsymbol{\Sigma}_{\boldsymbol{s}_{(k)}^{(c)}} = \left(\boldsymbol{\Sigma}_k'^{-1} + \sum_{r=1}^{R} \sum_{i=1}^{N_r} \left(\gamma_\epsilon^{(r)} \left(\boldsymbol{d}_{ki}^{(r)} \right)^2 \right) \left(\boldsymbol{b}_{(k)}^{(c)} \boldsymbol{b}_{(k)}^{(c)^{\mathrm{T}}} \right) \boldsymbol{I}_M \right)^{-1} \tag{A.15}$$

$$\boldsymbol{\mu}_{\boldsymbol{s}_{(k)}^{(c)}} = \boldsymbol{\Sigma}_{\boldsymbol{s}_{(k)}^{(c)}} \sum_{r=1}^{R} \sum_{i=1}^{N_r} \left(\gamma_\epsilon^{(r)} \boldsymbol{d}_{ki}^{(r)} \right) \left(\boldsymbol{b}_{(k)}^{(c)} \odot \hat{\boldsymbol{x}}_{(i)}^{-k,(r)} \right) \tag{A.16}$$

where

$$\hat{\boldsymbol{x}}_{(i)}^{-k,(r)} = \boldsymbol{x}_{(i)}^{(r)} - (\boldsymbol{S}^{(c)^{\mathrm{T}}} \odot \boldsymbol{B}^{(c)^{\mathrm{T}}}) \boldsymbol{d}_{(i)}^{(r)} + \left(\boldsymbol{s}_{(k)}^{(c)} \odot \boldsymbol{b}_{(k)}^{(c)} \right) \boldsymbol{d}_{ki}^{(r)} - \left(\boldsymbol{s}^{(r)^{\mathrm{T}}} \odot \boldsymbol{b}^{(r)^{\mathrm{T}}} \right) \boldsymbol{d}_{(i)}^{(r)} \tag{A.17}$$

Similarly, it can be shown that $\boldsymbol{s}_{(k)}^{(r)}$ is drawn from a normal distribution

$$p(\boldsymbol{s}_{(k)}^{(r)} | -) \sim \mathcal{N} \left(\boldsymbol{\mu}_{\boldsymbol{s}_k^{(r)}}, \boldsymbol{\Sigma}_{\boldsymbol{s}_k^{(r)}} \right) \tag{A.18}$$

where

$$\boldsymbol{\Sigma}_{\boldsymbol{s}_k^{(r)}} = \left(\boldsymbol{\Sigma}_k'^{-1} + \sum_{i=1}^{N_r} \left(\gamma_\epsilon^{(r)} \left(\boldsymbol{d}_{ki}^{(r)} \right)^2 \right) \left(\boldsymbol{b}_{(k)}^{(r)} \boldsymbol{b}_{(k)}^{(r)^{\mathrm{T}}} \right) \boldsymbol{I}_M \right)^{-1} \tag{A.19}$$

$$\boldsymbol{\mu}_{\boldsymbol{s}_k^{(r)}} = \boldsymbol{\Sigma}_{\boldsymbol{s}_k^{(r)}} \sum_{i=1}^{N_r} \left(\gamma_\epsilon^{(r)} \boldsymbol{d}_{ki}^{(r)} \right) \left(\boldsymbol{b}_{(k)}^{(r)} \odot \hat{\boldsymbol{x}}_{(i)}^{-k,(r)} \right) \tag{A.20}$$

where

$$\hat{\boldsymbol{x}}_{(i)}^{-k,(r)} = \boldsymbol{x}_{(i)}^{(r)} - (\boldsymbol{S}^{(r)^{\mathrm{T}}} \odot \boldsymbol{B}^{(r)^{\mathrm{T}}}) \boldsymbol{d}_{(i)}^{(r)} + \left(\boldsymbol{s}_{(k)}^{(r)} \odot \boldsymbol{b}_{(k)}^{(r)} \right) \boldsymbol{d}_{ki}^{(r)} - \left(\boldsymbol{S}^{(c)^{\mathrm{T}}} \odot \boldsymbol{B}^{(c)^{\mathrm{T}}} \right) \boldsymbol{d}_{(i)}^{(r)} \tag{A.21}$$

Sample π_k

$$p(\pi_k|-) \sim \text{Beta}(\pi_k; c\alpha, c(1-\alpha)) \prod_{i=1}^{M} \text{Bernoulli}(b_{ki}; \pi_k) \tag{A.22}$$

It can be shown that π_k may be drawn from a Beta distribution as

$$p(\pi_k|-) \sim \text{Beta}(c\alpha + \sum_{i=1}^{M} b_{ki}, c(1-\alpha) + M - \sum_{i=1}^{M} b_{ki}) \tag{A.23}$$

Sample $\gamma_\epsilon^{(r)}$

$$p(\gamma_\epsilon^{(r)}|-) \sim \prod_{i=1}^{M} \mathcal{N}\left(\boldsymbol{x}_i^{(r)}; \boldsymbol{D}(\boldsymbol{s}_i^{(c)} \odot \boldsymbol{b}_i^{(c)} + \boldsymbol{s}_i^{(r)} \odot \boldsymbol{b}_i^{(r)}), {\gamma_\epsilon^{(r)}}^{-1} \boldsymbol{I}_{N_r}\right) \text{Gamma}(\gamma_\epsilon^{(r)}, a_0, b_0) \tag{A.24}$$

It can be shown that $\gamma_\epsilon^{(r)}$ may be drawn from a Gamma distribution as,

$$p(\gamma_\epsilon^{(r)}|-) \sim \text{Gamma}\left(a_0 + \frac{1}{2} M N_r, b_0 + \frac{1}{2} \sum_{i=1}^{M} \left\| \boldsymbol{x}_i^{(r)} - \boldsymbol{D}^{(r)}(\boldsymbol{s}_i^{(c)} \odot \boldsymbol{b}_i^{(c)} + \boldsymbol{s}_i^{(r)} \odot \boldsymbol{b}_i^{(r)}) \right\|^2\right) \tag{A.25}$$

Appendix B

Appendix for Chapter 4

B.1 Posterior computation

Updated distribution for factor loadings $\beta_{g,k}$

For factor k, let $x^*_{g,d,j} = x_{g,d,j} - \mu_{g,d} - \sum_{l \neq k} \beta_{g,l} f_{l,d,j}$, so that $x^*_{g,d,j} \sim N(\beta_{g,k} f_{k,d,j}, \theta_{g,d})$. Therefore,

$$
\begin{aligned}
\beta_{g,k} | z_{g,k} = 1, - \quad &\propto \quad \prod_{d=1}^{D} \prod_{j=1}^{N_d} N(x^*_{g,d,j} | \beta_{g,k} f_{k,d,j}, \theta_{g,d}) N(\beta_{g,k} | m_{g,k}, \phi_{g,k}) \\
&\propto \quad \prod_{d=1}^{D} \prod_{j=1}^{N_d} exp(-\frac{(x^*_{g,d,j} - \beta_{g,k} f_{k,d,j})^2}{2\theta_{g,d}}) exp(-\frac{(\beta_{g,k} - m_{g,k})^2}{2\phi_{g,k}}) \\
&\propto \quad exp(-\frac{{\beta_{g,k}}^2}{2}(\frac{\sum_d \sum_j f^2_{k,d,j}}{\theta_{g,d}} + \frac{1}{\phi_{g,k}}) + \beta_{g,k}(\frac{x^*_{g,d,j} f_{k,d,j}}{\theta_{g,d}} + \frac{m_{g,k}}{\phi_{g,k}})) \\
&\sim \quad N(\eta_\beta, \upsilon_\beta)
\end{aligned}
$$

where $\upsilon_\beta = 1/(\sum_d \sum_j \frac{f^2_{k,d,j}}{\theta_{g,d}} + \frac{1}{\phi_{g,k}}), \eta_\beta = \upsilon_\beta(\sum_d \sum_j \frac{x^*_{g,d,j} f_{k,d,j}}{\theta_{g,d}} + \frac{m_{g,k}}{\phi_{g,k}})$.

if $z_{g,k} = 0$, then $\beta_{g,k} = 0$.

Updated distribution for $m_{g,k}$

$$\begin{aligned} Pr(m_{g,k} \mid -) &\propto N(\beta_{g,k}|m_{g,k}, \phi_{g,k})N(m_{g,k}; u_0, n_0) \\ &\propto exp(-\frac{(\beta_{g,k} - m_{g,k})^2}{2\phi_{g,k}})exp(-\frac{(m_{g,k} - u_0)^2}{2n_0}) \\ &\sim N(u_m, n_m) \end{aligned}$$

where $n_m = 1/(\frac{1}{n_0} + \frac{1}{\phi_{g,k}}), u_m = n_m(\frac{u_0}{n_0} + \frac{\beta_{g,k}}{\phi_{g,k}})$. m_0 and u_0 are respectively the prior mean and variance.

Updated distribution for $\phi_{g,k}$

$$\begin{aligned} Pr(\phi_{g,k} \mid -) &\propto N(\beta_{g,k}|m_{g,k}, \phi_{g,k})Ga(\phi_{g,k}^{-1}; c_0, d_0) \\ &\propto exp(-\frac{(\beta_{g,k} - m_{g,k})^2}{2\phi_{g,k}})\phi_{g,k}^{-0.5}\phi_{g,k}^{-(c_0-1)}exp(-\frac{d_0}{\phi_{g,k}}) \\ &\sim Ga(\phi_{g,k}^{-1}; c_\phi, d_\phi) \end{aligned}$$

where $c_\phi = c_0 + 0.5, d_\phi = d_0 + 0.5 \times (\beta_{g,k} - m_{g,k})^2$. c_0 and d_0 are the prior shape and rate parameters of the gamma distribution, respectively.

Updated distribution for factor scores $f_{k,d,j}$

$f_{k,d,j} = 0$ if $\gamma_{k,d,\cdot} = 0$, otherwise

$$
\begin{aligned}
f_{k,d,j}|\gamma_{k,d,\cdot} = 1, - &\propto Pr(\mathbf{x}_{d,j}|f_{k,d,j}, -)Pr(f_{k,d,j}) \\
&\propto N(\mathbf{x}^*_{d,j}; \boldsymbol{\beta}_k f_{k,d,j}, \boldsymbol{\theta}_d)N(f_{k,d,j}; 0, 1) \\
&\propto exp(-0.5 \times (\mathbf{x}^*_{d,j} - \boldsymbol{\beta}_k f_{k,d,j})^T \boldsymbol{\theta}_d^{-1}(\mathbf{x}^*_{d,j} - \boldsymbol{\beta}_k f_{k,d,j}))exp(-\frac{f^2_{k,d,j}}{2}) \\
&\propto N(f_{k,d,j}; \eta_f, \upsilon_f)
\end{aligned}
$$

where $\upsilon_f = 1/(1 + \boldsymbol{\beta}_k^T \boldsymbol{\theta}_d^{-1} \boldsymbol{\beta}_k)$, $\eta_f = \upsilon_f \times \boldsymbol{\beta}_k^T \boldsymbol{\theta}_d^{-1} \mathbf{x}^*_{d,j}$.

Updated distribution for $\gamma_{k,d,\cdot}$

$$
\begin{aligned}
\gamma_{k,d,\cdot} \mid - \quad &\propto \quad \prod_j \int Pr(\mathbf{x}_{d,j}|\gamma_{k,d,\cdot}, y_{k,d,j}, -)Pr(y_{k,d,j}|\gamma_{k,d,\cdot})Pr(\gamma_{k,d,\cdot})\mathrm{d}y_{k,d,j} \\
&= \quad \prod_d \int N(\mathbf{x}^*_{d,j}; \boldsymbol{\beta}_k \gamma_{k,d,\cdot} y_{k,d,j}, \boldsymbol{\theta}_d) N(y_{k,d,j}; 0, 1)((1-\pi_k)\delta_{\mathbf{0}} + \pi_k \delta_{\mathbf{1}})\mathrm{d}y_{k,d,j} \\
&= \quad (1-\pi_k)\delta_{\mathbf{0}} \prod_d N(\mathbf{x}^*_{d,j}; 0, \boldsymbol{\theta}_d) + \pi_k \prod_j \int \frac{1}{\sqrt{2\pi\theta_d}} exp(\\
&\quad -0.5 \times (\mathbf{x}^*_{d,j} - \boldsymbol{\beta}_k y_{k,d,j})^T \boldsymbol{\theta}_d^{-1} (\mathbf{x}^*_{d,j} - \boldsymbol{\beta}_k y_{k,d,j})) \frac{1}{\sqrt{2\pi}} exp(-\frac{y^2_{k,d,j}}{2})\mathrm{d}y_{k,d,j} \\
&\propto \quad (1-\pi_k)\delta_{\mathbf{0}} + \pi_k \int \frac{1}{2\pi} exp(-0.5 \times (y^2_{k,d,j}(\boldsymbol{\beta}_k^T \boldsymbol{\theta}_d^{-1} \boldsymbol{\beta}_k + 1) \\
&\quad -2 \times y_{k,d,j} \boldsymbol{\beta}_k^T \boldsymbol{\theta}_d^{-1} \mathbf{x}^*_{d,j}))\mathrm{d}y_{k,d,j} \\
&= \quad (1-\pi_k)\delta_{\mathbf{0}} + \\
&\quad \pi_k (\sqrt{V_{k,d}})^{N_d} exp(\frac{\sum_j {M_{k,d,j}}^2}{2V_{k,d}}) \prod_j \int exp(-\frac{(y_{k,d,j} - M_{k,d,j})^2}{2V_{k,d}}) \frac{1}{\sqrt{2\pi V_{k,d}}} \\
&\quad \mathrm{d}y_{k,d,j}
\end{aligned}
$$

where $V_{k,d} = \upsilon_f, M_{k,d,j} = \eta_f$. Therefore, sampling $\gamma_{k,d,\cdot} = 1$ with posterior odds

$$
\frac{\hat{\pi}_k}{1-\hat{\pi}_k} = \frac{\pi_k}{1-\pi_k} (\sqrt{V_{k,d}})^{N_d} exp(\frac{\sum_j {M_{k,d,j}}^2}{2V_{k,d}})
$$

Otherwise, $\gamma_{k,d,\cdot} = 0$.

Updated distribution for π_k

$$
\begin{aligned}
\pi_k \mid - &\propto \prod_d \prod_j Pr(\gamma_{k,d,\cdot}|\pi_k) Pr(\pi_k|\rho) \\
&= (1-\pi_k)^{N-S} \pi_k^S ((1-\rho) Beta(\alpha_0, \kappa_0) + \rho Beta(\kappa_0, \alpha_0)) \\
&\propto (1-\pi_k)^{N-S} \pi_k^S (1-\rho) \pi_k^{\alpha_0 - 1} (1-\pi_k)^{\kappa_0 - 1} \\
&\quad + (1-\pi_k)^{N-S} \pi_k^S \rho \pi_k^{\kappa_0 - 1} (1-\pi_k)^{\alpha_0 - 1} \\
&= (1-\rho) Beta(\pi_k; S+\alpha_0, N-S+\kappa_0) B(S+\alpha_0, N-S+\kappa_0) + \\
&\quad \rho Beta(\pi_k; S+\kappa_0, N-S+\alpha_0) B(S+\kappa_0, N-S+\alpha_0)
\end{aligned}
$$

Where $S = \sum_d \sum_j I(\gamma_{k,d,\cdot} = 1), N$ is the total sample size. Therefore, sampling π_k from $Beta(\pi_k; S+\kappa_0, N-S+\alpha_0)$ with posterior odds

$$
\frac{\hat{\rho}}{1-\hat{\rho}} = \frac{\rho}{1-\rho} \frac{B(S+\kappa_0, N-S+\alpha_0)}{B(S+\alpha_0, N-S+\kappa_0)}
$$

Otherwise, π_k is sampled from $Beta(\pi_k; S+\alpha_0, N-S+\kappa_0)$.$\kappa_0$ and α_0 are the prior shape parameters of the beta distribution.

Updated distribution for ρ

$$
\begin{aligned}
\rho \mid - &\propto \prod_k Pr(\pi_k|\rho) Pr(\rho) \\
&= (1-\rho)^{K-Q} \rho^Q Beta(e_0, l_0) \\
&\propto (1-\rho)^{K-Q} \rho^Q \rho^{e_0 - 1} (1-\rho)^{l_0 - 1} \\
&\sim Beta(\rho; Q+e_0, K-Q+l_0)
\end{aligned}
$$

where Q is the number of times π_k is sampled from $Beta(\pi_k; S+\kappa_0, N-S+\alpha_0).e_0$ and l_0 are the shape parameters of the beta distribution.

Updated distribution for noise variance $\boldsymbol{\theta}_d$

$$
\begin{aligned}
\boldsymbol{\theta}_d &\propto \prod_j N(\mathbf{x}_{d,j} - \boldsymbol{\mu}_d; \boldsymbol{\beta}\mathbf{f}_{d,j}, \boldsymbol{\theta}_d) Ga(\boldsymbol{\theta}_d^{-1}; a_0, b_0) \\
&\propto exp(-0.5 \times \sum_{j=1}^{N_d} (\mathbf{x}_{d,j} - \boldsymbol{\mu}_d - \boldsymbol{\beta}\mathbf{f}_{d,j})^T \boldsymbol{\theta}_d^{-1} (\mathbf{x}_{d,j} - \boldsymbol{\mu}_d - \boldsymbol{\beta}\mathbf{f}_{d,j})) \boldsymbol{\theta}_d^{-\frac{N_d}{2}} \boldsymbol{\theta}_d^{-(a_0-1)} \\
&\quad exp(-\frac{b_0}{\boldsymbol{\theta}_d}) \\
&\sim Ga(\boldsymbol{\theta}_d^{-1}; a_\theta, b_\theta)
\end{aligned}
$$

where $a_\theta = a_0 + \frac{N_d}{2}, b_\theta = b_0 + \frac{\sum_{j=1}^{N_d} (\mathbf{x}_{d,j} - \boldsymbol{\mu}_d - \boldsymbol{\beta}\mathbf{f}_{d,j})^T (\mathbf{x}_{d,j} - \boldsymbol{\mu}_d - \boldsymbol{\beta}\mathbf{f}_{d,j})}{2}$. a_0 and b_0 are the prior shape and rate parameters of the gamma distribution, respectively.

Updated distribution for sample mean $\boldsymbol{\mu}_d$

$$
\begin{aligned}
\boldsymbol{\mu}_d &\propto \prod_j N(\mathbf{x}_{d,j} - \boldsymbol{\mu}_d; \boldsymbol{\beta}\mathbf{f}_{d,j}, \boldsymbol{\theta}_d) N(\boldsymbol{\mu}_d; m_0, s_0) \\
&\propto exp(-0.5 \times \sum_{j=1}^{N_d} (\mathbf{x}_{d,j} - \boldsymbol{\mu}_d - \boldsymbol{\beta}\mathbf{f}_{d,j})^T \boldsymbol{\theta}_d^{-1} (\mathbf{x}_{d,j} - \boldsymbol{\mu}_d - \boldsymbol{\beta}\mathbf{f}_{d,j})) \\
&\quad exp(-\frac{(\boldsymbol{\mu}_d - m_0)^T (\boldsymbol{\mu}_d - m_0))}{2s_0}) \\
&\sim N(\boldsymbol{\mu_d}; m_\mu, s_\mu)
\end{aligned}
$$

where $s_\mu = 1/(\frac{1}{s_0} + \frac{N_d}{\boldsymbol{\theta}_d}), m_\mu = s_\mu \times (\frac{m_0}{s_0} + \boldsymbol{\theta}_d^{-1} \oplus \sum_{j=1}^{N_d}(\mathbf{x}_{d,j} - \boldsymbol{\beta}\mathbf{f}_{d,j}))$. s_0 and m_0 are respectively prior mean and variance parameters.

Bibliography

Airoldi, E. M., Huttenhower, C., Gresham, D., Lu, C., Caudy, A. A., Dunham, M. J., Broach, J. R., Botstein, D., and Troyanskaya, O. G. (2009), "Predicting Cellular Growth from Gene Expression Signatures," *PLoS Computational Biology*, 5, e1000257.

Akahira, J., Aoki, M., Suzuki, T., Moriya, T., Niikura, H., Ito, K., Inoue, S., Okamura, K., Sasano, H., and Yaegashi, N. (2004), "Expression of EBAG9/RCAS1 is associated with advanced disease in human epithelial ovarian cancer." *British Journal of Cancer*, 90, 2197–202.

Akaike, H. (1987), "Factor analysis and AIC," *Psychometrika*, 52, 317–332.

Akavia, U. D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H. C., Pochanard, P., Mozes, E., Garraway, L. A., and Pe'er, D. (2010), "An Integrated Approach to Uncover Drivers of Cancer," *Cell*, 143, 1005 – 1017.

Albertson, D. (2003), "Profiling breast cancer by array CGH." *Breast Cancer Research and Treatment*, 78, 289–298.

Anders, S. and Huber, W. (2010), "Differential expression analysis for sequence count data," *Genome Biology*, 11, R106.

Archambeau, C. and Bach, F. R. (2008), "Sparse probabilistic projections," *Advances in Neural Information Processing Systems 21*, pp. 73–80.

Auer, P. and Doerge, R. (2011), "A two-stage poisson model for testing RNA-seq data," *Statistical Application in Genetics and Molecular Biology*, 10, Article 26.

Bach, F. R. and Jordan, M. I. (2005), "A probabilistic interpretation of canonical correlation analysis," Tech. rep., University of California, Berkeley.

Beheshti, B., Braude, I., Marrano, P., Thorner, P., Zielenska, M., and Squire, J. (2003), "Chromosomal localization of DNA amplifications in neuroblastoma tumors using cDNA microarray comparative genomic hybridization." *Neoplasia*, 5, 53–62.

Beier, U. H., Wang, L., Bhatti, T. R., Liu, Y., Han, R., Ge, G., and Hancock, W. W. (March 1, 2011), "Sirtuin-1 Targeting Promotes Foxp3+T-Regulatory Cell Function and Prolongs Allograft Survival," *Molecular and Cellular Biology*, 31, 1022–1029.

Bhattacharya, A. and Dunson, D. B. (2011), "Sparse Bayesian infinite factor models," *Biometrika*, 98, 291–306.

Blei, D. M., Ng, A. Y., Jordan, M. I., and Lafferty, J. (2003), "Latent dirichlet allocation," *Journal of Machine Learning Research*, 3, 2003.

Borga, M. (1998), "Learning Multidimensional Signal Processing," Linkping Studies in Science and Technology. Dissertations No. 531, Linkping University, Sweden.

Bourgon, R., Gentleman, R., and Huber, W. (2010), "Independent filtering increases detection power for high-throughput experiments," *Proceedings of the National Academy of Sciences*, 107, 9546–9551.

Bühlmann, P. and Hothorn, T. (2005), "Spike and slab variable selection: Frequentist and Bayesian strategies," *Statistical Science*.

Buxant, F., Bucella, D., Anaf, V., Simon, P., and JC, N. (2009), "Glucocorticoid receptor expression in cervical intraepithelial neoplasia and invasive squamous cell carcinoma of the cervix." *European Journal of Gynaecology Oncology*, 30, 259–62.

Carvalho, C., Chang, J., Lucas, J., Nevins, J., Wang, Q., and West, M. (2008a), "High-Dimensional Sparse Factor Modelling: Applications in Gene Expression Genomics," *Journal of the American Statistical Association*, 103, 1438–1456.

Carvalho, C. M. (2006), "Structure and Sparsity in High-Dimensional Multivariate Analysis," Ph.D. thesis, Duke University, ISDS.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008b), "High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics," *Journal of the American Statistical Association*, 103, 1438–1456.

Chen, M., Zaas, A., Woods, C., Ginsburg, G., Lucas, J., Dunson, D., and Carin, L. (2011), "Predicting Viral Infection From High-Dimensional Biomarker Trajectories," *Journal of the American Statistical Association*, pp. 1–21.

Chung, E. and Kondo, M. (2011), "Role of Ras/Raf/MEK/ERK signaling in physiological hematopoiesis and leukemia development," *Immunologic Research*, 49, 248–268.

Cleveland, W. S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, pp. 829–836.

Congdon, K. L., Voermans, C., Ferguson, E. C., DiMascio, L. N., Uqoezwa, M., Zhao, C., and Reya, T. (2008), "Activation of Wnt Signaling in Hematopoietic Regeneration," *STEM CELLS*, 26, 1202–1210.

Coronnello, C. and Benos, P. V. (2013), "ComiR: combinatorial microRNA target prediction tool," *Nucleic Acids Research*.

Creighton, C. J., Fountain, M. D., Yu, Z., Nagaraja, A. K., Zhu, H., Khan, M., Olokpa, E., Zariff, A., Gunaratne, P. H., Matzuk, M. M., and Anderson, M. L. (2010), "Molecular Profiling Uncovers a p53-Associated Role for MicroRNA-31 in Inhibiting the Proliferation of Serous Ovarian Carcinomas and Other Cancers," *Cancer Research*, 70, 1906–1915.

De Zoeten, E. F., Wang, L., Butler, K., Beier, U. H., Akimova, T., Sai, H., Bradner, J. E., Mazitschek, R., Kozikowski, A. P., Matthias, P., and Hancock, W. W. (2011), "Histone Deacetylase 6 and Heat Shock Protein 90 Control the Functions of Foxp3+ T-Regulatory Cells," *Molecular and Cellular Biology*, 31, 2066–2078.

Dennis, G., Sherman, B., Hosack, D., Yang, J., Gao, W., Lane, H. C., and Lempicki, R. (2003), "DAVID: Database for Annotation, Visualization, and Integrated Discovery," *Genome Biology*, 4, P3.

Di, Y., Schafer, D., Cumbie, J., and Chang, J. (2011), "The NBP negative binomial model for assessing differential gene expression from RNA-seq," *Statistical Application in Genetics Molecular Biolology*, 10, Article 24.

Diebolt, J. and Robert, C. P. (1994), "Estimation of Finite Mixture Distributions through Bayesian Sampling," *Journal of the Royal Statistical Society. Series B (Methodological)*, 56, pp. 363–375.

Du, L., Chen, M., Lucas, J. E., and Carin, L. (2010), "Sticky hidden Markov modeling of comparative genomic hybridization." *IEEE Transactions on Signal Processing*, 58, 5353–5368.

Eder, M. and Scherr, M. (2005), "MicroRNA and Lung Cancer," *New England Journal of Medicine*, 352, 2446–2448, PMID: 15944431.

Edgar, R., Domrachev, M., and Lash, A. E. (2002), "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, 30, 207–210.

Emdad, L., Sarkar, D., Su, Z.-Z., Lee, S.-G., Kang, D.-C., Bruce, J. N., Volsky, D. J., and Fisher, P. B. (2007), "Astrocyte elevated gene-1: Recent insights into a novel gene involved in tumor progression, metastasis and neurodegeneration," *Pharmacology and Therapeutics*, 114, 155 – 170.

Enright, A., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D. (2003), "MicroRNA targets in Drosophila," *Genome Biology*, 5, R1.

Frank, B., Bermejo, J. L., Hemminki, K., Sutter, C., Wappenschmidt, B., Meindl, A., Kiechle-Bahat, M., Bugert, P., Schmutzler, R. K., Bartram, C. R., and Burwinkel, B. (2007), "Copy number variant in the candidate tumor suppressor gene MTUS1 and familial breast cancer risk," *Carcinogenesis*, 28, 1442–1445.

Fridlyand, J., Snijders, A. M., Pinkel, D., Albertson, D. G., and Jain, A. N. (2004), "Hidden Markov models approach to the analysis of array CGH data," *Journal of Multivariate Analysis*, 90, 132 – 153.

Garber, M. E., Troyanskaya, O. G., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., van de Rijn, M., Rosen, G. D., Perou, C. M., Whyte, R. I., Altman, R. B., Brown, P. O., Botstein, D., and Petersen, I. (2001), "Diversity of gene expression in adenocarcinoma of the lung," *Proceedings of the National Academy of Sciences*, 98, 13784–13789.

Garraway, L. A., Widlund, H. R., Rubin, M. A., Getz, G., Berger, A. J., Ramaswamy, S., Beroukhim, R., Milner, D. A., Granter, S. R., Du, J., Lee, C., Wagner, S. N., Li, C., Golub, T. R., Rimm, D. L., Meyerson, M. L., Fisher, D. E., and Sellers, W. R. (2005), "Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma," *Nature*, 436, 117–122.

Gentile, M., Latonen, L., and Laiho, M. (2003), "Cell cycle arrest and apoptosis provoked by UV radiation-induced DNA damage are transcriptionally highly divergent responses," *Nucleic Acids Research*, 31, 4779–4790.

Gentleman, R., Carey, V., Huber, W., Irizarry, R., and Dudoit, S. (2005), *Bioinformatics and Computational Biology Solutions Using R and Bioconductor (Statistics for Biology and Health)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Georgantas, R. W., Tanadve, V., Malehorn, M., Heimfeld, S., Chen, C., Carr, L., Martinez-Murillo, F., Riggins, G., Kowalski, J., and Civin, C. I. (2004), "Microarray and Serial Analysis of Gene Expression Analyses Identify Known and Novel Transcripts Overexpressed in Hematopoietic Stem Cells," *Cancer Research*, 64, 4434–4441.

George, E. I. and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.

Gerber, H.-P. and Ferrara, N. (2003), "The role of VEGF in normal and neoplastic hematopoiesis," *Journal of Molecular Medicine*, 81, 20–31.

Gower, A., Spira, A., and Lenburg, M. (2011), "Discovering biological connections between experimental conditions based on common patterns of differential gene expression," *BMC Bioinformatics*, 12, 381.

Griffiths, T. L. and Ghahramani, Z. (2005), "Infinite latent feature models and the Indian buffet process," in *Neural Information Processing Systems*, pp. 475–482.

Gui, J. and Li, H. (2005), "Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data," *Bioinformatics*, 21, 3001–3008.

Habermehl, D., Parkitna, J. R., Kaden, S., Brgger, B., Wieland, F., Grne, H.-J., and Schtz, G. (2011), "Glucocorticoid Activity during Lung Maturation Is Essential in Mesenchymal and Less in Alveolar Epithelial Cells," *Molecular Endocrinology*, 25, 1280–1288.

Han, D., Zhang, M., Ma, J., Hong, J., Chen, C., Zhang, B., Huang, L., Lv, W., Yin, L., Zhang, A., Zhang, H., Zhang, Z., Vidyasagar, S., Okunieff, P., and Zhang, L. (2012), "Transition Pattern and Mechanism of B-lymphocyte Precursors in Regenerated Mouse Bone Marrow after Subtotal Body Irradiation," *PLoS ONE*, 7, e46560.

Hanzelmann, S., Castelo, R., and Guinney, J. (2013), "GSVA: gene set variation analysis for microarray and RNA-Seq data," *BMC Bioinformatics*, 14, 7.

Hardcastle, T. and Kelly, K. (2010), "baySeq: empirical Bayesian methods for identifying differential expression in sequence count data," *BMC Bioinformatics*, 11, 422.

Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. (2004), "Canonical Correlation Analysis: An Overview with Application to Learning Methods," *Neural Computation*, 16, 2639–2664.

Henao, R., Thompson, J. W., Moseley, M. A., Ginsburg, G. S., Carin, L., and Lucas, J. E. (to be appear 2013), "Latent protein trees," *Annals of Applied Statistics*.

Hosack, D., Dennis, G., Sherman, B., Lane, H., and Lempicki, R. (2003), "Identifying biological themes within lists of genes with EASE," *Genome Biology*, 4, R70.

Hotelling, H. (1936), "Relations Between Two Sets of Variates," *Biometrika*, 28, 321–377.

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

Hu, G., Chong, R. A., Yang, Q., Wei, Y., Blanco, M. A., Li, F., Reiss, M., Au, J. L.-S., Haffty, B. G., and Kang, Y. (2009), "MTDH Activation by 8q22 Genomic Gain Promotes Chemoresistance and Metastasis of Poor-Prognosis Breast Cancer," *Cancer Cell*, 15, 9 – 20.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009a), "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Research*, 37, 1–13.

Huang, J., Zhang, Y., Bersenev, A., OBrien, W. T., Tong, W., Emerson, S. G., and Klein, P. S. (2009b), "Pivotal role for glycogen synthase kinase3 in hematopoietic stem cell homeostasis in mice," *The Journal of Clinical Investigation*, 119, 3519–3529.

Hup, 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.

Ishwaran, H. and Rao, J. S. (2005), "Spike and slab variable selection: Frequentist and Bayesian strategies," *Annals of Statistics*, 33, 730–773.

Iwasaki, T., Nakashima, M., Watanabe, T., Yamamoto, S., Inoue, Y., Yamanaka, H., Matsumura, A., Iuchi, K., Mori, T., and Okada, M. (2000), "Expression and prognostic significance in lung cancer of human tumor-associated antigen RCAS1," *International Journal of Cancer*, 89, 488–493.

Jaffe, I. Z., Newfell, B. G., Aronovitz, M., Mohammad, N. N., McGraw, A. P., Perreault, R. E., Carmeliet, P., Ehsan, A., and Mendelsohn, M. E. (2010), "Placental growth factor mediates aldosterone-dependent vascular injury in mice," *The Journal of Clinical Investigation*, 120, 3891–3900.

Jeong, J., Li, L., Liu, Y., Nephew, K., Huang, T., and Shen, C. (2010), "An empirical Bayes model for gene expression and methylation profiles in antiestrogen resistant breast cancer," *BMC Medical Genomics*, 3, 55.

John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., and Marks, D. S. (2004), "Human MicroRNA Targets," *PLoS Biology*, 2, e363.

Jönsson, G., Staaf, J., Vallon-Christersson, J., Ringnr, M., Holm, K., Hegardt, C., Gunnarsson, H., Fagerholm, R., Strand, C., Agnarsson, B., Kilpivaara, O., Luts, L., Heikkilä, P., Aittomäki, K., Blomqvist, C., Loman, N., Malmstrm, P., Olsson, H., Th Johannsson, O., Arason, A., Nevanlinna, H., Barkardottir, R., and Borg, . (2010), "Genomic subtypes of breast cancer identified by array-comparative genomic hybridization display distinct molecular and clinical characteristics," *Breast Cancer Research*, 12, 1–14.

Kanehisa, M. and Goto, S. (2000), "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, 28, 27–30.

Karoui, E. (2008), "Operator norm consistent estimation of large-dimensional sparse covariance matrices," *Annals of Statistics*, 36, 2717–2756.

Karpov, A., Semenova, Y., Takhauov, R., Litvinenko, T., and Kalinkin, D. (2012), "The Risk of Acute Myocardial Infarction and Arterial Hypertension in a Cohort of Male Employees of a Siberian Group of Chemical Enterprises Exposed to Long-Term Irradiation," *Health Physics*, 103, 15–23.

Kendziorski, C. M., Chen, M., Yuan, M., Lan, H., and Attie, A. D. (2006), "Statistical Methods for Expression Quantitative Trait Loci (eQTL) Mapping," *Biometrics*, 62, 19–27.

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007), "The role of site accessibility in microRNA target recognition," *Nature Genetics*, 39, 1278–1284.

Kiechle, M., Jacobsen, A., Schwarz-Boeger, U., Hedderich, J., Pfisterer, J., and Arnold, N. (2001), "Comparative genomic hybridization detects genetic imbalances in primary ovarian carcinomas as correlated with grade of differentiation." *Cancer*, 91, 534–40.

Kikuno, N., Shiina, H., Urakami, S., Kawamoto, K., Hirata, H., Tanaka, Y., Place, R. F., Pookot, D., Majid, S., and Igawa, M. (2007), "Knockdown of astrocyte-elevated gene-1 inhibits prostate cancer progression through upregulation of FOXO3a activity," *Oncogene*, pp. 7647–7655.

Kim, M.-S., Kim, Y., Lee, D., Seo, J., Cho, K., Eun, H., and Chung, J. (2009), "Acute exposure of human skin to ultraviolet or infrared radiation or heat stimuli increases mast cell numbers and tryptase expression in human skin *in vivo*," *British Journal of Dermatology*, 160, 393–402.

Klami, A. and Kaski, S. (2008), "Probabilistic approach to detecting dependencies between data sets," *Neurocomputing*, 72, 39–46.

Kothandaraman, N., Bajic, V., Brendan, P., Huak, C., Keow, P., Razvi, K., Salto-Tellez, M., and Choolani, M. (2010), "E2F5 status significantly improves malignancy diagnosis of epithelial ovarian cancer," *BMC Cancer*, 10, 64.

Lanckriet, G. R. G., De Bie, T., Cristianini, N., Jordan, M. I., and Noble, W. S. (2004), "A statistical framework for genomic data fusion," *Bioinformatics*, 20, 2626–2635.

Le, H.-S. and Bar-Joseph, Z. (2011), "Inferring interaction networks using the IBP applied to microRNA target prediction," *Advances in Neural Information Processing Systems 24*, pp. 235–243.

Lee, D. D. and Seung, H. S. (2000), "Algorithms for Non-negative Matrix Factorization," in *In NIPS*, pp. 556–562, MIT Press.

Lee, M. S., Hanspers, K., Barker, C. S., Korn, A. P., and McCune, J. M. (2004), "Gene expression profiles during human CD4+ T cell differentiation," *International Immunology*, 16, 1109–1124.

Lee, S.-G., Su, Z.-Z., Emdad, L., Sarkar, D., Franke, T. F., and Fisher, P. B. (2007), "Astrocyte elevated gene-1 activates cell survival pathways through PI3K-Akt signaling," *Oncogene*, pp. 1114–1121.

Leng, N., Dawson, J., Thomson, J., Ruotti, V., Rissman, A., Smits, B., Haag, J., Gould, M., Stewart, R., and Kendziorski, C. (2012), "EBSeq: an empirical bayes hierarchical model for inference in RNA-seq experiments," *Bioinformatics*.

Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005), "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets." *Cell*, 120, 15–20.

Li, J. and Tibshirani, R. (2013), "Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-seq data," *Statistical Methods in Medical Research*, 22, 519–536.

Li, M., Balch, C., Montgomery, J., Jeong, M., Chung, J., Yan, P., Huang, T., Kim, S., and Nephew, K. (2009), "Integrated analysis of DNA methylation and gene expression reveals specific signaling pathways associated with platinum resistance in ovarian cancer," *BMC Medical Genomics*, 2, 1–13.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdttir, H., Tamayo, P., and Mesirov, J. P. (2011), "Molecular signatures database (MSigDB) 3.0," *Bioinformatics*, 27, 1739–1740.

Lockwood, W. W., Chari, R., Chi, B., and Lam, W. L. (2005), "Recent advances in array comparative genomic hybridization technologies and their applications in human genetics," *European Journal of Human Genetics*, 14, 139–148.

Louhimo, R. and Hautaniemi, S. (2011), "CNAmet: an R package for integrating copy number, methylation and expression data," *Bioinformatics*, 27, 887–888.

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J., and West, M. (2006), "Sparse Statistical Modelling in Gene Expression Genomics," *Bayesian Inference for Gene Expression and Proteomics*, pp. 155–176.

Lucas, J., Carvalho, C., and West, M. (2009), "A Bayesian analysis strategy for cross-study translation of gene expression biomarkers." *Statistical applications in genetics and molecular biology*, 8.

Lucas, J. E., Kung, H.-N., and Chi, J.-T. A. (2010), "Latent Factor Analysis to Discover Pathway-Associated Putative Segmental Aneuploidies in Human Cancers," *PLoS Computational Biology*, 6.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., and Zisserman, A. (2008), "Supervised Dictionary Learning," in *Neural Information Processing Systems*, pp. 1033–1040.

Malumbres, M. (2012), "miRNAs versus oncogenes: the power of social networking," *Molecular Systems Biology*, 8.

McCubrey, J. A., Steelman, L. S., Abrams, S. L., Bertrand, F. E., Ludwig, D. E., Basecke, J., Libra, M., Stivala, F., Milella, M., Tafuri, A., Lunghi, P., Bonati, A., and Martelli, A. M. (2008), "Targeting survival cascades induced by activation of Ras/Raf/MEK/ERK, PI3K/PTEN/Akt/mTOR and Jak/STAT pathways for effective leukemia therapy," *Leukemia*, 22, 708–722.

McKay, S. C., Unger, K., Pericleous, S., Stamp, G., Thomas, G., Hutchins, R. R., and Spalding, D. R. C. (2011), "Array comparative genomic hybridization identifies novel potential therapeutic targets in cholangiocarcinoma," *HPB*, 13, 309–319.

Mihailidou, A. S., Loan Le, T. Y., Mardini, M., and Funder, J. W. (2009), "Glucocorticoids Activate Cardiac Mineralocorticoid Receptors During Experimental Myocardial Infarction," *Hypertension*, 54, 1306–1312.

Mitchell, T. J. and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83, pp. 1023–1032.

Miyamoto, K., Morishita, Y., Yamazaki, M., Minamino, N., Kangawa, K., Matsuo, H., Mizutani, T., Yamada, K., and Minegishi, T. (2001), "Isolation and Characterization of Vascular Smooth Muscle Cell Growth Promoting Factor from Bovine Ovarian Follicular Fluid and Its cDNA Cloning from Bovine and Human Ovary," *Archives of Biochemistry and Biophysics*, 390, 93 – 100.

Moreno-Asso, A., Castao, C., Grilli, A., Novials, A., and Servitja, J.-M. (2013), "Glucose regulation of a cell cycle gene module is selectively lost in mouse pancreatic islets during ageing," *Diabetologia*, 56, 1761–1772.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008), "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, 5, 621–628.

Nakajima, T., Matsumoto, K., Suto, H., Tanaka, K., Ebisawa, M., Tomita, H., Yuki, K., Katsunuma, T., Akasawa, A., Hashida, R., Sugita, Y., Ogawa, H., Ra, C., and Saito, H. (2001), "Gene expression screening of human mast cells and eosinophils using high-density oligonucleotide probe arrays: abundant expression of major basic protein in mast cells," *Blood*, 98, 1127–1134.

Nakashima, M., Sonoda, K., and Watanabe, T. (1999), "Inhibition of cell growth and induction of apoptotic cell death by the human tumor-associated antigen RCAS1," *Nature Medicine*, 5, 938–942.

Nouzova, M., Holtan, N., Oshiro, M. M., Isett, R. B., Munoz-Rodriguez, J. L., List, A. F., Narro, M. L., Miller, S. J., Merchant, N. C., and Futscher, B. W. (2004), "Epigenomic Changes during Leukemia Cell Differentiation: Analysis of Histone Acetylation and Cytosine Methylation Using CpG Island Microarrays," *Journal of Pharmacology and Experimental Therapeutics*, 311, 968–981.

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004), "Circular binary segmentation for the analysis of array?based DNA copy number data," *Biostatistics*, 5, 557–572.

Paisley, J. and Carin, L. (2009), "Nonparametric factor analysis with beta process priors," in *Proceedings of the 26th International Conference on Machine Learning*, pp. 777–784.

Parkitna, J. R., Bilbao, A., Rieker, C., Engblom, D., Piechota, M., Nordheim, A., Spanagel, R., and Schtz, G. (2010), "Loss of the serum response factor in the dopamine system leads to hyperactivity," *The FASEB Journal*, 24, 2427–2435.

Parsons, D. W., Jones, S., Zhang, X., Lin, J. C.-H., Leary, R. J., Angenendt, P., Mankoo, P., Carter, H., Siu, I.-M., Gallia, G. L., Olivi, A., McLendon, R., Rasheed, B. A., Keir, S., Nikolskaya, T., Nikolsky, Y., Busam, D. A., Tekleab, H., Diaz, L. A., Hartigan, J., Smith, D. R., Strausberg, R. L., Marie, S. K. N., Shinjo, S. M. O., Yan, H., Riggins, G. J., Bigner, D. D., Karchin, R., Papadopoulos, N., Parmigiani, G., Vogelstein, B., Velculescu, V. E., and Kinzler, K. W. (2008), "An Integrated Genomic Analysis of Human Glioblastoma Multiforme," *Science*, 321, 1807–1812.

Pe'er, D. and Hacohen, N. (2011), "Principles and Strategies for Developing Network Models in Cancer," *Cell*, 144, 864 – 873.

Pils, D., Horak, P., Gleiss, A., Sax, C., Fabjani, G., Moebus, V. J., Zielinski, C., Reinthaller, A., Zeillinger, R., and Krainer, M. (2005), "Five genes from chromosomal band 8p22 are significantly down-regulated in ovarian carcinoma," *Cancer*, 104, 2417–2429.

Pinkel, D. and Albertson, D. G. (2005), "Array comparative genomic hybridization and its applications in cancer." *Nature genetics*, 37 Suppl.

Pyeon, D., Newton, M. A., Lambert, P. F., den Boon, J. A., Sengupta, S., Marsit, C. J., Woodworth, C. D., Connor, J. P., Haugen, T. H., Smith, E. M., Kelsey, K. T., Turek, L. P., and Ahlquist, P. (2007), "Fundamental Differences in Cell Cycle Deregulation in Human PapillomavirusPositive and Human Papillomavirus-Negative Head/Neck and Cervical Cancers," *Cancer Research*, 67, 4605–4619.

Pyle-Chenault, R., Stolk, J., Molesh, D., Boyle-Harlan, D., McNeill, P., Repasky, E., Jiang, Z., Fanger, G., and Xu, J. (2005), "VSGP/F-spondin: a new ovarian cancer marker." *Tumor Biology*, 26, 245–257.

Qin, L.-X. (2008), "An Integrative Analysis of microRNA and mRNA Expression - A Case Study," *Cancer Informatics*, 6, 369–379.

Rai, P. and Daume, H. (2009), "Multi-Label Prediction via Sparse Infinite CCA," in *Advances in Neural Information Processing Systems 22*, eds. Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, pp. 1518–1526, Proceedings of the Conference on Neural Information Processing Systems (NIPS).

Rashi-Elkeles, S., Elkon, R., Weizman, N., Linhart, C., Amariglio, N., Sternberg, G., Rechavi, G., Barzilai, A., Shamir, R., and Shiloh, Y. (2005), "Parallel induction of ATM-dependent pro- and antiapoptotic signals in response to ionizing radiation in murine lymphoid tissue," *Oncogene*, 25, 0950–9232.

Ray, P. and Carin, L. (2011), "Non-parametric Bayesian modeling and fusion of spatio-temporal information sources," in *2011 Proceedings of the 14th International Conference on Information Fusion (FUSION)*, pp. 1 –7.

Rennstam, K., Ahlstedt-Soini, M., Baldetorp, B., Bendahl, P.-O., Borg, A., Karhu, R., Tanner, M., Tirkkonen, M., and Isola, J. (2003), "Patterns of Chromosomal Imbalances Defines Subgroups of Breast Cancer with Distinct Clinical Features and Prognosis. A Study of 305 Tumors by Comparative Genomic Hybridization," *Cancer Research*, 63, 8861–8868.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010), "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, 26, 139–140.

Rueda, O. and Uriarte, R. D. (2007), "Flexible and accurate detection of genomic copy-number changes from aCGH," *PLoS Computational Biology*, preprint, 122.

Sandy, A., Jones, M., and Maillard, I. (2012), "Notch Signaling and Development of the Hematopoietic System," in *Notch Signaling in Embryology and Cancer*, eds. J. Reichrath and S. Reichrath, vol. 727 of *Advances in Experimental Medicine and Biology*, pp. 71–88, Springer US.

Schmidt, M. N., Winther, O., and Hansen, L. K. (2009), "Bayesian non-negative matrix factorization," in *Independent Component Analysis and Signal Separation, International Conference on*, vol. 5441 of *Lecture Notes in Computer Science (LNCS)*, pp. 540–547, Springer.

Schuringa, J. J., Chung, K. Y., Morrone, G., and Moore, M. A. (2004), "Constitutive Activation of STAT5A Promotes Human Hematopoietic Stem Cell Self-Renewal and Erythroid Differentiation," *The Journal of Experimental Medicine*, 200, 623–635.

Schwarz, G. (1978), "Estimating the dimension of a model," *The Annals of Statistics*, 6, 461–464.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003), "Module Networks: Discovering Regulatory Modules and their Condition Specific Regulators from Gene Expression Data," *Nature Genetics*, 34, 2003.

Sesto, A., Navarro, M., Burslem, F., and Jorcano, J. L. (2002), "Analysis of the ultraviolet B response in primary human keratinocytes using oligonucleotide microarrays," *Proceedings of the National Academy of Sciences*, 99, 2965–2970.

Shamir, R. (2010), "Analysis of DNA chips and gene networks," *Lecture 14a*.

Shen, R., Olshen, A. B., and Ladanyi, M. (2009), "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis," *Bioinformatics*, 25, 2906–2912.

Smirnov, D. A., Brady, L., Halasa, K., Morley, M., Solomon, S., and Cheung, V. G. (2012), "Genetic variation in radiation-induced cell death," *Genome Research*, 22, 332–339.

Smith, L. (2005), "Exploratory Genomic Data Analysis," in *Medical Informatics*, eds. H. Chen, S. Fuller, C. Friedman, and W. Hersh, vol. 8 of *Integrated Series in Information Systems*, pp. 573–592, Springer US.

Stingo, F. C., Chen, Y. A., Vannucci, M., Barrier, M., and Mirkes, P. E. (2010), "A Bayesian Graphical Modeling Approach to MicroRNA Regulatory Network Inference." *The Annals of Applied Statistics*, 4, 2024–2048.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005), "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, 102, 15545–15550.

Talloen, W., Clevert, D.-A., Hochreiter, S., Amaratunga, D., Bijnens, L., Kass, S., and Gohlmann, H. (2007), "I/NI-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data," *Bioinformatics*, 23, 2897–2902.

Tamayo, P., Steinhardt, G., Liberzon, A., and Mesirov, J. P. (2012), "The limitations of simple gene set enrichment analysis assuming gene independence," *Statistical Methods in Medical Research*.

Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A., and Conesa, A. (2011), "Differential expression in RNA-seq: a matter of depth," *Genome Research*, 21, 2213–2223.

Taylor, K. H., Pena-Hernandez, K. E., Davis, J. W., Arthur, G. L., Duff, D. J., Shi, H., Rahmatpanah, F. B., Sjahputera, O., and Caldwell, C. W. (2007), "Large-Scale CpG Methylation Analysis Identifies Novel Candidate Genes and Reveals Methylation Hotspots in Acute Lymphoblastic Leukemia," *Cancer Research*, 67, 2617–2625.

TCGA (2008), "Comprehensive genomic characterization defines human glioblastoma genes and core pathways." *Nature*, 455, 1061–1068.

TCGA (2011), "Integrated genomic analyses of ovarian carcinoma," *Nature*, 474, 609–615.

Thibaux, R. and Jordan, M. I. (2007), "Hierarchical beta processes and the Indian buffet process," in *Proceedings of the 11th Conference on Artificial Intelligence and Statistic*.

Tipping, M. E. (2001), "Sparse bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, 1, 211–244.

Trowbridge, J. J., Scott, M. P., and Bhatia, M. (2006), "Hedgehog modulates cell cycle regulators in stem cells to control hematopoietic regeneration," *Proceedings of the National Academy of Sciences*, 103, 14134–14139.

Tseng, C.-W., Lin, C.-C., Chen, C.-N., Huang, H.-C., and Juan, H.-F. (2011), "Integrative network analysis reveals active microRNAs and their functions in gastric cancer," *BMC Systems Biology*, 5, 99.

Vastrik, I., D'Eustachio, P., Schmidt, E., Joshi-Tope, G., Gopinath, G., Croft, D., de Bono, B., Gillespie, M., Jassal, B., Lewis, S., Matthews, L., Wu, G., Birney, E., and Stein, L. (2007), "Reactome: a knowledge base of biologic pathways and processes," *Genome Biology*, 8, R39.

Wang, C. (2007), "Variational Bayesian Approach to Canonical Correlation Analysis," *IEEE Transactions on Neural Networks*, 18, 905 –910.

Wang, Y., Krivtsov, A. V., Sinha, A. U., North, T. E., Goessling, W., Feng, Z., Zon, L. I., and Armstrong, S. A. (2010), "The Wnt/β-Catenin Pathway Is Required for the Development of Leukemia Stem Cells in AML," *Science*, 327, 1650–1653.

Wesolowski, S., Birtwistle, M. R., and Rempala, G. A. (2013), "A Comparison of Methods for RNA-Seq Differential Expression Analysis and a New Empirical Bayes Approach," *Biosensors*, 3, 238–258.

Wilusz, M. and Majka, M. (2008), "Role of the Wnt/β-catenin network in regulating hematopoiesis," *Archivum Immunologiae et Therapiae Experimentalis*, 56, 257–266.

Wu, H., Chen, Y., Liang, J., Shi, B., Wu, G., Zhang, Y., Wang, D., Li, R., Yi, X., Zhang, H., Sun, L., and Shang, Y. (2005), "Hypomethylation-linked activation of PAX2 mediates tamoxifen-stimulated endometrial carcinogenesis," *Nature*, 438, 981–987.

Zhang, L., Huang, J., Yang, N., Greshock, J., Megraw, M. S., Giannakakis, A., Liang, S., Naylor, T. L., Barchetti, A., Ward, M. R., Yao, G., Medina, A., OBrien-Jenkins, A., Katsaros, D., Hatzigeorgiou, A., Gimotty, P. A., Weber, B. L., and Coukos, G. (2006), "microRNAs exhibit high frequency genomic alterations in human cancer," *Proceedings of the National Academy of Sciences*, 103, 9136–9141.

Zheng, L. and Lucas, J. (2012), *Aneuploidy in Health and Disease*, chap. Uncover Cancer Genomics by Jointly Analysing Aneuploidy and Gene Expression, pp. 22–41, InTech.

Zhou, M., Hannah, L., Dunson, D., and Carin, L. (2012), "Beta-Negative Binomial Process and Poisson Factor Analysis," *AISTATS*.

Biography

Lingling Zheng was born on November 30, 1985 in Wenzhou, a place best known for its private enterprises in China. She received her B.S. in biotechnology from Zhejiang University, Hangzhou, China in 2008. In the fall of the same year, she enrolled in the Computational Biology and Bioinformatics program at Duke University as a Ph.D. student. She is expected to receive her Ph.D. from Duke University in December 2013.

Publication

- Chapter 2

  Zheng, L. and Lucas, J. (2012), *Aneuploidy in Health and Disease*, book chapter. Uncover Cancer Genomics by Jointly Analyzing Aneuploidy and Gene Expression, pp. 22-41, no. 2, InTech. available at http://www.intechopen.com/books/aneuploidy-in-health-and-disease/ joint-analysis-of-aneuploidy-and-gene-expression

Unpublished

- Chapter 4

  Zheng, L., Yan, X., Suchindran, S., Dressman, H., Chute, J. and Lucas, J. (2013), *Statistical Applications in Genetics and Molecular Biology*, Biological Pathway Selection through Bayesian Integrative Modeling, *under review*.

- Chapter 3

Ray, P., **Zheng, L.**, Wang, Y., Lucas, J., Dunson, D., and Carin, L. (2012), *Bayesian Analysis*, Bayesian Joint Analysis of Heterogeneous Data, *resubmission in preparation.*

- **Zheng, L.**, Chen, M., Carin, L., and Lucas, J. (2011), *BMC Bioinformatics*, Bayesian Elastic Net for Multi-Class Classification and Survival Analysis, *resubmission in preparation.*