Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Ben Li Date

Novel Model-based Methods for High-throughput Genomics Data Analysis

By

Ben Li

Doctor of Philosophy

Biostatistics

Zhaohui Steve Qin, Ph.D.

Advisor

Hao Wu, Ph.D.

Advisor

Lance A. Waller, Ph.D.

Committee Member

Timothy D. Read, Ph.D.

Committee Member

Tianwei Yu, Ph.D.

Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Novel Model-based Methods for High-throughput Genomics Data Analysis

By

Ben Li

M.S., Emory University, 2016

B.S., Nanjing University, 2012

Advisor: Zhaohui Steve Qin, Ph.D. and Hao Wu, Ph.D.

An abstract of

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University

in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics

2017

Abstract

Novel Model-based Methods for High-throughput Genomics Data Analysis

By Ben Li

In this dissertation, I propose three model-based methods for improving genomics data analysis by utilizing existing external datasets ("Historical Data").

In the first topic, I propose a Bayesian inference framework with historical data-based informative priors to improve detection of differentially expressed (DE) genes. To evaluate the feasibility and effectiveness of my Bayesian framework, I use a normal-inv$-\chi^2$ model on gene expression microarray data and Bayes factors (BF) are calculated to rank the top DE genes. Extensive real data-based simulations and real data analyses are conducted to illustrate the advantages of the proposed method.

In my second topic, I propose rank-based strategies to incorporating historical information into new experimental datasets. Ranks from historical data are used to determine groups or windows for new experimental datasets. I also propose a group dividing metric (GDM) to determine the optimal number of groups or size of windows. Through real data-based simulations and real data analysis, I demonstrate that proposed strategies can be easily applied to gene expression microarray data and methylation array data. I also showed the potential of borrowing information across different platforms for the proposed method by applying new strategies to BS-Seq data.

In the third topic, I propose a two-step strategy to summarize and borrow information from historical data by "gene panels". In the first step, I use a penalized EM algorithm to define gene panels, which summarizing information of target gene, from historical data. In the second step, tasks could be accomplished with better accuracy or previously impossible tasks could be possible when incorporating gene panels. By simulation studies and real data examples, I demonstrate that the use of gene panels improves data analytics results in detecting DE genes, especially with extremely few or no replicates available.

Novel Model-based Methods for High-throughput Genomics Data Analysis

By

Ben Li

M.S., Emory University, 2016

B.S., Nanjing University, 2012

Advisor: Zhaohui Steve Qin, Ph.D. and Hao Wu, Ph.D.

A dissertation submitted to the Faculty of the

James T. Laney School of Graduate Studies of Emory University

in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics

2017

Acknowledgements

I would like to thank my advisors, Dr. Zhaohui Steve Qin and Dr. Hao Wu, for shedding light upon how to become an outstanding researcher, for encouraging me to explore research ideas, for guiding me to be a better person and for supporting me about academics and for supporting me unreservedly during my time at Emory. They are my lifetime mentors and role models.

I would like to thank my dissertation committee members, Dr. Lance A. Waller, Dr. Timothy D. Read and Dr. Tianwei Yu, for their invaluable comments, constructive criticism and insightful suggestions. They helped me polish up my dissertation work significantly.

I would like to thank all members of Bioinformatics Interest Group (BIG) – in particular, Dr. Li Chen, Tianlei Xu and Hao Feng, who provided helpful suggestions and consulting for a variety of research topics. I would also like to extend my gratitude to all the people in the biostatistics and bioinformatics department at Emory. I could not complete this work without their continuous support.

I would also like to thank my parents for their support and understanding. I dedicate this work to them.

Finally, I want to take a chance to salute to the era. It may not be the best time nor the worst time. But it is a time you can see things differently and you can dream to change the world. The best time is yet to come.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Recent advancements in high-throughput experiments such as gene expression microarrays, methylation arrays, RNA-Seq and BS-Seq have provided abundant information and extensive resources for biomedical researchers studying genetics, genomics and other biomedical fields. These high-throughput technologies have become indispensable tools in a variety of biomedical research areas. These technologies are able to generate a rich set of information for each biological sample, which can be summarized into a comprehensive picture of the underlying biological processes or systems. However, due to the relatively high cost and complexity in sample preparation, the number of samples surveyed in each experiment is much smaller than the number of features surveyed in each sample. The key characteristic of such dataset can be summarized as 'large p, small n' (Fan & Lv, 2010). This presents tremendous challenges when conducting statistical inference on these data, e.g. to detect DE genes and find differential methylated loci (DML) which are fundamental problems in genomic data analysis affecting downstream analysis. Many traditional statistical methods have been modified to tackle this problem and statistical inferences have been improved. Nevertheless, without a rich set of historical data, all existing practical methods only improve inferences from current datasets. With further accumulated publicly available datasets in the big data era, I believe incorporating the information from historical data could lead to practical methods greatly improving statistical inferences on genomics

data. Therefore, this dissertation is dedicated to developing novel practical model-based methods that can reasonably incorporate historical data to improve genomic data analysis.

1.2 Literature Review

In this dissertation, I apply my Bayesian framework to both gene expression and DNA methylation data. Both types of data could be obtained from array or sequencing technology. Here I introduce data formats and review existing methods for gene expression and DNA methylation, respectively.

1.2.1 Gene Expression

Gene expression is the process of using a gene's information in the synthesis of a functional gene product (usually a protein). Gene expression is the most fundamental level in genetics since the genotype gives rise to the phenotype through gene expression. By the definition of gene expression, the amount of functional gene products (usually proteins) should be measured. However, the measurement for functional gene products is difficult and often the abundance of messenger RNA (mRNA), an intermediate product positively correlated with functional gene product, is measured to determine the intensity for gene expression.

Gene expression microarray and RNA-Seq are actually two different methods measuring the intensity of mRNA. Gene expression microarray uses an "array", a collection of microscopic DNA spots attached to a solid surface, to hybridize cDNAs (converted from mRNAs) for target genes. On the other hand, RNA-Seq sequences cDNA and all the sequence fragments (Reads) will be aligned to a reference genome to reflect the intensity of gene expression. Preprocessing and normalization are extremely important

topics for both microarray and RNA-Seq data (Ghosh & Qin, 2010). In this dissertation, we skip the preprocessing and normalization steps and focus on the downstream analyses, assuming our data have already been properly preprocessed and normalized.

After appropriate preprocessing and normalization, microarray data can be summarized into an I by K matrix that stores log transformed gene expression levels across I genes and K samples.

Basic statistical framework: Microarray analysis

An important task of analyzing gene expression data from different conditions is to identify DE genes in an experiment that compares two groups (conditions) of samples. We define the two groups as the control group and the treatment group. Let X_{ijk} denotes the normalized log expression value, where i denotes different genes, j denotes different conditions (control group or treatment group), k denotes different replicates. i = 1, 2…I, j = 1, 2. k = 1, 2….n. The basic assumption for the log gene expression value is:

$$X_{ijk}|\mu_{i,j}, \sigma_i^2 \sim N(\mu_{i,j}, \sigma_i^2) \quad (1)$$

where $\mu_{i,j}$ denotes the mean for the ith gene in the jth group and σ_i^2 is the variance for the ith gene. We test whether the mean expression for a certain gene is significantly different between the two groups. For the ith gene, the hypotheses are: $H_0: \mu_{i,1} = \mu_{i,2}$ versus $H_A: \mu_{i,1} \neq \mu_{i,2}$.

A natural statistical tool for detecting DE genes is to apply the two sample Student's t-test to each gene and calculate the t statistics: $t_i = (\overline{X}_{i1} - \overline{X}_{i2}) / \sqrt{(S_{i1}^2 + S_{i2}^2)/n}$ where $\overline{X}_{ij}$ and S_{ij}^2 are the sample mean and variance of X_{ijk}. Genes can be ranked by their t statistics and DE genes are defined by associated p-values. However, the limited sample size in

microarray studies may lead to underestimation of SDs yielding an increase in false positives for DE genes. To overcome this, various methods have been proposed striving to obtain a more accurate estimate of SDs. An adjusted *t*-test then will be conducted after substituting for regular variance by adjusted variance in the student's *t*-test: $t_i^* = (\overline{X}_{i1} - \overline{X}_{i2})/SE^*$. Here SE^* denotes adjusted variance produced from different methods. We next survey two widely used state-of-the-art methods.

Significance Analysis of Microarrays (SAM)

SAM was proposed by Tusher, Tibshirani, and Chu (2001). To avoid the problems caused by inaccurate SD estimates, SAM attempts to remove or minimize the test statistics' dependence on variances by adding a small constant to adjust the estimated variance when performing student's *t*-test.

Linear Models for Microarray Data (Limma)

Limma, an empirical Bayesian method, utilizes such standard hierarchical model to borrow information from other genes so that the estimate of variance can be improved (Smyth, 2004). In essence, the variance estimate for each gene can be regarded as the weighted average of the sample variance of this gene and the overall sample variance observed across all genes. The underlying assumption is that all genes share some commonalities, so much so that the prior distributions of the model parameters of their gene expression values can be regarded as random samples from a single distribution (hyper-prior).

Basic statistical framework: RNA-Seq

RNA-Seq is regarded as a better alternative for cost effective microarray analysis if not considering the cost of experiments themselves since RNA-Seq could be more accurate

and can provide addition information for gene fusion, alternative splicing, etc. In this dissertation, I only focus on DE gene detection and will not discuss the additional information from RNA-Seq. For DE gene detection purposes, RNA-Seq data can also be summarized into an I by K matrix after proper preprocessing. Different assumptions relating to the data lead to different methods. Some methods assume that the logarithms of reads/fragments per kilobase of gene per million mapped reads (RPKM/FPKM) or transcript per million (TPM) follow normal distribution. Hence, all methods for Microarray data with normal assumption can be easily modified for RNA-Seq. The most popular RNA-Seq data analysis method based on a normal assumption is limma-voom (Law, Chen, Shi, & Smyth, 2014) and will be discussed in the next section. Many other methods work with raw reads directly. These methods assume the raw reads follow negative binomial distributions. Based on negative binomial distribution, RNA-Seq could gain additional information to help further inference from the relationship between genes means and variances. One important difference worth noting between Microarray and RNA-Seq is that Microarray only covers some genes in the genome while RNA-Seq covers genes across the whole genome. The difference in coverage could cause matching issues when comparing results between microarray and RNA-Seq.

We will survey these state-of-the-art methods below. We keep notations the same as it is in the method's original paper for consistency. However, the same notations may have different definitions between different methods and what we proposed in page 3.

Limma-voom

Limma-voom uses its "variance modeling at the observational level" (voom) method to estimate the mean-variance relationship of the log-counts (Law et al., 2014). This will

assign a precision weight for each observation and then the Limma pipeline for Microarray can be used for RNA-Seq. The core part of limma-voom is to use a linear model to account for RNA-seq experiments' arbitrary complexity including multiple treatment factors, batch effects or other related numerical covariates:

$$E(y_{gi}) = \mu_{gi} = x_i^T \beta_g$$

where y_{gi} is the log-counts per million (log-cpm): $y_{gi} = \log_2 \frac{r_{gi}+0.5}{\sum_{g=1}^{G} r_{gi}} \times 10^6$, x_i is a vector of covariates indicating experimental design or other factors and β_g is a vector of unknown coefficients representing log2-fold-changes between different experimental conditions.

DEseq

DEseq assumes that counts follow Negative Binomial (NB) distribution (Anders & Huber, 2010).

$$K_{ij} \sim NB(mean = \mu_{ij}, variance = \sigma_{ij}^2)$$
$$\mu_{ij} = q_{i,\rho(j)} s_j$$
$$\sigma_{ij}^2 = \mu_{ij} + s_j^2 v_{i,\rho(j)}$$

The variance for the NB distribution is assumed to be the sum of shot noise term μ_{ij} and raw variance term $s_j^2 v_{i,\rho(j)}$ while the raw variance $v_{i,\rho(j)}$ is estimated using the mean and variance relationship: a smooth function of the mean: $v_{i,\rho(j)} = v_\rho(q_{i,\rho(j)})$.

DEseq2

DEseq2 is a disruptively improved version of DEseq (Love, Huber, & Anders, 2014) in terms of its stability and interpretability of estimates. Although additional shrinkage techniques involved, basic assumptions are similar to DEseq:

$$K_{ij} \sim NB(mean = \mu_{ij}, dispersion = \alpha_i)$$
$$\mu_{ij} = s_{ij} q_{ij}$$
$$\log(q_{ij}) = \sum_r x_{jr} \beta_{ir}$$

K_{ij} is the read count for gene *i* in sample *j*, which is assumed to follow a negative binomial distribution and further modeled with a generalized linear model.

edgeR

edgeR also assumes that counts follow Negative Binomial distribution (Robinson, McCarthy, & Smyth, 2010). It uses empirical Bayes to "shrink" estimations for variance which is the same idea Limma uses for Microarray. An empirical Bayes procedure is used for the shrinkage.

DSS

DSS also starts analysis with raw counts and assume the counts follow NB distributions (H. Wu, Wang, & Wu, 2013). DSS is based on noting that the shrinkages in DEseq and edgeR are too strong and presented a new empirical Bayes shrinkage estimate for the dispersion parameters to improve DE detection:

$$Y_{gi} | \theta_{gi} \sim Poisson(\theta_{gi} s_i)$$
$$\theta_{gi} | \phi_g \sim Gamma(\mu_{g,k(i)}, \phi_g)$$
$$\phi_g \sim log - normal(m_0, \tau^2)$$

1.2.2 DNA methylation

DNA methylation indicates the process that methyl groups are added to DNA. This process serves as the foundation of epigenomics. There are many different technologies for obtaining DNA methylation information including many different array and sequencing

based technologies. In this dissertation, I will focus on the widely used Infinium HumanMethylation450 BeadChip array and Bisulfite sequencing (BS-Seq). Similar to gene expression, array-based methods for DNA methylation cover fewer CpG sites than BS-Seq but are more cost-effective. The expensive BS-Seq can cover more sites and is regarded as a more accurate method.

Basic statistical framework: Methylation Array

Array-based approaches rely on bisulfite treatment of DNA converting unmethylated cytosines to uracils and keeping 5-methylcytosines unaffected. The converted uracils are amplified as thymines during subsequent amplification and then the bisulfite-treated DNA can be quantitatively measured to assess the proportion of DNA methylation levels in each sample at single-CpG resolution. After proper preprocessing and normalization, each of these arrays allows for the estimation of a methylated (M) and an unmethylated (U) signal intensities. Then these signals can be used to calculate the β-value: M/(M+U). The β-value indicates the proportion of methylated cells in a sample. The β-value can be further converted into the M-value, which is logit transform of the β-value. Hence, Methylation array data can also be summarized into an I*K matrix of β-values or M-values. Each row of the matrix indicates a CpG site while each column indicates a sample. Similar to gene expression, finding differential methylated loci (DML) is also an important task for DNA methylation data analysis (Qin et al., 2016). Since the log of M-values can be reasonably assumed to be normal distributed, all methods relating to gene expression microarray with normal assumption can be applied on methylation array data. Here we introduce a typical method dealing with methylation array:

Minfi

Minfi is a suite of computational tools for preprocessing, quality assessment and detection of DML. Since this dissertation focuses on detection of DML, I focus on Minfi's detection of DML. Minfi provides two different options for users. Minfi runs an *F*-test on multiple groups or conditions if one does not wish to shrink the variances. This is equivalent to a *t*-test for two-group comparison. Otherwise Minfi would run Limma's procedure to shrink the variances.

Basic statistical framework: BS-Seq

While array-based methods only cover part of CpG sites in the genome, BS-seq could cover the whole genome and produce single-base resolution information about the methylation status for the entire genome. For each CpG site, one obtains two numbers: one counts the occurrences of methylated reads and the other counts the unmethylated reads at a certain CpG site. Hence for I CpG sites and K samples, one will obtain an I*(2K) matrix after proper preprocessing and normalization. This characteristic makes BS-Seq have a quite different basic statistical framework than the preceding methods. One can summarize the data from two conditions into a 2-by-2 table for each CpG site and apply a χ^2-test or Fisher's exact test for each of the 2-by-2 tables. An alternative is to use a Beta-binomial model (H. Wu et al., 2015). Here we review the model from dispersion shrinakge for sequencing data (DSS):

DSS

The sequencing counts of BS-Seq are described by a lognormal-beta-binomial hierarchical model (Feng, Conneely, & Wu, 2014). The hierarchical model helps stabilize the variance

estimate for each CpG site. A Wald test is then used for hypothesis testing at each CpG site.

1.2.3 Hierarchical Models

Most of state-of-the-art methods mentioned above use a hierarchical model structure to stabilize SD estimates and the topics in this dissertation are built based on such hierarchical models. Therefore, I review hierarchical models in this section.

Hierarchical models (Good, 1965), which are conceptually related to regularization techniques (Hastie, Tibshirani, & Friedman, 2009), can be a valuable statistical tool for addressing "large *p*, small *n*" problems. A variety of efforts have been made by statisticians to show the effectiveness of hierarchical models in the analysis of microarray gene expression data (Kerr & Churchill, 2001; Newton, Kendziorski, Richmond, Blattner, & Tsui, 2001; Parmigiani, Garett, Irizarry, & Zeger, 2003; Smyth, 2004). In addition, the genomics research community, facing massive datasets produced by high-throughput technologies, has enthusiastically embraced hierarchical models (Ji & Liu, 2010). Examples of hierarchical model applications include Limma (Smyth, 2004) for Microarray, edgeR (Robinson et al., 2010), DSS (H. Wu et al., 2013) for RNA-seq, TileMap (Ji & Wong, 2005) for ChIP-chip, Minfi (Aryee et al., 2014) and DSS-single (H. Wu et al., 2015) for methylation array and whole genome bisulphite sequencing WGBS data.

The key benefit of the hierarchical model lies in the fact that it enables "borrowing" information from other features (e.g. genes/probes in gene expression microarray or CpG sites in methylation array) to stabilize and improve the inference results for individual features. Such a strategy has been shown to be much more reliable over naïve inferences

especially when the sample size is limited, thus leads to more accurate downstream analyses.

For completeness, I review a typical hierarchical model for gene expression microarray data. The probability models for gene expression values under the two conditions can be written as follows: (Note that the following models are adapted from the ones originally proposed by Ji and Wong for modeling tiling array data (Ji & Wong, 2005)):

$$X_{ijk} \ |\mu_{i,j}, \sigma_i^2 \sim N(\mu_{i,j}, \sigma_i^2) \quad (2)$$

$$\mu_{ij} \mid \mu_0, \tau_0^2 \ \propto 1 \quad (3)$$

$$\sigma_i^2 \mid \nu_0, \omega_0^2 \sim \text{Inv} - \chi^2(\nu_0, \omega_0^2) \quad (4)$$

X_{ijk} denotes the normalized gene expression values, where *i* denotes different genes, *j* denotes different conditions, *k* denotes different replicates. $i = 1, 2 \ldots I$, $j = 1, 2$. $k = 1, 2 \ldots n$. The mean parameter μ_{ij} is assumed to be uniform and variance parameter σ_i^2 is assumed to follow an inverse-χ^2 distribution with hyper-parameters ν_0 and ω_0^2. An empirical Bayes shrinkage estimator for σ_i^2 is then used as the variance estimator $\widehat{\sigma_i^2}$:

$$\widehat{\sigma_i^2} = \left(1 - \hat{B}\right) s_i^2 + \hat{B}\overline{s^2}$$

$$\hat{B} = \frac{2/v}{1 + 2/v} \frac{I - 1}{I} + \frac{1}{1 + 2/v} \left(\frac{2}{v}\right) \left(\overline{s^2}\right)^2 \frac{I - 1}{S}$$

where $\nu = 2(n-1), s_i^2 = 2\sum_k \frac{\left(X_{ijg} - \bar{x}_{ij}\right)^2}{v}$, $\overline{s^2} = \frac{\sum_i s_i^2}{I}, S = \sum_i \left[s_i^2 - \overline{s^2}\right]^2$

$\widehat{\sigma_i^2}$ is then subsequently used to perform an adjusted *t*-test:

$$t_i = \frac{\bar{X}_{i1} - \bar{X}_{i2}}{\hat{\sigma}_i \sqrt{(2/n)}}$$

1.3 Outline

In Chapter 2, Section 2.1 introduces a Bayesian inference framework with historical data-based informative priors including motivation, model building, inferences and tests. Section 2.2 reports simulation results for comparing IPBT with existing state-of-the-art methods. Section 2.3 focuses on real data analysis for further demonstrating the usage of Informative Priors Bayesian Test(IPBT). Discussions and Conclusions are presented in Section 2.4.

In Chapter 3, Section 3.1 introduces stratified hierarchical model (stHM) and sliding window hierarchical model (swHM) including model building, group choosing, inference and test. Section 3.2 reports simulation results for comparing stHM and swHM with IPBT and other existing state-of-the-art methods. Section 3.3 focuses on real data analysis for further demonstrating the usage of stHM and swHM, especially on DNA methylation data and the usage for crossing different platforms. Discussions and Conclusions are presented in Section 3.4.

In Chapter 4, Section 4.1 introduces the concept of gene panels and how to construct them using penalized Expectation–maximization (EM) algorithm. Section 4.2 shows simulation studies validating the consistency of gene panels and examples of applying gene panels on detecting different expressions. Section 4.3 uses real data to demonstrate properties of gene panels and how can they be used in real data analysis.

Chapter 5 concludes the dissertation with summary and discussion about potential future work.

Chapter 2

Bayesian inference with historical data-based informative priors improves detection of differentially expressed genes

2.1 Methods

This section introduces a Bayesian inference framework with historical data-based informative priors including motivation, model building, inference and test.

2.1.1 Motivation

Although hierarchical models stabilize inferences by “shrinking” all the estimates toward their means, they also inevitably bring over-correction problems. For genes whose intrinsic variances are far lower or higher from the mean level, the inferences from hierarchical models could be biased. In fact, the over-correction is not unexpected since all the genes involved in a typical microarray or RNA-Seq study perform rather diversely and the traditional exchangeability assumption of hierarchical models usually does not hold. Therefore, “borrowing” information from all genes (including the ones with different properties) may not be the best strategy and the strategy could be a double-edged sword in many scenarios. It is a reasonable strategy if no additional information except the current experimental data is available. However, in reality, given the explosion of genomic datasets that are publicly available, there is abundant information that can be utilized and should be

considered. A unique and fundamental advantage of the Bayesian inference framework lies in its capability to incorporate existing prior information. Bayesian inference achieves seamless integration of prior knowledge and observed data hence is desirable in solving real practical problems (Gelman, 2004). Because technologies like microarray have been widely adopted, there are plenty of publicly available data (referred to as historical data hereafter). We believe such information should be taken advantage of, and the Bayesian framework provides an attractive avenue for implementing such a strategy. Although historical data have been exploited in other contexts (for example, Sui, et al. (2009) applied a historical database of microarray experiments to adjust background for DNA microarrays), we found none of the existing methods for detecting DE genes explicitly utilizes historical data.

2.1.2 Informative prior Bayesian test (IPBT)

In this topic, I propose an alternative approach for the classical hierarchical model, which is in some sense “perpendicular” to the Bayesian hierarchal model for detecting DE genes. Instead of borrowing information from different genes measured in the same experiment, our proposed approach borrows information from the measurements of the same gene in different experiments conducted in the past, using the same technology, same type of chip, on the same type of cells (or similar). The idea of utilizing past experience can be readily achieved under a Bayesian inference framework in the form of prior distributions.

The key idea of our approach is to specify an informative, gene-specific prior distribution for each gene based on abundant historical data and then conduct Bayesian hypothesis testing. Hence, we name our approach the informative prior Bayesian test

(IPBT). Because different genes have different biological functions, it is often the case that their expression quantities display rather diverse distributions. Therefore, in contrast to the traditional Bayesian hierarchical model, IPBT assumes that each gene has its own unique prior distribution. The full model is:

$$X_{ijk} \mid \mu_{i,j}, \sigma_i^2 \sim N(\mu_{i,j}, \sigma_i^2) \tag{5}$$

$$\mu_{i,j} \mid \mu_{i0}, \frac{\sigma_i^2}{k_i} \sim N(\mu_{i0}, \frac{\sigma_i^2}{k_i}) \tag{6}$$

$$\sigma_i^2 \mid \nu_i, \omega_i^2 \sim \text{Inv} - \chi^2(\nu_i, \omega_i^2) \tag{7}$$

where (μ_{i0}, k_i) and (ν_i, ω_i^2) are the hyper-parameters. The main difference between IPBT

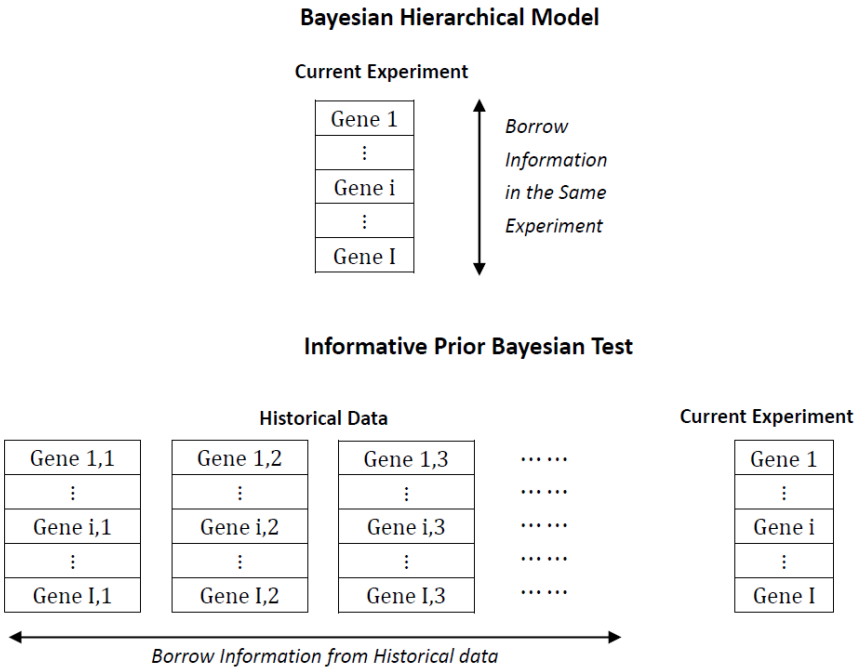

Figure 1 Main differences between IPBT and standard hierarchical model. "Gene i,1", "Gene i,2", "Gene i,3" indicate gene i's expression in first, second, third historical experiment, respectively.

and the hierarchical model in equations (2) to (4) is that here hyper parameters (ν_i, ω_i^2) for the variance $\sigma_i^2 s$ are gene-specific. This gives each gene its specific prior distribution and allows more flexibility.

Figure 1 summarizes the difference between IPBT and traditional Bayesian hierarchical model. In IPBT, the parameter of interest for each gene is σ_i^2 for which we infer using a Bayesian procedure.

The full model (5)-(7) in the main text can be rewritten as follows:

$$X_{ijk} \ |\mu_{i,j}, \sigma_i^2 \sim N(\mu_{i,j}, \sigma_i^2) \tag{8}$$

$$(\mu_{i,j}, \sigma_i^2)|\ \mu_{i0}, k_i, \nu_i, \omega_i^2 \sim N\text{Inv} - \chi^2(\mu_{i0}, k_i, \nu_i, \omega_i^2) \tag{9}$$

$N\text{Inv} - \chi^2$ denotes normal-inverse-χ^2 distribution. There are four hyper-parameters $\mu_{i0}, k_i, \nu_i, \omega_i^2$ for each gene. μ_{i0} is the location parameter for $\mu_{i,j}$ and k_i is how strongly one believes in μ_{i0}. ω_i^2 is the scale parameter for σ_i^2 and ν_i is how strongly one believes in ω_i^2. When equation (9) is written as equation (6) and (7), we can also interpret μ_{i0} and σ_i^2/k_i as $\mu_{i,j}$'s location and scale parameters while ν_i and ω_i^2 as σ_i^2's degrees of freedom and scale parameter. We use the sample size of historical data (n_0) to denote how strongly we believe in μ_{i0} and ω_i^2. That is to say, $k_i = \nu_i = n_0$. We also use *i*th gene's variance from historical data ($S_{i,0}^2$) to estimate ω_i^2: $\widehat{\omega_i^2} = S_{i,0}^2$.

The joint prior distribution of $\mu_{i,j}$ and σ_i^2 is

$$Pr(\mu_{i,j}, \sigma_i^2) = NInv - \chi^2(\mu_{i0}, k_i, \nu_i, \omega_i^2) \tag{10}$$

$$= N\left(\mu_{i,j}\middle|\mu_{i0}, \frac{\sigma_i^2}{k_i}\right) \times \text{Inv} - \chi^2(\sigma_i^2|\nu_i, \omega_i^2) \tag{11}$$

$$= \left[\frac{\sqrt{2\pi}}{\sqrt{k_i}}\Gamma\left(\frac{\nu_i}{2}\right)\left(\frac{2}{\nu_i\omega_i^2}\right)^{\frac{\nu_i}{2}}\right]^{-1} \times \sigma_i^{-1}(\sigma_i^2)^{-\left(\frac{\nu_i}{2}+1\right)} \times \exp\left(-\frac{1}{2\sigma_i^2}\left[\nu_i\omega_i^2 + k_i(\mu_{i0} - \mu_{i,j})^2\right]\right) \tag{12}$$

$Pr(X_{ijk} \mid \mu_{i,j}, \sigma_i^2)$, the likelihood for X_{ijk} given $\mu_{i,j}, \sigma_i^2$ is:

$$\frac{1}{(2\pi)^{\frac{n}{2}}}(\sigma_i^2)^{-\frac{n}{2}}\exp\left(-\frac{1}{2\sigma_i^2}\left[\sum_{i=1}^{n}\left(X_{ijk}-\bar{X}_{ij}\right)^2+n\left(\bar{X}_{ij}-\mu_{i,j}\right)^2\right]\right) \quad (13)$$

where $\bar{X}_{ij}$ denotes sample group mean.

2.1.3 Inference and Testing

Our Bayesian inference framework incorporating historical data with informative priors can be applied to many distributional assumptions. However, most of them can be extremely computational intensive since many of the distributions are not conjugate. Some distributions may be intractable and Markov chain Monte Carlo (MCMC) may be required especially considering the huge number of genes or CpG sites involved in the test. This is one important reason we use normal-inverse-χ^2 distribution in our model: normal-inverse-χ^2 distribution is conjugate and provides a closed form solution. Thus, the inference could avoid using MCMC and be time and computational efficient.

Due to conjugacy, the joint posterior distribution for $\mu_{i,j}$ and σ_i^2 is also normal-inverse-χ^2 distribution, we have:

$$Pr\left(\mu_{i,j},\sigma_i^2|X_{ijk}\right) \quad \propto \quad N\text{Inv}-\chi^2(\mu_{i0}^*,k_i^*,\nu_i^*,\omega_i^{*2}) \quad (14)$$

$$=\sigma_i^{-1}(\sigma_i^2)^{-\left(\frac{\nu_i^*}{2}+1\right)}\exp\left(-\frac{1}{2\sigma_i^2}\left[\nu_i^*\,\omega_i^{*2}+\,k_i^*\left(\mu_{i0}^*-\,\mu_{i,j}\right)^2\right]\right) \quad (15)$$

where $\mu_{i0}^*, k_i^*, \nu_i^*, \omega_i^{*2}$ are parameters for the posterior distributions.

Then we compare corresponding terms for posterior distributions from two different forms, we can obtain the equations (16) to (18) for $\nu_i^*, k_i^*, \mu_{i0}^*$ and ω_i^{*2}:

$$\nu_i^* = \nu_i + n, \qquad k_i^* = k_i + n \quad (16)$$

$$\mu_{i0}^{*} = \frac{k_i}{k_i + n}\mu_{i0} + \frac{n}{k_i + n}\bar{x} \tag{17}$$

$$\omega_i^{*2} = \frac{\nu_i}{\nu_i + n}\omega_i^2 + \frac{n-1}{\nu_i + n}S^2 + \frac{n}{\nu_i + n}\frac{k_i}{k_i + n}(\mu_{i0} - \bar{x})^2 \tag{18}$$

where $\bar{x}$ is the sample mean and S^2 is the sample variance for the current control data.

We then can calculate Bayes factor for *i*th gene as:

$$\mathrm{BF_i} = \frac{Pr(X_{ijk} \mid H_0 \; is \; true)}{Pr(X_{ijk} \mid H_A \; is \; true)} \tag{19}$$

$$= \frac{\int_{-\infty}^{+\infty} Pr(\mu_i^*, \sigma_i^2)\, Pr(X_{i1*}|\mu_i^*, \sigma_i^2)\, Pr(X_{i2*}|\mu_i^*, \sigma_i^2)\, d\mu_i^*}{\int_{-\infty}^{+\infty} Pr(\mu_{i1}, \sigma_i^2)\, Pr(X_{i1*}|\mu_{i1}, \sigma_i^2)\, d\mu_{i1} \int_{-\infty}^{+\infty} Pr(\mu_{i2}, \sigma_i^2)\, Pr(X_{i2*}|\mu_{i2}, \sigma} \tag{20}$$

Detailed formulas can be obtained after plugging (12) and (13) into (20) and proper algebraic manipulation. For computational purposes, we use $\log(BF_i)$ in IPBT to rank genes:

$$\log(BF_i) = \frac{1}{2\widehat{\sigma_i^2}}\left[k_i\mu_{i0}^2 + \frac{2n^2\bar{X}_{i1}\bar{X}_{i2} - \mu_{i0}^2 k_i^2}{2n + k_i} - \frac{n}{(2n + k_i)(n + k_i)}\Delta_{\mathrm{i1}}\right] + \Delta_{i2} \tag{21}$$

where $\Delta_{\mathrm{i1}} = n^2(\bar{X}_{i1}^2 + \bar{X}_{i2}^2) + 2n\mu_{i0}k_i(\bar{X}_{i1} + \bar{X}_{i2}) + 2\mu_{i0}^2k_i^2$ and $\Delta_{\mathrm{i2}} = \frac{1}{2}\log(\frac{k_i}{2n+k_i}) - \log\left(\frac{k_i}{n+k_i}\right)$. The posterior mean $\nu_i^*/(\nu_i^* - 2)\omega_i^{*2}$ for σ_i^2 is used for $\widehat{\sigma_i^2}$ in (21).

Most widely used state-of-the-art methods adopt an adjusted *t*-test for detection DE genes. We also perform statistical hypothesis testing to detect DE genes in the form of student's *t*-test (with adjusted variance estimates) in IPBT to allow a direct and fair performance comparison with other existing methods. The test statistics is:

$$t_i^* = \frac{\bar{X}_{i1} - \bar{X}_{i2}}{\sqrt{\frac{2\widehat{\nu_i^*}/(\widehat{\nu_i^*} - 2)}{n}\,\widehat{\omega_i^*}^2}} \tag{22}$$

where $\bar{X}_{i1}$ and $\bar{X}_{i2}$ are sample means for control and treatment group, respectively. $\widehat{\nu_i^*}$ and $\widehat{\omega_i^*}$ are estimates of the posterior distribution parameters.

The adjusted variance estimate is essentially the weighted average of the estimated variances obtained from historical data and current data, respectively. This indicates that IPBT indeed enables natural integration of historical data into the current experiment to assist in DE gene detection. Next, we will show that using adjusted *t*-test is equivalent to using Bayes factor in terms of ranking DE genes in IPBT.

We use the posterior mean $\nu_i^* /(\nu_i^* - 2)\omega_i^{*2}$ as the point estimator for σ_i^2 when calculating test statistics in equation (22). Hence, t_i^* can be further written as:

$$t_i^* = \frac{\bar{X}_{i1} - \bar{X}_{i2}}{\sqrt{\frac{2(n_0+n)/(n_0+n-2)}{n}\left[\frac{n_0}{n_0+n}S_{i,0}^2 + \frac{n-1}{n_0+n}S^2 + \frac{n \quad n_0}{n_0+n n_0+n}(\mu_{i0} - \bar{X}_{i1})^2\right]}} \tag{23}$$

Larger values of $|t_i^*|$ indicates the gene is more likely to be differentially expressed.

We next prove that the Bayes factor version and the adjusted *t*-test version of IPBT are equivalent in terms of ranking DE genes. That is to say, the two different versions of IPBT have exactly the same ranks for all the genes. To prove this, we only need to show that for two different genes *i* and *j* ($i \neq j$), the following condition holds:

$$\log(BF_i) > \log(BF_j) \Leftrightarrow |t_i^*| < |t_j^*| \tag{24}$$

Rewriting (21), we have

$$\log(BF_i) = \frac{1}{2\widehat{\sigma_i^2}}\left[\frac{-n^3(\bar{X}_{i1} - \bar{X}_{i2})^2 + \Delta_{i3}}{(2n + k_i)(n + k_i)}\right] + \Delta_{i2} \tag{25}$$

where $\Delta_{i3} = 2n^2\bar{X}_{i1}\bar{X}_{i2}k_i - 2n^2\mu_{i0}k_i\bar{X}_{i1} - 2n^2\mu_{i0}k_i\bar{X}_{i2} + 2n^2k_i\mu_{i0}^2$

Because the historical data's mean value might not be very consistent with current data and the final results are robust to the plug-in mean estimator, we use the control group's sample mean $\bar{X}_{i1}$ as μ_{i0} in equation (25). Therefore, $\Delta_{i3} = 0$ and we have:

$$\log(BF_i) = \frac{1}{2\widehat{\sigma_i^2}}\left[\frac{-n^3(\bar{X}_{i1}-\bar{X}_{i2})^2}{(2n+k_i)(n+k_i)}\right]+\Delta_{i2} \quad (26)$$

$$= \frac{-n^2}{(2n+k_i)(n+k_i)}t_i^{*2}+\Delta_{i2} \quad (27)$$

We use the sample size of the historical data as k_i. Hence, the k_is are the same for different genes ($k_i = k_j = k_0$). *n*, as the sample size for current experiment, is also the same for different genes. Then we have $\Delta_{i2} = \Delta_{j2}$ since Δ_{i2} only involves k_i and n. Therefore, we have:

$$\log(BF_i) > \log(BF_j) \Leftrightarrow \log(BF_i) - \log(BF_j) > 0 \quad (28)$$

$$\Leftrightarrow \frac{-n^2}{(2n+k_0)(n+k_0)}\left(t_i^{*2} - t_j^{*2}\right) > 0 \quad (29)$$

$$\Leftrightarrow |t_i^*| < |t_j^*| \quad (30)$$

This concludes the proof for equation (24) and shows that two different versions for IPBT are equivalent in terms of ranking DE genes.

2.1.4 Informative Priors

Reliable informative priors are essential for IPBT's performance. Without reliable historical data, it is impossible to obtain informative priors. Even if many publicly available datasets exist in the literature, it remains a difficult task to find and process reliable historical datasets for one's own experimental purpose. Fortunately, Lukk et al. (2010) built a global gene expression map which includes microarray data from 5,372 human samples and contains 369 different tissues, cell lines and disease states. All the samples can be divided in to 4, 15 or 369 different groups of various levels. Among them, we calculate informative priors for 96 out of 369 groups with at least ten samples including

normal solid brain tissue, normal solid heart tissue, etc. The dataset (processed and normalized by robust multiarray analysis (RMA)(Irizarry et al., 2003)) was downloaded from arrayExpress (ID: E-MTAB-62). The 96 groups have a median sample size of 25.5 and a mean sample size of 48. More details about the historical datasets used to calculate informative priors can be found in Table 1 and Table 2.

Table 1 Sample size in each meta-groups

4 categories		**15 meta groups**	
Group	# of samples	Group	# of samples
cell line	1259	blood neoplasm cell line	166
		non neoplastic cell line	262
		solid tissue neoplasm cell line	831
disease	765	blood non neoplastic disease	388
		solid tissue non neoplastic disease	377
neoplasm	2315	breast cancer	672
		germ cell neoplasm	71
		leukemia	567
		nervous system neoplasm	112
		non breast carcinoma	288
		non leukemic blood neoplasm	334
		other neoplasm	167
		sarcoma	104
normal	1033	normal blood	467
		normal solid tissue	566

Table 2 Summary statistics for sample size in 96 groups

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
10.00	17.00	25.50	48.00	41.25	672

2.2 Simulation Study

I conducted four sets of real-data based simulation studies to demonstrate advantages of IPBT over existing methods. 1) In the first simulation study, I used 566 normal solid tissue microarray datasets obtained by Affymetrix GeneChip U133A from the global gene expression map to show a general trend between mean value and SD for genes in microarray. All the following simulation are generated with the parameters obtained from these 566 normal samples. We also show different SD estimates from different methods versus their truth to illustrate the over-shrinkage phenomenon and how IPBT can avoid the over-shrinkage. 2) In the second set of simulation, I show the false discovery rates (FDR) and receiver operating characteristic (ROC) curves for IPBT and competing methods. I also show the consistency of IPBT and other existing methods on independent datasets 3) In the last simulation, I show that IPBT can be robust even if the historical data has some noise.

2.2.1 Simulation Study I: Alleviation of Over-shrinkage

One fundamental hypothesis for IPBT is that the expression value of each gene has its unique distribution which reflects its intrinsic biological properties. For example, when historical data collected under diverse conditions were aggregated together, compounded with limited signal range of microarray technology, measurements of house-keeping genes tend to show high means but relatively small variances across conditions; whereas genes responding to stimuli tend to have large variances since their expression values can go either way. Therefore, assuming proper normalization has been performed across samples, to perform statistical inference, we believe it is perhaps a better strategy to use data that

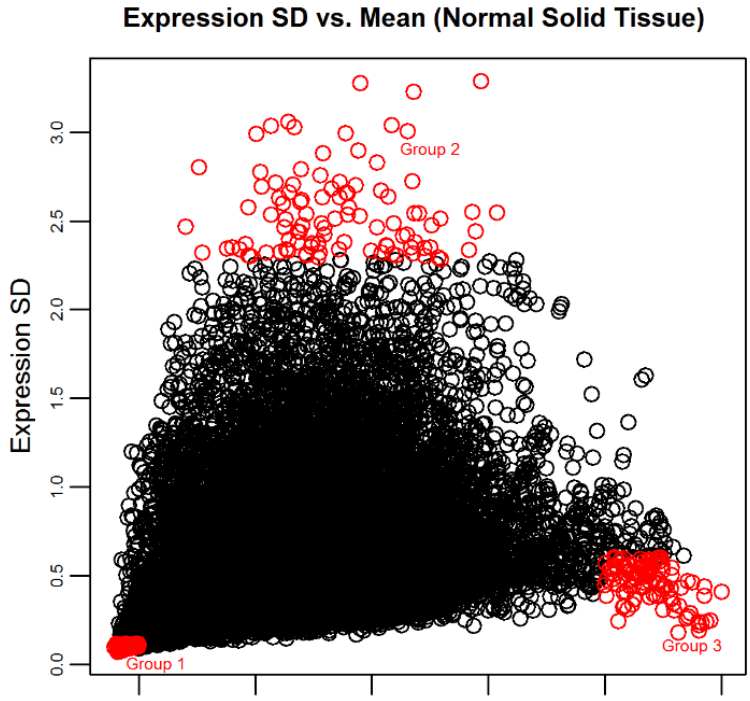

Figure 2 Standard deviation (SD) versus mean for each probe across 566 normal solid tissue samples. The red zones for Group 1, 2, and 3 represent probes with low means and small SDs, probes with mid-level means and large SDs, and probes with high means and small SDs, respectively. The respective GO term enrichment results are presented in the Supplementary Materials.

are collected from different experiments but the same gene, than data collected from the same experiment but different genes. To illustrate the point, using 566 normal solid tissue microarray datasets obtained by Affymetrix GeneChip U133A from the global gene expression map of microarray data, we plot standard deviation versus mean on 22,283 genes of their normalized and log-transformed expression values (Figure 2). We observe a crescent shape in the plot, probes with low or high means tend to have small variance (measured by standard deviation in figures and tables), while genes with mid-level means tend to have large variance.

We choose 100 genes from each of the three spots that correspond to low mean/small variance, mid-level mean/large variance, and high mean/small variance, respectively, and perform a Gene Ontology (GO) (Ashburner et al., 2000) enrichment analysis on each set of the corresponding genes using DAVID (Huang da, Sherman, & Lempicki, 2009a, 2009b). More details can be found in the Tables 3-5 and Figure 3.

Table 3 GO term enrichment analysis, Group 1

Cluster	Enrichment Score	Functional annotation
1	1.10	chromosome organization/chromatin organization/chromatin modification/chromatin binding
2	0.94	purine nucleotide binding/purine ribonucleotide binding/ribonucleotide binding/nucleotide binding/adenyl nucleotide binding/purine nucleoside binding/nucleoside binding/ATP binding/adenyl ribonucleotide binding
3	0.80	ion binding/zinc ion binding/metal ion binding/cation binding/transition metal ion binding
4	0.80	DNA binding/regulation of transcription, DNA-dependent/regulation of RNA metabolic process/regulation of transcription/transcription/ sequence-specific DNA binding/transcription factor activity/transcription regulator activity

Table 4 GO term enrichment analysis, Group 2

Cluster	Enrichment Score	Functional annotation
1	6.94	response to inorganic substance/response to metal ion/response to calcium ion
2	3.59	extracellular matrix organization/extracellular structure organization/extracellular matrix structural

		constituent/peptide cross-linking/collagen fibril organization/growth factor binding/ blood vessel development/vasculature development/platelet-derived growth factor binding/protein binding, bridging/epidermis development/ ectoderm development/skeletal system development/skin development/ integrin binding
3	3.19	response to organic substance/cell adhesion/biological adhesion/cell-cell adhesion
4	2.72	response to steroid hormone stimulus/response to organic substance/response to hormone stimulus/response to endogenous stimulus/response to abiotic stimulus/response to mechanical stimulus/response to extracellular stimulus/ response to nutrient/skeletal system development/response to nutrient levels/ossification/bone development/cartilage development/skeletal system morphogenesis

Table 5 GO term enrichment analysis, Group 3

Cluster	Enrichment Score	Functional annotation
1	44.55	translational elongation/structural constituent of ribosome/translation/structural molecule activity/RNA binding
2	5.66	ribosomal small subunit biogenesis/ribosome biogenesis/ rRNA processing/rRNA metabolic process/ribonucleoprotein complex biogenesis/ncRNA processing/ncRNA metabolic process/RNA processing/erythrocyte homeostasis/homeostasis of number of cells/homeostatic process
3	1.1	response to calcium ion/response to metal ion/response to inorganic substance
4	0.6	anti-apoptosis/negative regulation of apoptosis/negative regulation of programmed cell death/negative regulation of cell death/regulation of apoptosis/regulation of programmed cell death/regulation of cell death/positive regulation of apoptosis/positive regulation of programmed cell death/ positive regulation of cell death

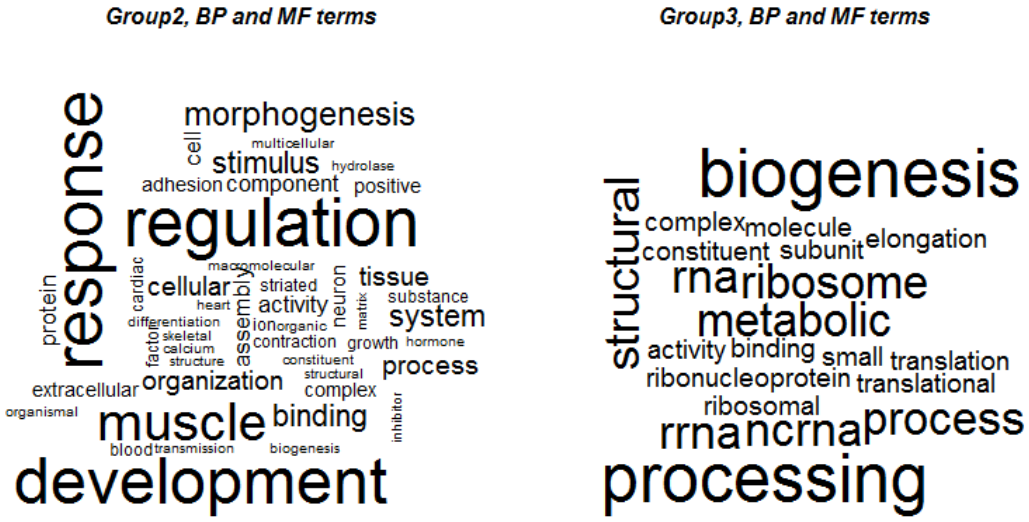

Figure 3 shows the word cloud for GO terms whose P value <= 0.01 in group 2 and group 3, respectively. We do not show such figure for group 1 since there are no terms in group 1 that has p-value less than 0.01. The word cloud supports the conclusion that genes in group 3 are mostly involved in housekeeping activities and genes in group 2 are mostly known for response to stimuli.

The result appears to support our hypothesis. We find that the genes in Group 3 are mostly involved in housekeeping activities evidenced by enriched functional categories such as translation elongation or ribosome-related. Genes in Group 2 are mostly known for being responsive to stimuli. Genes in Group 1 show no functional enrichment, perhaps because they are barely expressed.

Comparison of SD estimation between hierarchical model and IPBT

To illustrate the impact of different methods on genes' variance estimation, we conduct the following simulation study. Using the mean and standard deviation obtained from 566 normal solid tissue samples in the global gene expression map of microarray data, we simulate two samples of expression data and treat them as current control data. We randomly select ten samples from normal solid tissue samples and use them as historical data when estimating standard deviation with IPBT. Figure 4 shows the plots of standard deviations obtained using various methods versus means of the two "current" samples.

Figure 4(a) shows the pre-specified true standard deviation of each gene versus its true

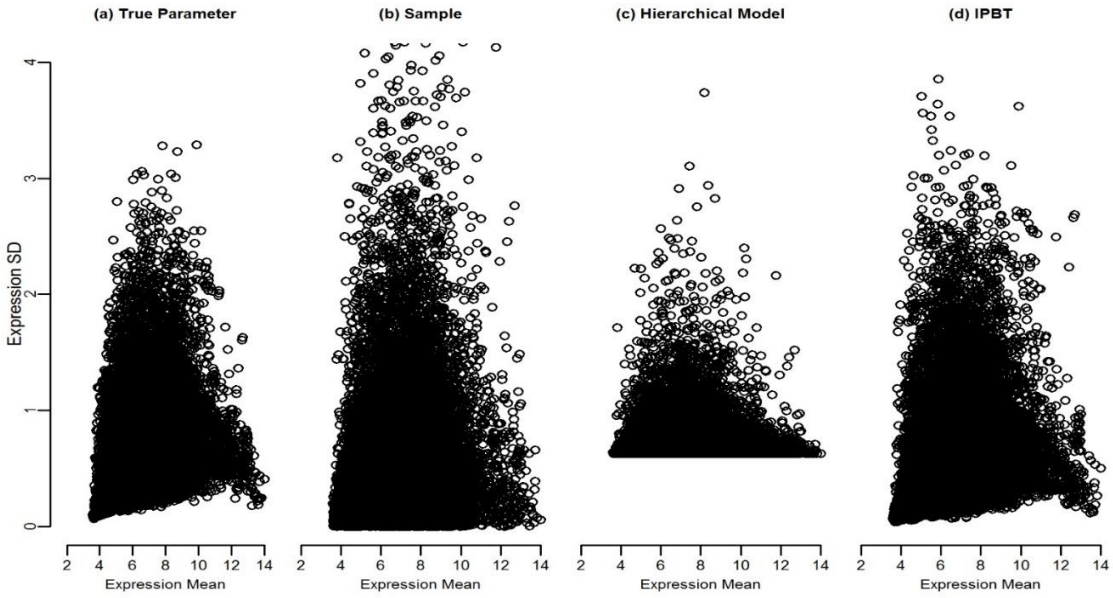

Figure 4 Standard Deviation (SD) Estimates generated from different methods with probes sorted by their true expression mean values. (a) True SDs (b) Sample SD of the current samples (c) SD estimates from Bayesian hierarchical model (d) SD estimates from IPBT

mean expression value. Figure 4(b) shows the sample standard deviations calculated from the two "current" samples, which include extreme small standard deviations caused by limited sample size. Figure 4(c) gives the standard deviations estimated from the Bayesian

hierarchical model, which show shrinkage towards the middle effect compared to Figure 4(b) and clearly suffer from the over-shrinkage problem. Figure 4(d) shows the variance estimates from IPBT, which show little over-shrinkage.

2.2.2 Simulation Study II: DE Gene Detection Performances

Simulation strategy

This simulation study considers 1,000 genes and *k* (ranging from 2 to 5) samples for both the treatment and control groups. We randomly select 10% of the 1,000 genes (i.e. 100 genes) as designated DE genes. Gene expression values in both the treatment and control groups are assumed to follow normal distributions. The distribution parameters are obtained from real data in the global gene expression map. First, 1,000 genes are randomly selected (without replacement) genome-wide. Then for each gene, we derive its sample mean and sample variance from the 566 normal samples in the collection. For the treatment group, the mean and variance of a gene's expression value are assumed to be equal to their counterparts in the control group except for the 100 DE genes for which the mean expression values are set to be two standard deviations higher. For historical data used by IPBT, we first randomly select 188 normal samples out of 566 (without replacement) from the global gene expression map, then obtain their gene expression values corresponding to the 1,000 genes selected earlier.

We compare IPBT with four alternative methods for detecting DE genes: (i) Student's *t*-test, (ii) SAM, achieved by R package "siggenes"; (iii) Limma, achieved by R package "Limma"; and (iv) Z test using the true variance (This is regarded as the best possible method).

DE gene detection result

To evaluate the performance, we calculate the empirical false discovery rate (FDR) (Benjamini & Hochberg, 1995; Tusher et al., 2001) (also known as false discovery proportion--the proportion of incorrect DE calls among all the ones called) from the top 100 genes ranked by the test statistics. The simulation procedure is repeated 500 times for each method. The distributions of the 500 FDRs of the methods are summarized using box plots and shown in Figure 5(a). Our method clearly outperforms all other methods except for the Z test using true variances (considered the gold standard). The performances of our method and Z test are fairly close. Remarkably, the FDR of DE genes detected by IPBT is even smaller than the FDR of DE genes detected by the Student's *t*-test with larger sample size (i.e. increased by one).

We use Receiver Operator Characteristic (ROC) curves to further compare IPBT with the other methods. Figure 5(c) shows a typical ROC curve for one single simulation with two replicates. Detailed area under the curve (AUC) corresponding to Figure 5(c) is listed in Table 6 ("Random Choice" column). The ROC curves again show that IPBT performs better than all the other methods in detecting DE genes except for the Z test, and the performances of our method and the Z test are similar. Additional results can be found in Figure 6 and Table 7.

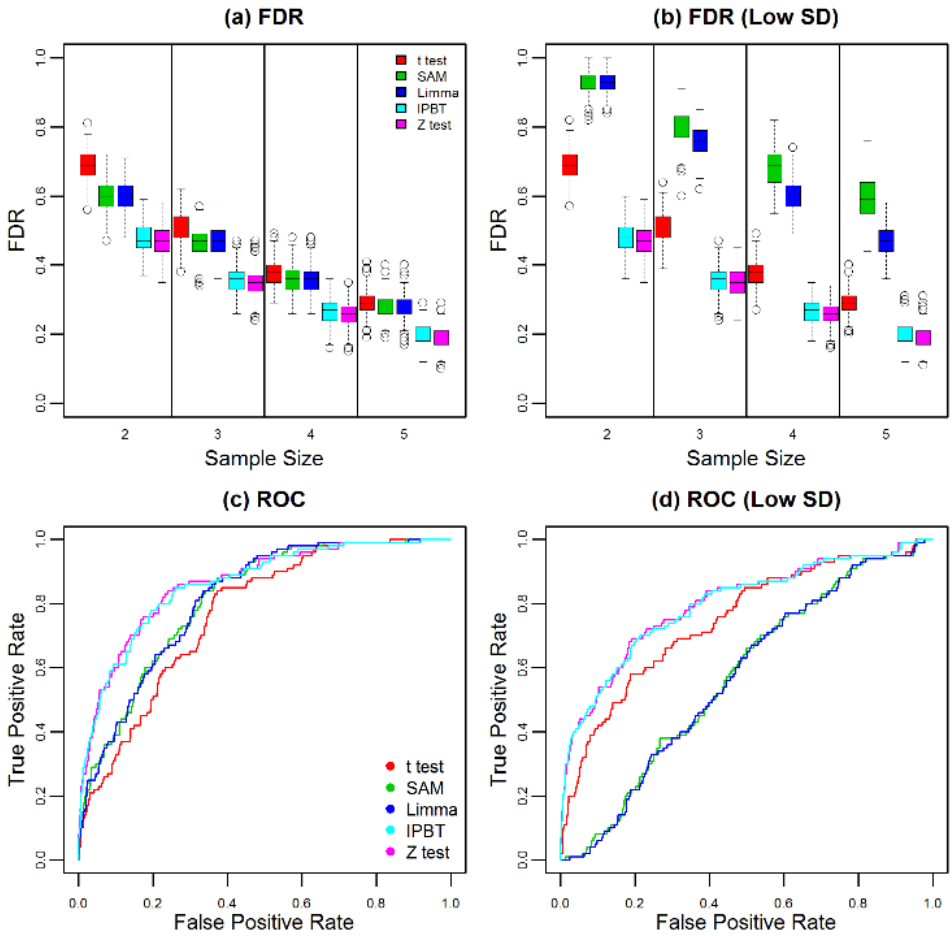

Figure 5 FDR for detecting DE genes comparing various methods with different sample size for (a) random chosen DE genes and (b) low standard deviation DE genes. ROC curves for detecting DE genes comparing different methods in one simulation for (c) random chosen DE genes and (d) low standard deviation DE genes.

Table 6 AUC of Detecting DE Genes Comparing Different Methods in Simulation

Method	Random Choice	Low Variance
student's *t*-test	0.770	0.747
SAM	0.814	0.573
Limma	0.813	0.570
IPBT	**0.861**	**0.798**
Z test	0.864	0.800

* The best results (after excluding Z test) are in **bold**.

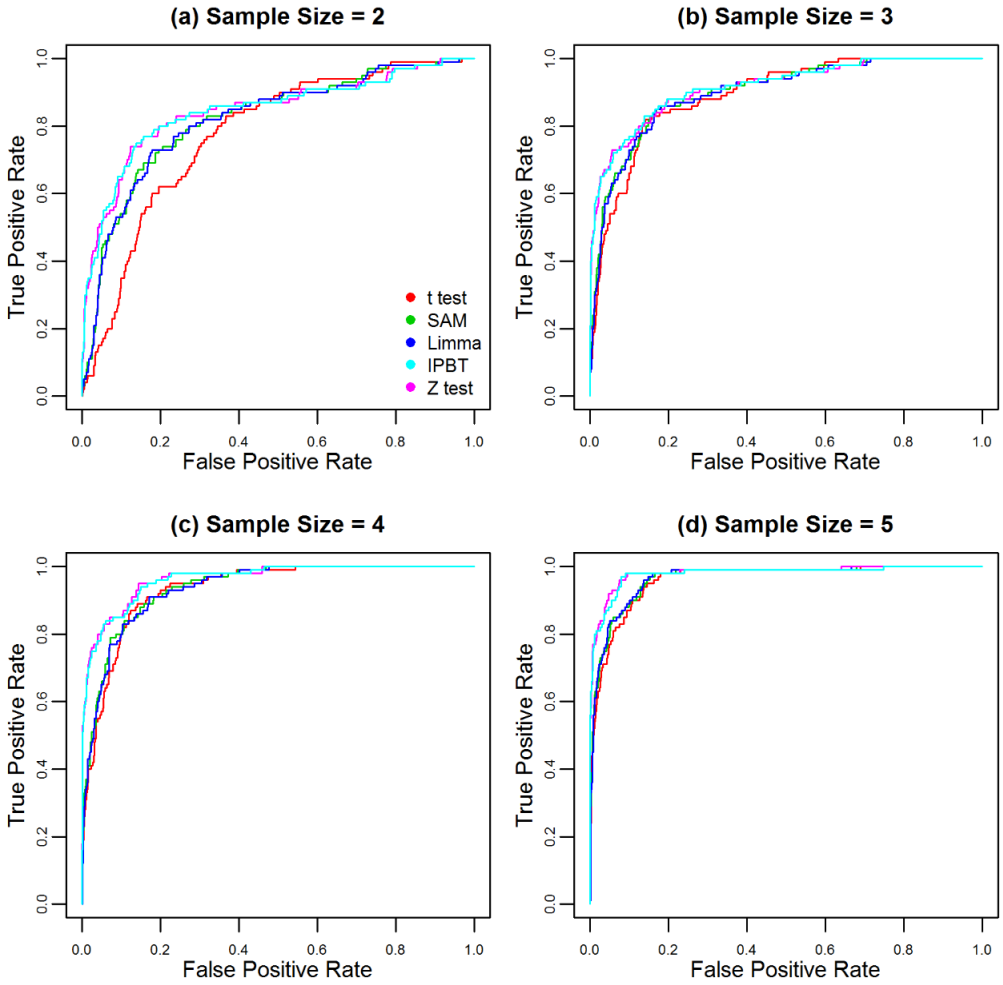

Figure 6 ROC curves for random chosen DE genes with different sample sizes

Table 7 Corresponding AUC for Figure 6 and 7

AUC for Figure 6				
Sample Size	2	3	4	5
student's *t*-test	0.776	0.892	0.933	0.960
SAM	0.820	0.900	0.938	0.965
Limma	0.818	0.898	0.935	0.966
IPBT	**0.845**	**0.915**	**0.962**	**0.979**
Z test	0.846	0.917	0.964	0.981
AUC for Figure 7				
Sample Size	2	3	4	5
student's *t*-test	0.776	0.882	0.947	0.962
SAM	0.614	0.747	0.839	0.899
Limma	0.653	0.794	0.893	0.930
IPBT	**0.870**	**0.925**	**0.970**	**0.978**
Z test	0.871	0.926	0.972	0.980

* The best results (after excluding Z test) are in **bold**.

We further compare the consistency and stability of these methods in detecting DE genes. In each simulation, historical data remain unchanged, but five different sets of the control and treatment data were generated from the same underlying distributions. For each set of control and treatment data, we apply all four methods to detect DE genes. We summarize the number of overlaps among the five lists of DE genes. The simulation procedure is repeated 500 times, and the average number of overlaps is used as a measure of consistency in detecting DE genes. We consider an average number of overlaps greater than or equal to four as an indication of high consistency and greater than or equal to three as moderate consistency. The average numbers of overlaps are reported in Table 8 which again shows that IPBT outperforms other methods except for the Z test in consistency, and the performances of IPBT and the Z test are close.

Table 8 Consistency for Detecting DE Genes

Overlap times	3	4	5	High (4+5)	Moderate (3+4+5)
student's *t*-test	31.20	14.64	2.92	17.56	48.76
SAM	28.43	25.24	10.55	35.79	64.22
Limma	28.63	25.67	10.56	36.23	64.86
IPBT	**33.12**	**30.39**	**11.54**	**41.93**	**75.05**
Z test	33.31	31.47	11.23	42.70	76.01
Overlap times (Low variance)	3	4	5	High (4+5)	Moderate (3+4+5)
student's *t*-test	21.55	6.87	0.83	7.70	29.25
SAM	17.28	5.73	0.83	6.56	23.84
Limma	18.02	6.13	0.93	7.06	25.08
IPBT	**33.57**	**28.67**	**10.69**	**39.36**	**72.93**
Z test	33.28	32.49	12.71	45.20	78.48

* The best results (after excluding Z test) are in **bold**.

Detect DE genes with low intrinsic variance

As Figure 4 demonstrates, Hierarchical model-based methods inflate the variance of the genes which have intrinsic low variance hence lower power to detect DE genes of this kind. IPBT, on the other hand, does not suffer from this shortcoming. To further investigate how over-correction affects the detection of DE genes, we conduct another simulation study

under the scenario that the DE genes have low intrinsic variance, and the results are reported in Figure 5(b), 5(d), Table 6 ("Low Variance" column), and Table 8 (Low variance). All the results show that the standard Bayesian hierarchical model performs even worse than Student's *t*-test, whereas IPBT maintains superior performance that is similar to the performance of the Z test. These results confirm the robustness of IPBT because it avoids the "over-correction" issues for those genes with low intrinsic variance. Additional ROC curves are shown in Figure 7 and Table 7.

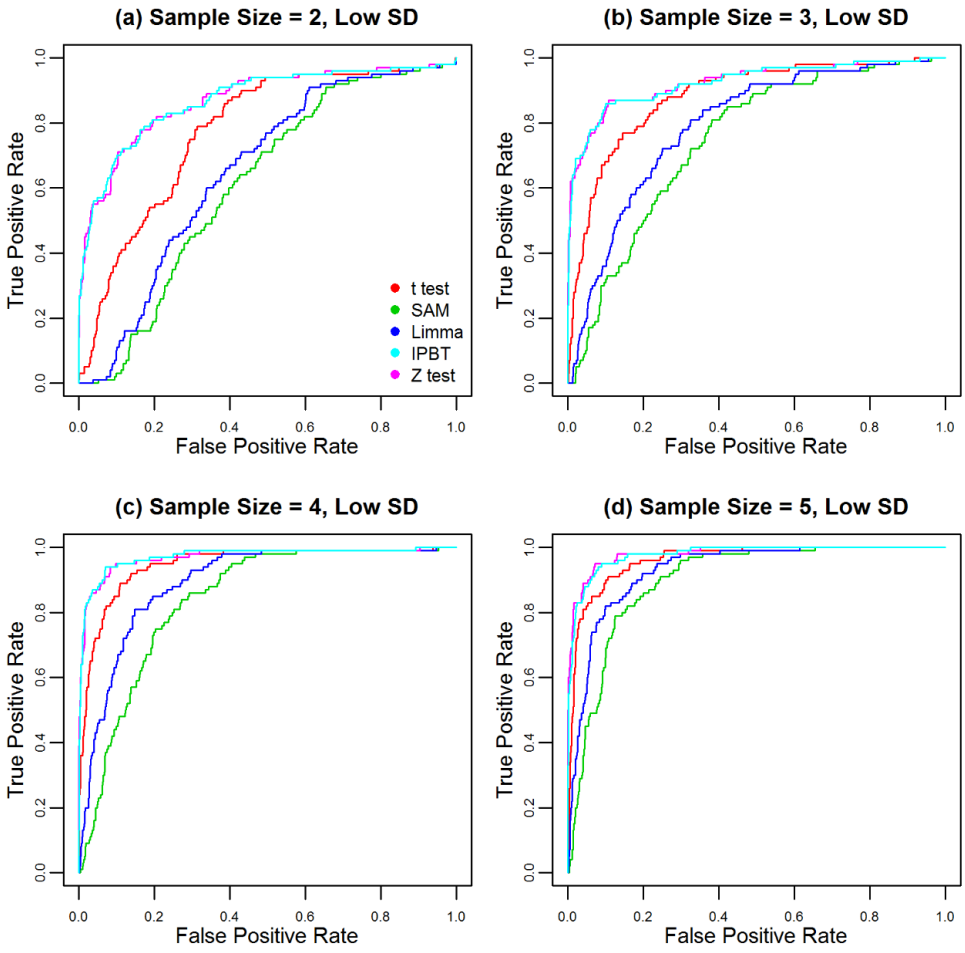

Figure 7 ROC curves for low standard deviation DE genes with different sample sizes

2.2.3 Simulation Study III: Impact of Inaccurate Historical Data

In the previous simulation study, for each gene, we use the same distribution to generate current data and historical data. This represents an idealistic scenario and may not hold true in reality. To examine the robustness of our method, we conduct an additional simulation study in which both parameters in the normal distribution that produces the historical data

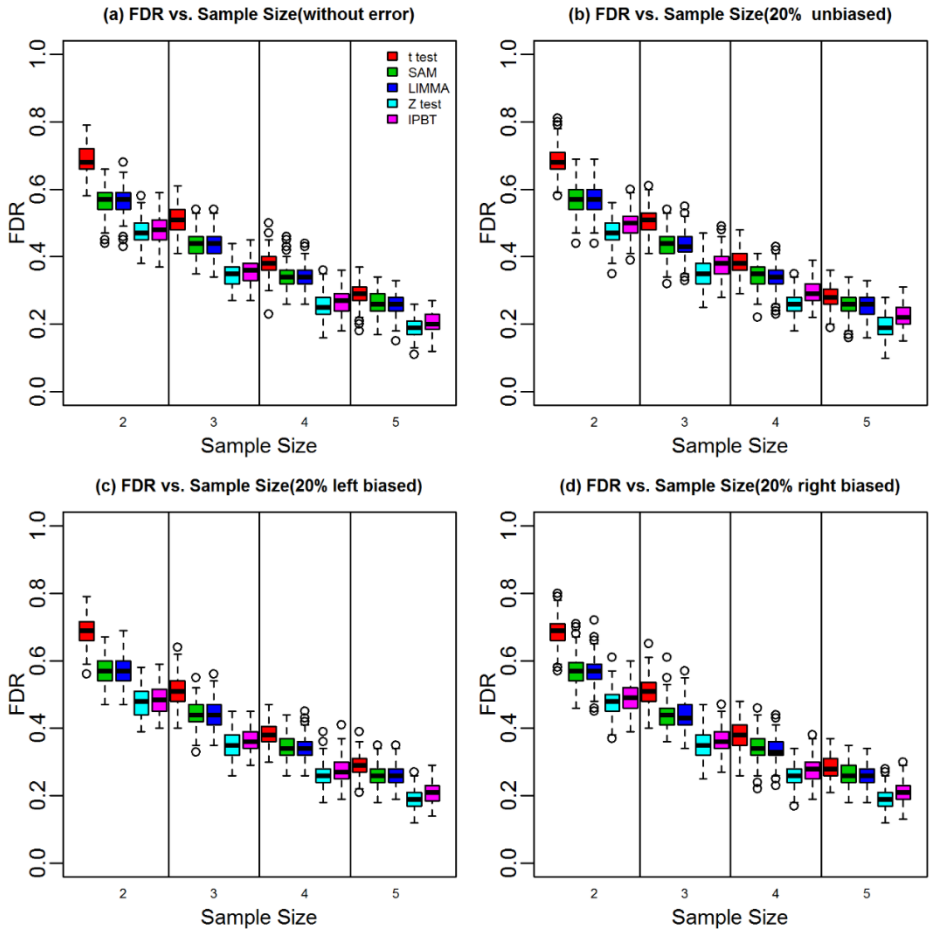

Figure 8 FDR for detecting DE genes using noisy historical data. FDR for detecting DE genes comparing various methods with different sample size using noise historical data of (a) no noise (b) 20% unbiased noise (c) 20% left biased noise (d) 20% right biased noise.

are shifted such that the distributions that generate historical data and current data are no longer identical.

The amount of shift is randomly drawn from a uniform distribution in the interval of (-20%, 20%). We investigate three types of noise-added historical data: unbiased, over-dispersion and under-dispersion. Figure 8 shows that IPBT with noisy historical data still outperforms other methods. Although the performance of IPBT deteriorated when noisy historical data are used, it is still better than Student's *t*-test, SAM and Limma in terms of FDR and is close to the gold standard Z test result in all scenarios. This result demonstrates the robustness of IPBT and implies its broad applicability even with potentially noisy historical data.

2.3 Real Data Analysis

In this section, I will use IPBT in real data analysis for DE gene detection. The historical data to build informative priors are from the global gene expression map. In the first real data analysis, "current" experimental data are also from global gene expression. In the second analysis, "current" experimental data are from Latin Square hgu133a Spike-in experiment.

2.3.1 Real Data Study I: Global Gene Expression Map

Comparison of current and historical data

Our model assumes that historical data are informative for estimating gene expression variance. We validate this using two sets of data from the global gene expression map. One set contains all the data from heart samples, and the other one contains all the data from

brain samples. For each set, we download the raw data (CEL files) and subsequently process and normalize the data using RMA by R package "oligo". For heart data, we randomly choose five normal heart samples (out of 36) and five disease heart samples (out

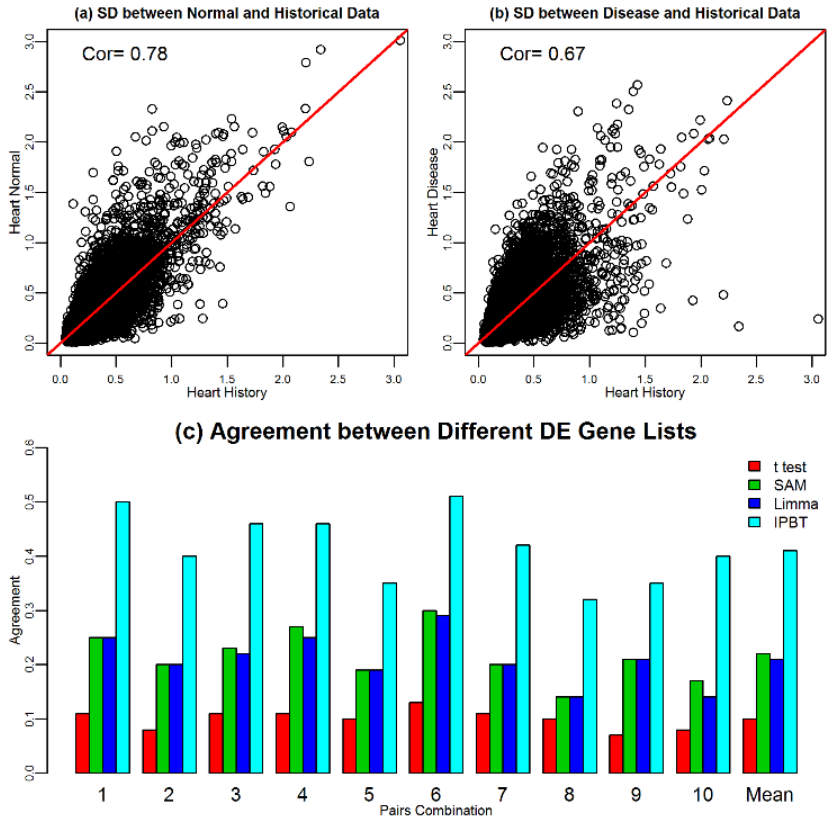

Figure 9 Real data analysis for heart data (a) Comparison of standard deviations (SD) obtained from the five heart normal samples and that obtained from the heart historical data. (b) Comparison of SDs obtained from the five heart disease samples and that obtained from the heart historical data. (c) Agreements be-tween all pair combinations of top 1,000 genes from all 5 DE gene lists.

of 51) and use them as the current data. Data from the 31 remaining normal heart samples are used as historical data. Figures 9(a) and (b) show the standard deviations of the genes in the control group (normal samples) and treatment group (disease samples) against historical data, respectively. The strong positive correlation patterns demonstrated in the

plot confirm that using historical information as informative priors in the inference procedure is feasible.

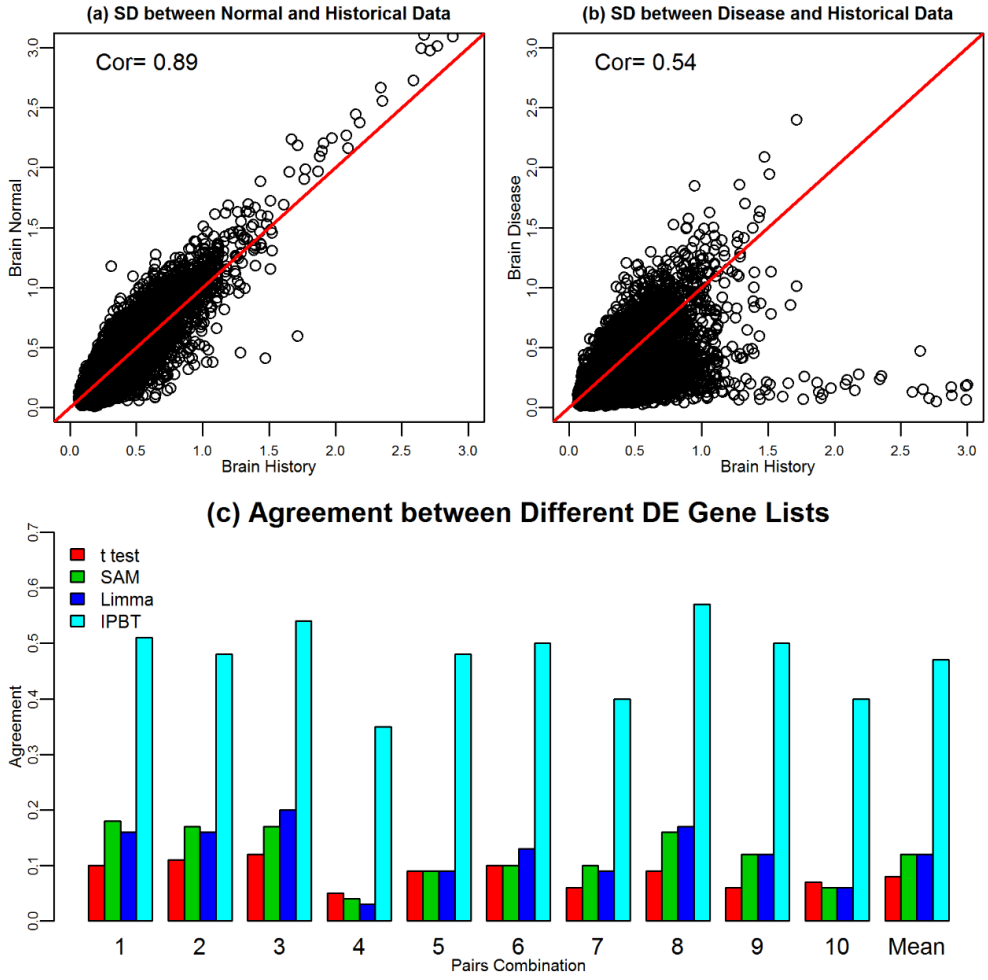

Figure 10 Real data analysis for brain data (a) Comparison of standard deviations (SD) obtained from the five brain normal samples and that obtained from the brain historical data. (b) Comparison of standard deviations obtained from the five brain disease samples and that obtained from the brain historical data. (c) Agreements between all pair combinations of top 1000 genes from all 5 DE gene lists.

We also conduct similar analyses on the brain samples. We randomly choose five normal brain samples (out of 39) as controls and five disease brain samples (out of 31) and use them as the current data. Data from the 34 remaining normal brain samples are used as

historical data. Corresponding results which display similar pattern are shown in Figure 10 (a) and (b).

DE gene detection

For real data analysis, since it is extremely difficult to know what the real DE genes are, we use agreement as the measurement of performance. This strategy has been commonly used in microarray data analysis studies (Lim, Li, Choi, & Wong, 2015; Lim & Wong, 2014). In this study, again using the global gene map data, we randomly select two normal heart samples and two disease heart samples. Data from the remaining 34 normal heart samples are used as historical data. We then apply IPBT and competing methods on these data to obtain a list of top 1,000 DE genes for each method. We repeat the above sampling and testing procedure five times. Then for each method, we calculate the agreement between every pair of the 1,000 DE gene lists. Figure 9(c) summarizes the results, which shows significant higher agreement for our IPBT method compared to others. We also compared the DE gene calling consistency as we did in the simulation study, and the results are summarized in Table 9. Again, IPBT performs the best among all methods tested. The procedure is repeated for brain data, comparing two normal brain samples and two disease brain samples. The results are shown in Figure 10(c) and Table 10. IPBT again achieves the best agreement and consistency.

Table 9 Consistency for Detecting DE Genes (Heart)

Overlap times	3	4	5	High (4+5)	Moderate (3+4+5)
student's t-test	108	15	1	16	124
SAM	219	68	48	116	335
Limma	203	69	45	114	317
IPBT	**291**	**189**	**164**	**353**	**644**

* The best results are in **bold**.

Table 10 Consistency for Detecting DE Genes (Brain)

Overlap times	3	4	5	High	Moderate

				(4+5)	(3+4+5)
student's t-test	72	7	3	10	82
SAM	128	27	6	33	161
Limma	118	36	4	40	158
IPBT	**275**	**215**	**213**	**428**	**703**

* The best results (after excluding Z test) are in **bold**.

To get a comprehensive picture of performance, we also conduct performance comparison on each of the five testing sets individually. We notice that different methods perform extremely similarly when sample sizes are large. Even the *t*-test could have an AUC more than 0.95 with 5 samples in the simulation study. Since we have more than 30 samples for heart data, we use the Student's *t*-test to compare the whole set of control with the whole set of treatment samples to define a gold standard DE gene list. In Figure 9, we

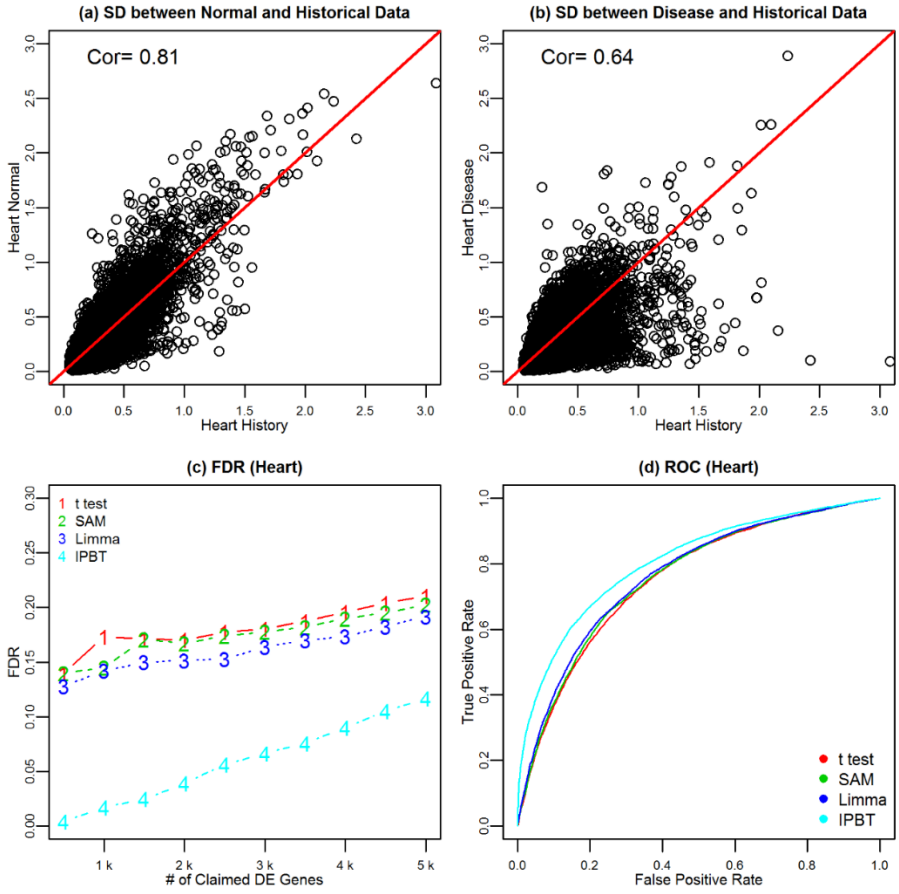

Figure 11 Real data analysis for heart dataset 1.

show the agreement of five independent datasets. Here, using our pre-defined gold standard, we show the performance for different methods on each individual dataset in Figure 11 to

15. We apply student's *t*-test, SAM, Limma and IPBT on each independent dataset. Figure 11-15 (c) and (d) shows the performances of different methods by their FDRs and ROC curves. IPBT achieves the lowest FDR and highest AUC for ROC curve. In particular, the top ranked genes in our DE gene list have a fairly low FDR.

Figure 11-15 share the same legend: (a) Comparison of standard deviations (SD) obtained from the five normal heart samples and that obtained from the heart historical data. (b) Comparison of standard deviations obtained from the five heart disease samples and that obtained from the heart historical data. (c) the FDR for detecting DE genes in the top ranked genes obtained using four different methods in the heart study. (d) ROC curves comparing four different methods for detecting DE genes in the heart study.

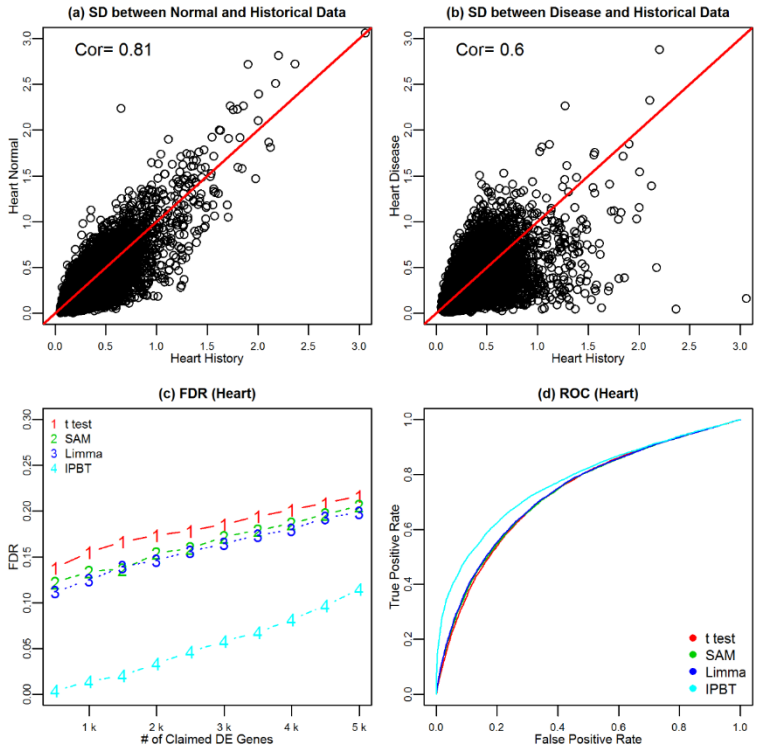

Figure 13 Real data analysis for heart dataset 2.

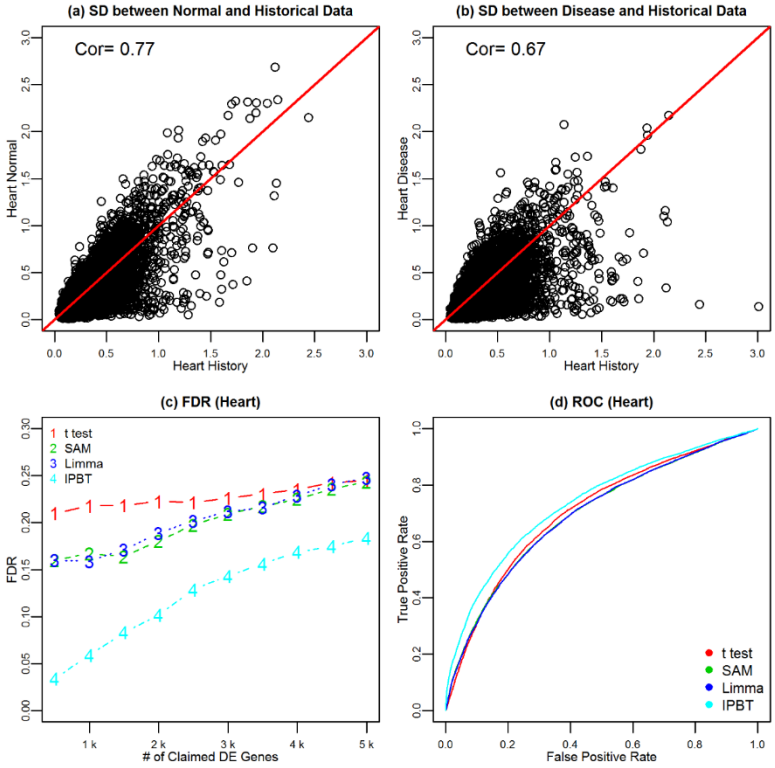

Figure 12 Real data analysis for heart dataset 3.

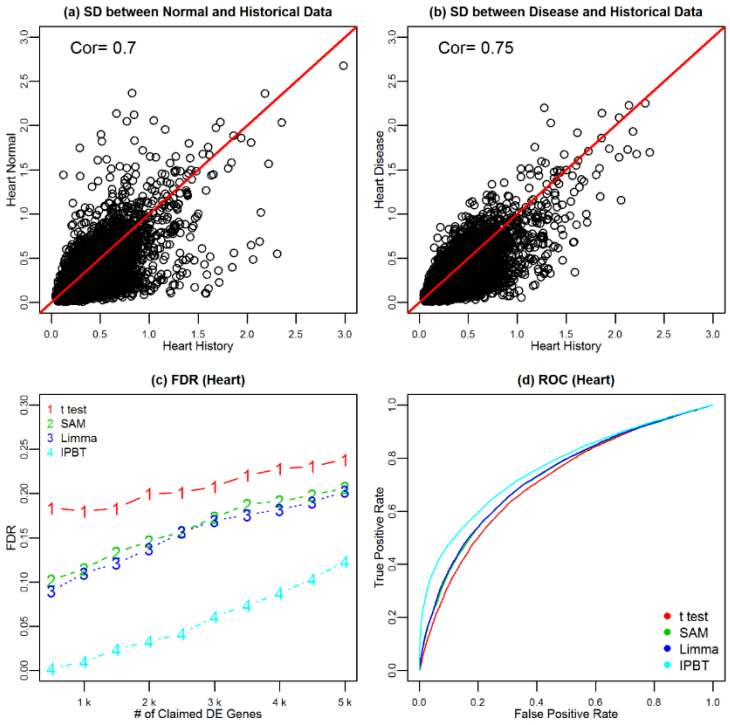

Figure 15 Real data analysis for heart dataset 4.

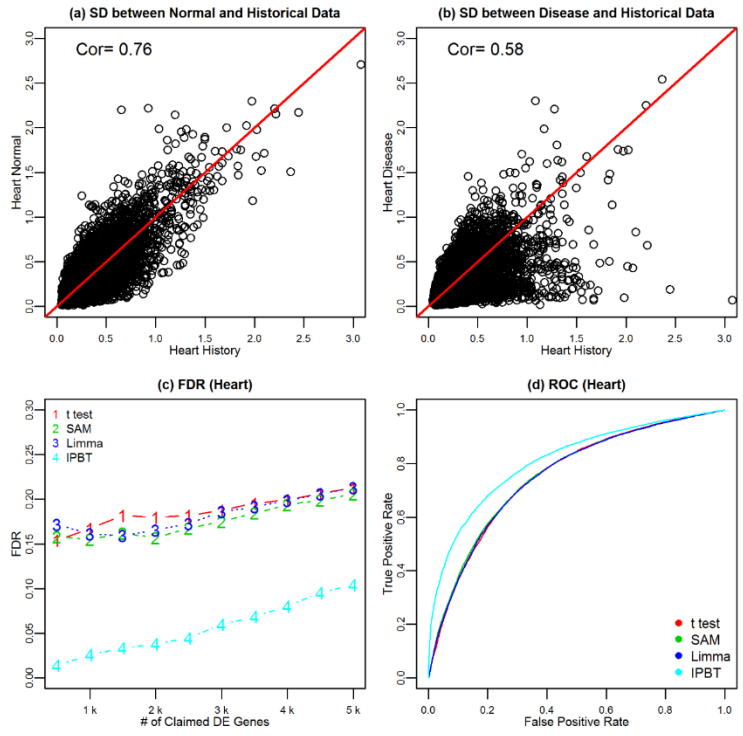

Figure 14 Real data analysis for heart dataset 5.

2.3.2 Real Data Study II: Latin Square Hgu133a Spike-in Experiment Data

This data set consists of three replicates of 14 separate hybridizations of 42 spiked transcripts in a complex human background (HeLa cells) at concentrations ranging from 0.125pM to 512pM (Affymetrix, Santa Clara, CA). Since the spike-in genes are known, this dataset has been widely used in evaluating the performance of Microarray preprocessing algorithms (McCall, Bolstad, & Irizarry, 2010; Z. Wu & Irizarry, 2004) and DE gene analysis methods (Lo & Gottardo, 2007). In our study, each time we select two out of the 14 separate hybridizations as the control and treatment groups (each group has three replicates) respectively. All 91 pairs are tested for DE gene detection. After excluding the probes that do not exist in Affymetrix GeneChip U133A, 34 probes are *bona fide* differentially expressed each time. We use 42 datasets from HeLa cells (cervical adenocarcinoma cell line) from the global gene expression map as the historical data.

In this study, each method generates a DE probe list (ranked by the test statistics) in every pair of the control and treatment groups and we obtain the proportion of correct DE calls (match the 34 *bona fide* DE probes). Table 11 summarizes the average number of correctly identified DE probes among the top k ($k = 5, 10, \ldots, 40$) probes across all 91 control and treatment combinations. Figure 16 shows the box plots of FDRs for the top k probes called significant. IPBT consistently detects more *bona fide* DE probes hence has a lower FDR in terms of the median across all the experiments. In addition, IPBT is more robust since it consistently shows the smallest interquartile ranges in the boxplot. All these results show that IPBT performs better than other methods.

Table 11 Average number of correctly identified DE probes across all 91 group pairs on Spike-in Experiments data among the top k probes.

Top *k*	5	10	15	20	25	30	35	40
student's *t*-test	3.1	5.9	8.6	11.2	13.8	16.3	18.7	21.0
SAM	3.3	6.2	8.9	11.6	14.3	16.7	19.5	22.3
Limma	3.3	6.2	8.8	11.6	14.1	16.8	19.5	22.3
IPBT	**3.9**	**7.4**	**10.4**	**13.3**	**16.4**	**19.1**	**22.0**	**25.0**

* The best results are in **bold**.

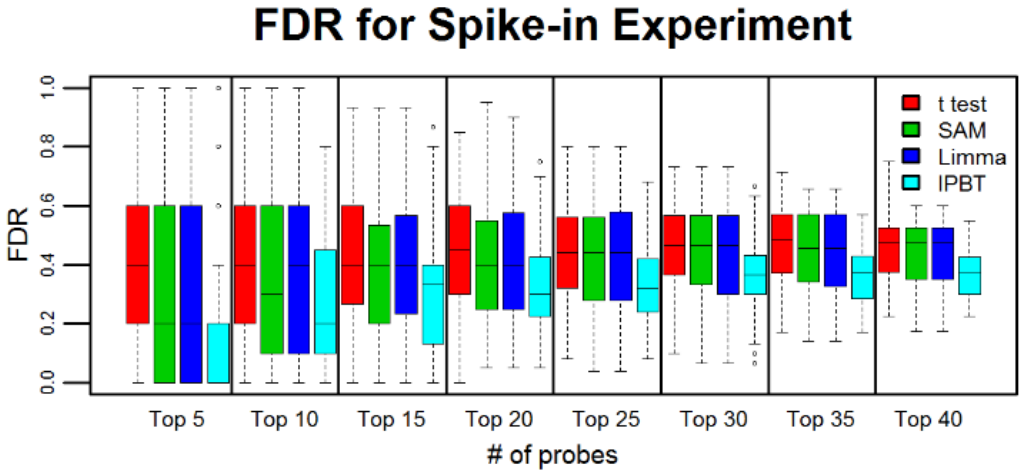

Figure 16 All the detection methods are applied to all 91 pairs of hybridizations. Box plots of FDRs are shown for all 91 group pairs when calling top k probes significant.

2.4 Discussion and Conclusion

Discussion

In this topic, I present a novel strategy of reutilizing relevant information contained in historical data to improve DE gene detection. Simulation studies and real data applications show that our method IPBT significantly outperforms other existing methods in terms of both accuracy and consistency in detecting DE genes. In particular, when the DE genes have relatively low intrinsic variances, methods based on the standard Bayesian

hierarchical models, which borrow information across genes in the same experiment, perform poorly whereas IPBT maintains its superior performance by borrowing information across experiments on the same gene.

In general, Bayesian hierarchical models provide an attractive statistical framework for handling 'large *p*, small *n*' inference problems. Because they can "borrow" information from all genes in the genome to aid the inference on a single gene so that the poor performance due to limited sample size can be improved. However, as we showed in this study, the traditional Bayesian hierarchical model approach can suffer from an "over-correction" problem and produce false negatives. In addition, the empirical Bayesian approach assumes a common prior for every gene, which will limit the effectiveness of the approach for genes with dramatically different behaviors. In contrast, IPBT assumes gene-specific, informative priors. With the rapid proliferation of high-throughput genomics big data, deriving these informative priors is no longer an issue.

Meta-analysis is a powerful tool for combining multiple studies of a related hypothesis and has been applied to microarray data (Conlon, Song, & Liu, 2007; Tseng, Ghosh, & Feingold, 2012). Our approach is different from meta-analysis because historical data used in IPBT may come from experiments with a different hypothesis, and the historical data are used indirectly in the form of informative priors in Bayesian inference.

There is much room for improvement in IPBT. First, the informative prior used in IPBT is gene-specific so DE gene analysis is done gene-by-gene. In reality we know some genes are correlated with each other such as genes located in the same pathway or sharing similar biological functions. A potential extension of IPBT is to introduce correlation among genes. Correlation information can be derived from biological knowledge or

historical data. Recent studies have demonstrated the benefit of incorporating correlation information in the inference of DE genes (Lim & Wong, 2014; Soh, Dong, Guo, & Wong, 2011).

Second, the current IPBT method uses normal distributions to model log transformed expression measures. The distribution choice is made mainly for mathematical convenience. One can replace normal distribution with other non-normal ones to achieve robustness in inference in the same way as Ganjali, Baghfalaki, and Berridge (2015) have done in their study of DE gene detection.

Third, we assume the expression values used by IPBT have already been background-corrected and normalized. This is possible with the powerful normalization techniques such as RMA. It is however, desirable if additional consideration is factored in the model to account for subtle experiment-to-experiment biases in the data as shown in studies such as Arima, Liseo, Mariani, and Tardella (2011) and Lewin, Richardson, Marshall, Glazier, and Aitman (2006). This will potentially make IPBT more flexible and further improve its performance.

Conclusion

In conclusion, we investigate the feasibility and effectiveness of deriving informative priors from historical microarray data and using them to help detect DE genes in studies with limited sample size. Through simulation and real data analysis, we show that our method significantly outperforms competing methods including the popular and state-of-the-art standard Bayesian hierarchical model-based approaches. The study has been published in Bioinformatics. (B. Li, Sun, He, Zhu, & Qin, 2015)

Taking advantage of the resource of global gene expression map developed by Lukk et al. (2010), we have calculated informative priors for 96 different groups of cell types using the Affymetrix U133A GeneChip as a community resource for DE gene study (all groups in the global gene expression map with at least 10 samples). We made the calculated informative priors freely available for the research community, which can be downloaded from https://github.com/benliemory/IPBT.

The strategy we propose in this paper is not limited to the microarray platform. RNA-Seq (Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008) is considered a better alternative for measuring gene expression because it can provide more information about the transcriptome (alternative splicing, gene fusion, etc.). We did not use RNA-Seq data since currently much less "historical" data is available compared to microarray, due to the comparatively higher cost and shorter time of the adoption of RNA-Seq. As the total volume of RNA-Seq data increases, the IPBT framework can be applied to RNA-Seq as well. Cross-platform models may also be considered.

Our work illustrates the feasibility and benefits of exploiting the increasingly available genomics big data in statistical inference and presents a promising strategy for dealing with the ‘large p, small n’ problem.

Chapter 3

Improving hierarchical models using rank information from historical data with applications in high throughput genomics data analysis

3.1 Methods

This section introduces model building, group choosing, inferences, and tests for two new improved hierarchical models using rank information from historical data-- a stratified hierarchical model (stHM) and a sliding window hierarchical model (swHM).

3.1.1 Motivation

A crucial assumption made by hierarchical models is that some or all features are considered exchangeable. That is to say, one is unable to distinguish any given feature from the others given the data observed since these features are regarded homogeneous. We believe this can be a rather strong assumption and it is often violated. Genes in the genome are designed to carry out different tasks. For example, the diverse biological features of developmental genes, housekeeping genes, response genes are reflected in their expression profiles measured under many different conditions. As shown in Figure 2, there are substantial differences in terms of the mean and variances from gene to gene. Given the

heterogeneity in high-throughput genomics data, it is counter-productive for a highly and stably expressed housekeeping gene to borrow information from developmental genes with a bimodal expression pattern across experiments.

Just like many other fields, the total amount of available genomic data is enormous and is still growing rapidly in the era of Big Data (Fan, Han, & Liu, 2014). As shown in chapter 2, there is a massive collection of publicly-available datasets (historical data) produced by gene expression microarray technology. The collection of historical microarray data is so rich that for a given new experiment, oftentimes one is able to find datasets under similar conditions or of the same cell/tissue type from the collection. Therefore, it is highly desirable if we can improve statistical inference of new experimental data by utilizing these historical data.

Over the years, numerous strategies have been proposed to leverage historical data to help analysis of new experimental data under various scenarios and contexts. As early as 2004, Kim and Park proposed to utilize historical data to obtain an improved estimate of the sample variance under a Student *t*-test framework for detecting differentially expressed (DE) genes (Kim & Park, 2004). However, they use historical variance directly in an adjusted *t*-test without incorporating current information. Altman and colleagues presented the singular value decomposition (SVD) Augmented Gene expression Analysis Tool (SAGAT) to increase the discovery power of microarray experiments by using publicly available microarray datasets (Daigle et al., 2010). They only use co-expression information from historical data and do not utilize other historical information. Wu and colleagues utilize a database of historical experiments to adjust background for the DNA microarray (Sui et al., 2009) but the historical data are not explored in a DE gene detection

setting. Therefore, with further accumulation of historical data, a model-based method better incorporating both historical and current data for DE gene detection may offer advantages.

On the other hand, statisticians increasingly recognize the importance of incorporating historical data into inference procedures. In particular, the Bayesian framework has been identified as the ideal vehicle that can be utilized to achieve this goal. Among them, the power prior (Chen & Ibrahim, 2006; Ibrahim, Chen, Gwon, & Chen, 2015), which has been used in various fields including clinical trials, health care, etc., has been proposed to construct informative priors from historical data to improve the inference for current data (Duan, Ye, & Smith, 2006; Hobbs, Carlin, Mandrekar, & Sargent, 2011).

In chapter 2, I proposed a Bayesian strategy to incorporate historical data to help detect DE genes in microarray experiments. The main idea behind the proposed method, named the informative prior Bayesian test (IPBT), is the construction of gene-specific, informative priors for the variance of each gene using historical data. IPBT is perhaps the first method that incorporates historical data in a formal Bayesian framework to detect DE genes. Despite its significant improvement over standard hierarchical model-based methods demonstrated using both simulated as well as real benchmark datasets, the success of IPBT hinges on the availability of large quantity of high quality historical data produced from the same platform, which limits the applicability of IPBT.

It is highly desirable to utilize historical data generated from different platforms. This is not possible using IPBT because this method relies on an inexplicit assumption that the current data and historical data (for each gene's expression measure) share similar a distribution. To overcome this limitation, in this topic, I propose a novel strategy to

incorporate historical data into the hierarchical model framework. The central idea is to "partially" utilize historical data: instead of numerical values, I only retain the order of the genes in the genome ranked by their variances estimated from historical data. Thus under the hierarchical model framework, when borrowing strength from other genes, instead of all genes in the genome, our approach select a subset of genes that have the closest variances according to historical data. To be specific, we proposed two different approaches, a stratified hierarchical model and a sliding window hierarchical model. In the first approach, we decompose all genes into disjoint groups such that borrowing strength only occurs among genes in the same group. The gene groups are determined by historical data such that the expressions of genes within a group are exchangeable. In the second approach, instead of fixed windows, we use a sliding window approach to group neighboring genes.

3.1.2 stHM and swHM

Here I start with basic notation and assumptions in equation (1). stHM and swHM are modified from the classical hierarchical model (HM) described as in equation (2)-(4).

stHM

Let $X_{i(g)jk}$ denotes *k*th replicate of log-transformed expression for *i*th gene in group g (g = 1, 2,… G) under condition *j*. We have:

$$\mathrm{X}_{i(g)jk} \mid \mu_{i(g),j}, \sigma_i^2 \sim \mathrm{N}(\mu_{i(g),j}, \sigma_{i(g)}^2) \tag{31}$$

$$\mu_{i(g)j} \mid \mu_g, \tau_g^2 \propto 1 \tag{32}$$

$$\sigma_{i(g)}^2 \mid \nu_g, \omega_g^2 \sim \mathrm{Inv} - \chi^2(\nu_g, \omega_g^2) \tag{33}$$

where mean parameter $\mu_{i(g),j}$ and variance parameter $\sigma^2_{i(g)}$ for genes in the same group are assumed to follow the same distribution with hyper-parameters ν_g and ω^2_g. Similarly, an empirical Bayes estimator $\widehat{\sigma^2_{i(g),B}}$ for $\sigma^2_{i(g)}$ is used for the subsequent adjusted *t*-test. The main difference between stHM and HM is that stHM "borrows" information only from genes in the same group instead of all genes in the same experiment. With appropriately identified groups, stHM "borrows" information from more similar genes and could alleviate the over-shrinkage suffered by HM. All genes in an experiment are divided into G disjoint subsets based on the order of their standard deviations estimated from the collection of historical data. More details about determining the number of groups (G) will be discussed separately.

swHM

In this approach, borrowing strength for each particular gene under the hierarchical framework is restricted to its "neighboring" genes, again determined by the standard deviations estimated from the historical data. Following the notations in (31)--(33), swHM can be described as:

$$\mathrm{X}_{i(g_i)jk} \mid \mu_{i(g_i),j}, \sigma_i^2 \sim \mathrm{N}(\mu_{i(g_i),j}, \sigma^2_{i(g_i)}) \tag{34}$$

$$\mu_{i(g_i)j} \mid \mu_{g_i}, \tau^2_{g_i} \propto 1 \tag{35}$$

$$\sigma^2_{i(g_i)} \mid \nu_{g_i}, \omega^2_{g_i} \sim \mathrm{Inv}-\chi^2(\nu_{g_i}, \omega^2_{g_i}) \tag{36}$$

Where g_i indicate *i*th gene's sliding window (its own group). The swHM strategy enables the identification of a group of more homogeneous genes to estimate the gene's adjusted standard deviation at the cost of more computation burden. More details about determining the size of each group will be discussed separately.

Group dividing

Our main purpose is to divide genes into subsets in which genes are consider homogeneous. A straightforward strategy is to use each gene's mean expression level estimated from the current data to select subsets. This strategy has been used in methods developed to detect DE genes from RNA-Seq data (Robinson & Smyth, 2007; H. Wu et al., 2013). In our stHM and swHM, we use the order of standard deviation estimated from historical data to determine subsets. We define the "Group Dividing Metric" (GDM) to indicate whether the number of groups is optimal:

$$\text{GDM(stHM)} = [\sum_g \frac{\sum_{i(g)}\left(S_{i(g)} - \overline{S_g}\right)^2}{I(g)}]/\text{G} \tag{37}$$

$$\text{GDM(swHM)} = [\sum_i \frac{\sum_{i(g_i)}\left(S_{i(g_i)} - \overline{S_{g_i}}\right)^2}{I(g_i)}]/\text{I} \tag{38}$$

where $S_{i(g)}$ is adjusted SD estimate from stHM or swHM for *i*th gene in group *g*, $\overline{S_g}$ is the mean of SD estimates in group *g*, *I(g)* is the total number of genes in group *g*, G is current number of groups and I is total number of genes. $S_{i(g)}$ is obtained by applying a classical hierarchical model within each group. For completeness, we list the empirical bayes estimator for SD below (Ji & Wong, 2005):

$$S_{i(g)} = \sqrt{\left(1 - \widehat{B_g}\right)sd_{i(g)}^2 + \widehat{B_g}\overline{sd_g^2}} \tag{39}$$

$$\widehat{B_g} = \frac{2/v}{1 + 2/v}\frac{I(g) - 1}{I(g)} + \frac{1}{1 + 2/v}\left(\frac{2}{v}\right)\left(\overline{sd_g^2}\right)^2\frac{I(g) - 1}{S_g} \tag{40}$$

where $v = 2(K - 1)$ and $S_g = \sum_{i(g)}\left(sd_{i(g)}^2 - \overline{sd_g^2}\right)^2$

One issue worth noting is that with increased group number (fewer genes in a group), the empirical Bayesian estimate might be inappropriate (yielding negative values

for SD) when the expression values for all the genes in a group are too close. We avoid such inappropriate scenarios by using the group mean SD as the estimate for all the genes in the group.

Models for DNA methylation data

Log ratios of methylated to unmethylated intensities (M value) are more widely used than the ratio of the methylated to the total of methylated and unmethylated intensities (beta value) for 450K methylation arrays because the M value performs similarly to gene expression data measured by microarray and all the methods on gene expression microarray can be applied almost identically to M values estimated from 450K array (Aryee et al., 2014; Robinson et al., 2014). We also apply our new approaches on M value directly (formulas (31-33) for stHM and (34-36) for swHM) and do not explicitly write out the models again. However, for sequencing based DNA methylation profiling approach (BS-Seq), the basic model assumption is completely different. This paper mainly discusses the improvement of HM in array data, thus we only show one state-of-the-art beta-binomial Bayesian hierarchical model (DSS) for differential methylated locus (DML) calling of BS-Seq data (Feng et al., 2014) and our stratified strategy (stDSS) to explore the possibility to borrow information across platforms. For the sake of completeness, we here rewrite the distribution assumptions made in DSS as follows:

$$\mathrm{X}_{ijk} \mid p_{ijk}, N_{ijk} \sim \mathrm{Binomial}(N_{ijk}, p_{ijk}) \tag{41}$$

$$p_{ijk} \sim \mathrm{Beta}(\mu_{ij}, \Phi_{ij}) \tag{42}$$

$$\Phi_{ij} \sim \log - \mathrm{normal}(m_{0j}, r_{0j}^2) \tag{43}$$

where X_{ijk}, N_{ijk} denote methylation reads and total reads for *i*th CpG site, *j*th group and kth replicate, respectively. p_{ijk} is the underlying true methylation proportion. μ_{ij}, Φ_{ij} are the mean and dispersion parameter for beta distribution, respectively. And m_{0j}, r_{0j}^2 are mean and variance parameter for the log-normal distribution.

Similarly, we modified formulas (41-43) into (44-46) for stDSS:

$$X_{i(g)jk} \mid p_{i(g)jk}, N_{i(g)jk} \sim \text{Binomial}(N_{i(g)jk}, p_{i(g)jk}) \quad (44)$$

$$p_{i(g)jk} \sim \text{Beta}(\mu_{i(g)j}, \Phi_{i(g)j}) \quad (45)$$

$$\Phi_{i(g)j} \sim \log - \text{normal}(m_{gj}, r_{gj}^2) \quad (46)$$

3.2 Simulation Study

I conducted two sets of real-data based simulation studies to demonstrate the advantages of our new approaches. 1) In the first set of simulations, I use 566 normal solid tissue microarray datasets obtained by Affymetrix GeneChip U133A from the global gene expression map to show the correlation between SD estimates and the true SDs. All the following simulation are generated with the parameters obtained from these 566 normal samples. We also show that GDM is a good indicator for group dividing by simulation. 2) In the second set of simulation, I show the false discovery rates (FDR) and Receiver operating characteristic (ROC) curves for our new approaches are almost as good as IPBT and outperform all other competing methods. I also show that our new approaches could be more robust than IPBT when historical data does not have high quality.

3.2.1 Simulation Study I: SD Estimate and Group Dividing

We conduct a simulation study to demonstrate the accuracy of standard deviations (SD) estimated by different strategies. Figure 17 shows that HM merely shrinks SD estimate without changing the order of them and the straightforward strategy of using current

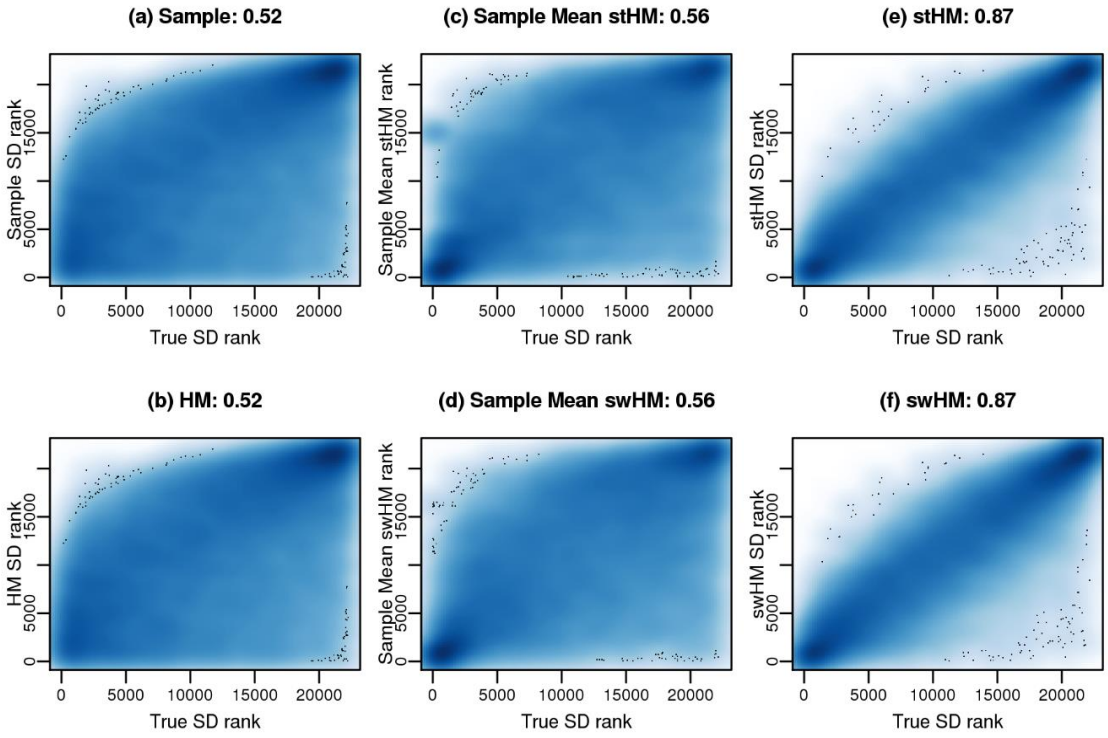

Figure 17 Standard deviation (SD) ranks between different strategies. True SD ranks V.S. (a) Sample SD rank (b) Standard HM SD rank (c) Sample mean stHM rank (d) Sample mean swHM rank (e) stHM SD rank (f) swHM

expression mean to choose subset genes improves SD estimates. The shrinkage of SD may change the order of SD in that scenarios, but here the orders remain since all genes have the same sample size. Using historical data's SD rank could significantly improve SD estimates. Our two approaches (stHM and swHM) have similar performance with the help of historical data.

We also tested how changing the number of groups could affect the estimation of standard deviations with six different settings (Figure 18). The figure shows that for low numbers of groups, the correlation between SD estimate and true SD will dramatically

increase. But with increasing of numbers of groups, there are few additional gains in the correlation.

We also show the trend for GDM with different number of genes or different sample sizes in Figure 19. The blue line shows that GDM decreases with the increase of

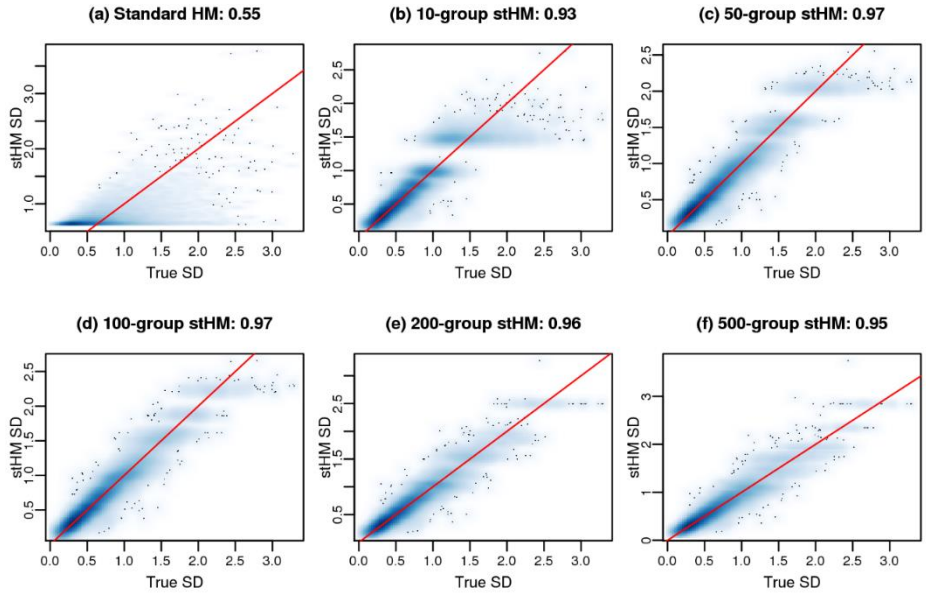

Figure 18 True SD V.S. stHM with different group numbers. (a) without grouping (standard HM) (b) 10 groups (c) 50 groups (d) 100 groups (e) 200 groups (f) 500 groups

number of groups. The decrease is very fast at the beginning but becomes stable soon. The red line is the correlation between estimated SD and true SD in the corresponding number of group. We can see that the correlation increases at the beginning but also becomes steady soon. Actually, GDM and the correlation between estimated SD and true SD have strong negative correlation. Table 12 lists corresponding correlations for Figure 19.

Table 12 Correlation between GDM and correlation between SD estimate and true SD

Correlation	2 Samples	5 Samples	10 Samples
1000 Genes	-0.863	-0.813	-0.896

5000 Genes	-0.985	-0.974	-0.989
10000 Genes	-0.982	-0.993	-0.975

Hence, we find GDM is a good indicator for deciding optimal group numbers. Actually, when GDM is stable, different number of groups will lead to similar results. Therefore, our

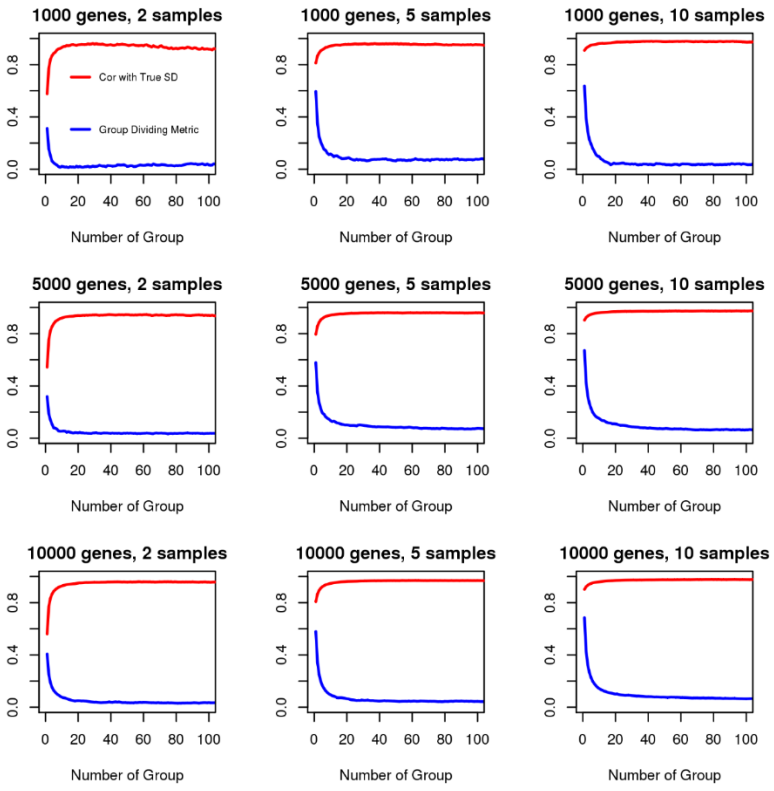

***Figure 19** GDM in different scenarios*

rule of thumb for choosing number of groups is to find the "turning point" in the curve for GDM. A conservative choice is to pick a number slightly larger than the "turning point". For example, we can choose 15 to 20 groups for 1000 genes, 2 samples.

3.2.2 Simulation Study II: DE Gene Detection Performances

Similar to section 2.2.2, I conduct a simulation study to compare stHM and swHM with IPBT and other well-established methods for detecting DE genes: (i) Student's *t*-test, (ii)

SAM (R package ‘siggenes’); (iii) Limma, (R package ‘Limma’); (iv) Z test using the true variance; and (v) IPBT (R package ‘IPBT’).

Expressions for 1000 genes in *k* (ranging from 2 to 5) samples are simulated for both the control and treatment groups. 10% of the 1000 genes (i.e. 100 genes) are randomly selected as designated DE genes. Gene expression values in both the control and treatment groups are assumed to follow normal distributions. We derive each gene’s sample mean and variance from the 566 normal samples in the collection. For the treatment group, the

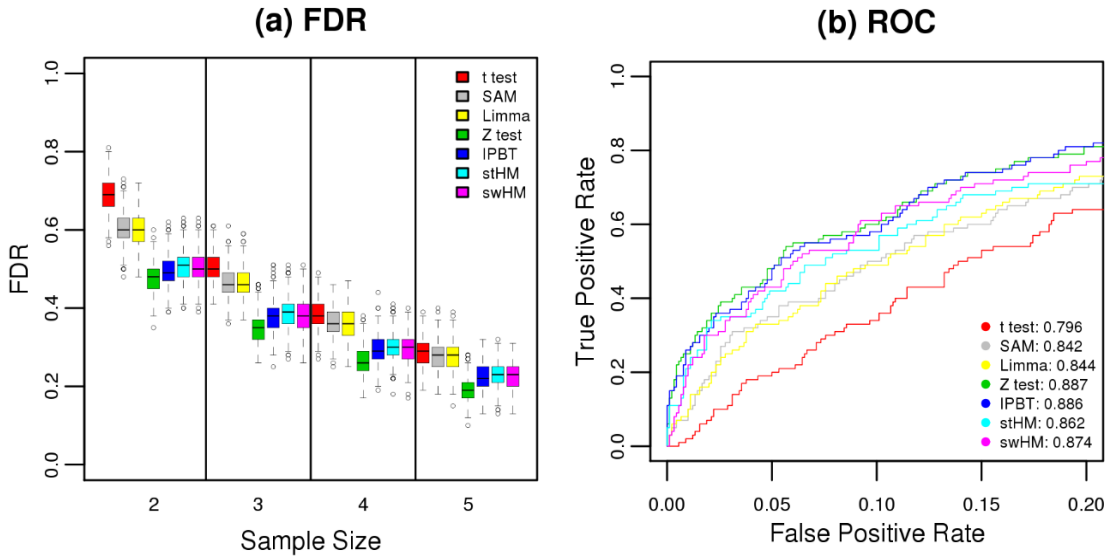

Figure 20 Simulation with accurate historical data (a) FDR (b) a typical ROC curve

mean and variance of a gene’s expression value are assumed to be the same as their counterparts in the control group except for the DE genes. The mean expression values for DE genes in the treatment group are two standard deviations higher. For historical data used by IPBT, stHM, swHM, 10 normal samples are randomly chosen out of 566 (without replacement) from the global gene expression map.

We use the empirical FDR (Benjamini & Hochberg, 1995; Tusher et al., 2001) to evaluate the performance for the top 100 genes ranked by the test statistics. The simulation

is repeated 500 times for each method. Figure 20(a) summarizes the distributions of the 500 FDRs for different methods by box plots. All methods using historical data clearly

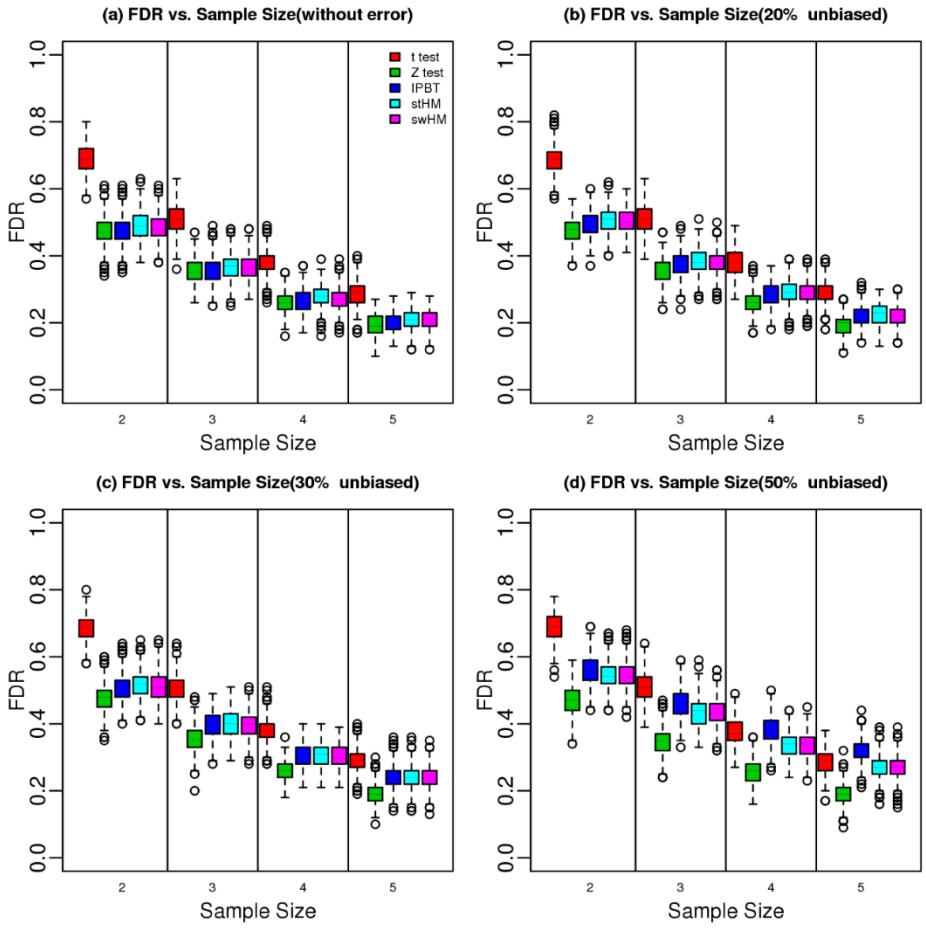

Figure 21 Simulation with inaccurate historical data (a) accurate historical data (b) historical data with 20% unbiased noise (c) historical data with 30% unbiased noise (d) historical data with 50% unbiased noise

performs better than methods without using historical data except for the Z test with true variances (considered the gold standard). The methods using historical data and Z test have fairly close performances. We also use Receiver Operator Characteristic (ROC) curves to compare different methods. Figure 20(b) shows a typical example of ROC curve for one single simulation with sample size k = 2. The ROC curves again show that methods with

historical data perform better than methods without historical data except for the Z test, and the performances of methods with historical data and Z test are similar.

We also repeat the simulation with a noisier historical data. Figure 21 shows that IPBT's performance started to deteriorate with noisier historical data while our new strategies maintain its performance advantage. Figure 20 and 21 together demonstrate that our new strategies could be almost as good as IPBT with accurate historical data and perform more robust than IPBT when the historical data becomes noisier.

3.3 Real Data Analysis

In this section, I will use stHM and swHM in real data analysis for DE gene detection. I will also apply stHM and swHM to 450K Methylation data to detect DML. I will further show an example of analyzing BS-Seq data using 450K array historical data, which reveals the possibility of borrowing information across different platforms.

3.3.1 Real Data Study I: Global Gene Expression Map

Similar to Section 2.3.1, we use the heart and brain datasets from the global gene expression map to compare the performance of different methods detecting DE genes. We apply all the methods except the Z test used in our simulation studies to real gene expression microarray data contained in the global map of gene expression.

Heart data

Since we do not know the true DE genes in the real data, we use agreement as the performance measure which is defined to be the proportion of overlap between two lists

(equal length) of genes. Two normal (out of 36) and two disease (out of 51) heart samples are randomly selected and we treat the remaining 34 normal heart samples as historical data. We then apply stHM, swHM and competing methods on these data to obtain a list of top 1000 DE genes for each method, respectively. We repeat the above sampling and

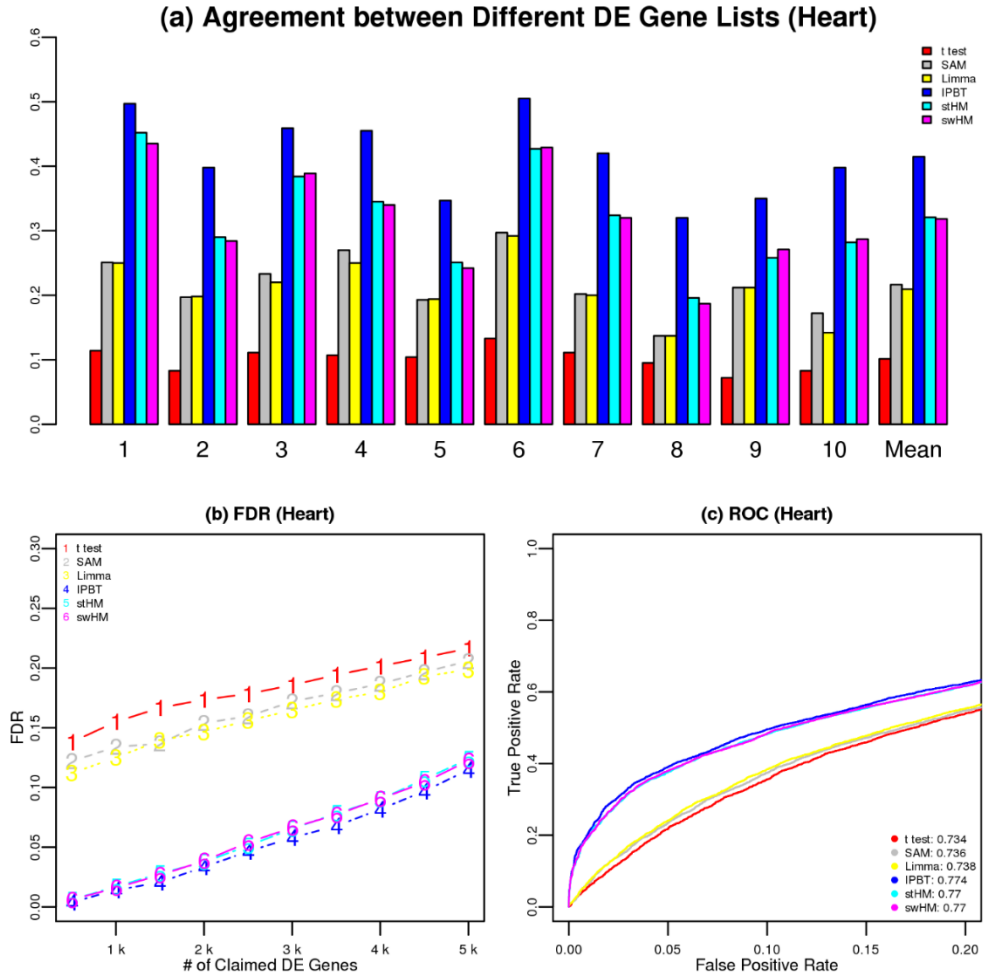

Figure 22 (a) Agreement for heart data (b) FDR for heart data (c) ROC curve for heart data

testing procedure five times. Then we calculate the agreement between every pair of the 1000 DE gene lists for each method. Figure 22(a) summarizes the agreement results, which shows that stHM and swHM have a higher agreement than methods that do not use historical data (*t*-test, SAM and Limma) but not as good as IPBT.

We also conduct performance comparison on each of the five testing sets individually. As different methods perform almost the same with sufficiently large sample size, we define the true DE genes by applying a *t*-test to all the available heart data. Figure 22 (b) and (c)

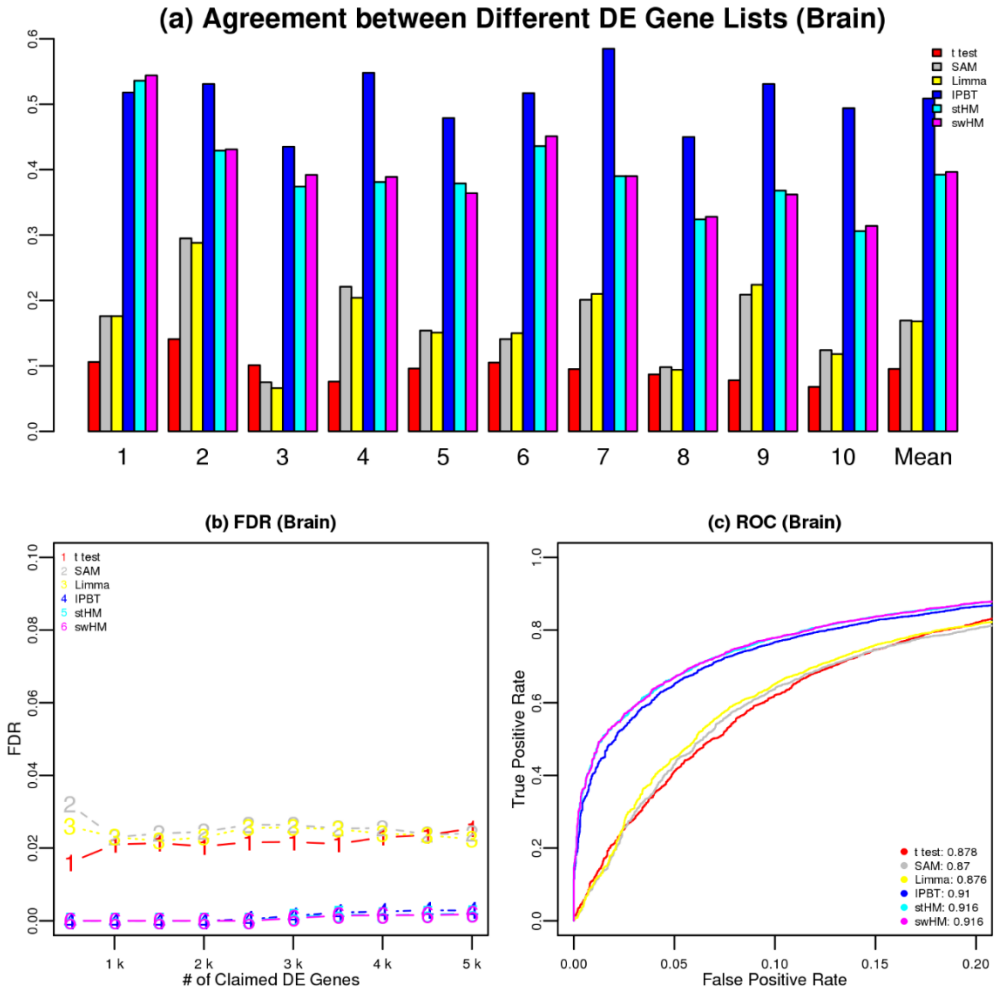

Figure 23 (a) Agreement for brain data (b) FDR for brain data (c) ROC curve for brain data

shows the performances of different methods by their FDRs and ROC curves. Again, methods with historical data perform similarly and much better than methods that do not use historical data.

Brain data

The analysis procedure for the heart data is also repeated, for the brain data, comparing two normal brain samples (out of 39) and two disease brain samples (out of 31). Figure 23 shows the corresponding results for brain data. Again, our new approaches perform similar to IPBT and much better than other methods.

3.3.2 Real Data Study II: DNA Methylation Data

Datasets

DNA methylation 450K array is an array-based technology measuring more than 485,000 CpG sites. On the other hand, BS-Seq, covering the whole genome (around 28 million CpG sites), is considered a better technology for measuring DNA methylation.

Here we use 50 liver cancer (LIHC) and matched normal control samples from The Cancer Genome Atlas (Cancer Genome Atlas 2012). Detailed barcodes of all these samples can be found in Section 3.5. For BS-Seq data, we use data from liver and hippocampus samples from the Roadmap Epigenomics project (Bernstein et al. 2010) (GEO accession number GSE64577).

Analyze 450K array data using 450K array historical data

Similar to Section 3.3.1 on gene expression microarray data, we take two normal and two cancer datasets and treat them as being collected from the current experiment. All other normal data are used as historical data. Figure 24 (a) shows that stHM and swHM achieve a better agreement than LIMMA and *t*-test. FDR and ROC curves in Figure 24 (b) and (c) again illustrate that methods with historical data could benefit from historical data.

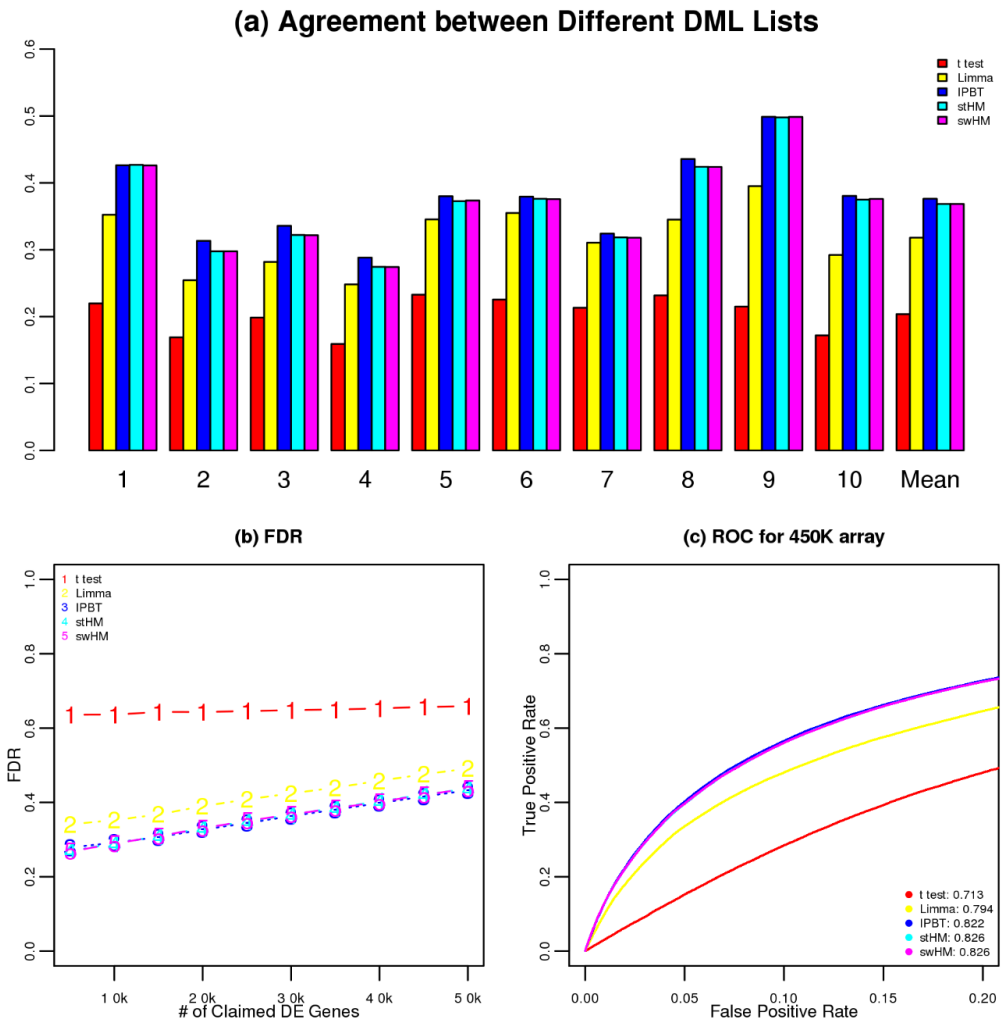

Figure 24 (a) Agreement for Methylation 450K array (b) FDR for Methylation 450K array (c) ROC curve for Methylation 450K array

Analyze BS-seq data using 450K array historical data

We next examine whether historical data generated from a different platform can be utilized effectively within our approach. Here we use 450K array data to group all the CpG sites. And then compare DSS and stDSS (We only include the CpG sites appeared in the 450K array). We adopted the same procedure as H. Wu et al. (2015) did to preprocess the BS-seq data. Since it is not possible to know which loci are bona fide DMLs, we use the FDR estimates from DSS to compare the number of DMLs identified after controlling

FDRs. Table 13 shows that how many DMLs are identified when controlling FDR at 0.01, 0.05 and 0.10, respectively. There are about 420,000 CpG sites involved in the analysis after quality control with Minfi excluding low quality CpG sites. For stDSS, we include two different group schemes (100 groups, each group has about 4,200 CpG sites and 4,500 groups, each group has fewer than 100 CpG sites). We can see that with the help of historical data, more DMLs can be identified while controlling FDR. With more groups, this advantage could be even more obvious.

Table 13 Number of DMLs identified when controlling FDR at 0.01, 0.05 and 0.10.

# of DML	FDR < 0.01	FDR < 0.05	FDR < 0.10
DSS	1,305	1,992	2,528
stDSS with 100 groups	1,312	2,032	2,567
stDSS with 4,500 groups	1,797	2,819	3,534

3.4 Discussion and Conclusion

In this topic, I introduce two new approaches (stHM, swHM) to detect DE genes or DMLs by improving the state-of-the-art hierarchical model with the aid of historical data. The simulation studies show that these two new approaches outperform the standard HM approach as expected and are more robust than IPBT, another method that utilizes historical data. Our data analyses demonstrate that our new approaches could be applied to a variety of datasets such as gene expression microarrays and methylation arrays. We further show that our new approaches make it possible to borrow data from different platforms. This feature could be extremely useful since more and more array and sequencing data

measuring similar underlying biological phenomena are accumulating but can hardly be effectively analyzed together. In summary, standard HM is the most efficient method when historical data are not available. However, IPBT could be a better alternative than HM with available high qualify historical data. When only historical data from other platforms are available or the historical data are noisy, we believe stHM and swHM are better choices and we highly recommend them. The study has been published in Statistics in Biosciences (B. Li, Li, & Qin, 2016).

Our main purpose in this paper is to introduce the framework of improved hierarchical models with targeted shrinkage and to illustrate that the framework can be generally applied to different types of data. However, we note that methylation 450K data and BS-Seq data have their own specific characteristics. In future work, we will explore further tailoring our framework to 450K methylation array and BS-Seq to obtain a better performance. In addition, our idea can also be extended to detect differential methylated regions (DMR) and to borrow information between gene expression microarray and RNA-Seq technology.

3.5 Appendices

Table 14 Barcodes for 450K array data used in real data analysis

Normal sample barcode	Cancer sample barcode
TCGA-BC-A10Q-11A-11D-A132-05	TCGA-BC-A10Q-01A-11D-A132-05
TCGA-BC-A10R-11A-11D-A132-05	TCGA-BC-A10R-01A-11D-A132-05
TCGA-BC-A10S-11A-11D-A132-05	TCGA-BC-A10S-01A-22D-A132-05
TCGA-BC-A10T-11A-11D-A132-05	TCGA-BC-A10T-01A-11D-A132-05
TCGA-BC-A10U-11A-11D-A132-05	TCGA-BC-A10U-01A-11D-A132-05
TCGA-BC-A10W-11A-11D-A132-05	TCGA-BC-A10W-01A-11D-A132-05
TCGA-BC-A10X-11A-11D-A132-05	TCGA-BC-A10X-01A-11D-A132-05
TCGA-BC-A10Y-11A-11D-A132-05	TCGA-BC-A10Y-01A-11D-A132-05
TCGA-BC-A10Z-11A-11D-A132-05	TCGA-BC-A10Z-01A-11D-A132-05

TCGA-BC-A110-11A-11D-A132-05	TCGA-BC-A110-01A-11D-A132-05
TCGA-BC-A112-11A-11D-A132-05	TCGA-BC-A112-01A-11D-A132-05
TCGA-BC-A216-11A-11D-A153-05	TCGA-BC-A216-01A-11D-A153-05
TCGA-BD-A2L6-11A-21D-A20Z-05	TCGA-BD-A2L6-01A-11D-A20Z-05
TCGA-BD-A3EP-11A-12D-A22H-05	TCGA-BD-A3EP-01A-11D-A22H-05
TCGA-DD-A113-11A-12D-A132-05	TCGA-DD-A113-01A-11D-A132-05
TCGA-DD-A114-11A-12D-A132-05	TCGA-DD-A114-01A-11D-A132-05
TCGA-DD-A115-11A-12D-A132-05	TCGA-DD-A115-01A-11D-A132-05
TCGA-DD-A116-11A-11D-A132-05	TCGA-DD-A116-01A-11D-A132-05
TCGA-DD-A118-11A-11D-A132-05	TCGA-DD-A118-01A-11D-A132-05
TCGA-DD-A119-11A-11D-A132-05	TCGA-DD-A119-01A-11D-A132-05
TCGA-DD-A11A-11A-11D-A132-05	TCGA-DD-A11A-01A-11D-A132-05
TCGA-DD-A11B-11A-11D-A132-05	TCGA-DD-A11B-01A-11D-A132-05
TCGA-DD-A11C-11A-11D-A132-05	TCGA-DD-A11C-01A-11D-A132-05
TCGA-DD-A11D-11A-12D-A132-05	TCGA-DD-A11D-01A-11D-A132-05
TCGA-DD-A1E9-11A-11D-A153-05	TCGA-DD-A1E9-01A-21D-A153-05
TCGA-DD-A1EB-11A-11D-A132-05	TCGA-DD-A1EB-01A-11D-A132-05
TCGA-DD-A1EC-11A-11D-A132-05	TCGA-DD-A1EC-01A-21D-A132-05
TCGA-DD-A1ED-11A-11D-A153-05	TCGA-DD-A1ED-01A-11D-A153-05
TCGA-DD-A1EE-11A-11D-A132-05	TCGA-DD-A1EE-01A-11D-A132-05
TCGA-DD-A1EF-11A-11D-A132-05	TCGA-DD-A1EF-01A-11D-A132-05
TCGA-DD-A1EG-11A-11D-A20Z-05	TCGA-DD-A1EG-01A-11D-A20Z-05
TCGA-DD-A1EH-11A-11D-A132-05	TCGA-DD-A1EH-01A-11D-A132-05
TCGA-DD-A1EI-11A-11D-A132-05	TCGA-DD-A1EI-01A-11D-A132-05
TCGA-DD-A1EJ-11A-11D-A153-05	TCGA-DD-A1EJ-01A-11D-A153-05
TCGA-DD-A1EL-11A-11D-A153-05	TCGA-DD-A1EL-01A-11D-A153-05
TCGA-DD-A39V-11A-11D-A20Z-05	TCGA-DD-A39V-01A-11D-A20Z-05
TCGA-DD-A39W-11A-11D-A20Z-05	TCGA-DD-A39W-01A-11D-A20Z-05
TCGA-DD-A39X-11A-11D-A20Z-05	TCGA-DD-A39X-01A-11D-A20Z-05
TCGA-DD-A39Z-11A-21D-A20Z-05	TCGA-DD-A39Z-01A-11D-A20Z-05
TCGA-DD-A3A1-11A-11D-A20Z-05	TCGA-DD-A3A1-01A-11D-A20Z-05
TCGA-DD-A3A2-11A-11D-A20Z-05	TCGA-DD-A3A2-01A-11D-A20Z-05
TCGA-DD-A3A3-11A-11D-A22H-05	TCGA-DD-A3A3-01A-11D-A22H-05
TCGA-EP-A12J-11A-11D-A132-05	TCGA-EP-A12J-01A-11D-A132-05
TCGA-EP-A26S-11A-12D-A16X-05	TCGA-EP-A26S-01A-11D-A16X-05
TCGA-ES-A2HS-11A-11D-A17Z-05	TCGA-ES-A2HS-01A-11D-A17Z-05
TCGA-ES-A2HT-11A-11D-A17Z-05	TCGA-ES-A2HT-01A-12D-A17Z-05
TCGA-FV-A23B-11A-11D-A16X-05	TCGA-FV-A23B-01A-11D-A16X-05
TCGA-FV-A2QR-11A-11D-A20Z-05	TCGA-FV-A2QR-01A-11D-A20Z-05
TCGA-G3-A25W-11A-12D-A16X-05	TCGA-G3-A25W-01A-11D-A16X-05
TCGA-G3-A25X-11A-11D-A16X-05	TCGA-G3-A25X-01A-11D-A16X-05

Chapter 4

Using historical data inferred gene panels to improve statistical inference on high throughput genomics data

4.1 Methods

This section introduces the general workflow on how to define gene panels from historical data and utilize them in a variety of high throughput genomics data analysis tasks.

4.1.1 Motivation

In our first topic, we developed IPBT, which utilize a Bayesian framework to detect DE genes from gene expression microarray data by borrowing information from the same gene across historical datasets. IPBT has demonstrated excellent performance for Affymetrix GeneChip microarray data as large amount of data generated using this platform are publicly available. A natural extension is to apply a similar framework to RNA-Seq data. One major barrier is that different published RNA-seq data were processed differently, not to mention how to normalize different types of processed data. Thanks to recount2 (Collado-Torres et al., 2017), it is plausible to explore strategies of borrowing information for RNA-Seq from similar samples. Recount2 processed and normalized all available RNA-seq samples consistently, which makes samples collected from different experiments comparable.

In my second topic, I devised an adaptive hierarchical model to borrow gene-specific external information as ranks instead of values to stabilize variance estimate when the external data do not have the highest quality or the historical data are produced using different platform. The core idea is to identify “similar” genes in the genome using historical data. For detecting DE genes, we use variances of genes to measure similarity. When handling RNA-seq samples, we need to consider both means and variances to group genes since there are a significant mean and dispersion trend for RNA-seq samples. It is even more attractive if we can pre-define consistent gene panels using historical data for different genomics data analysis tasks and detecting DE genes is only one application for this general gene panel framework.

4.1.2 Overview of IPBTSeq

In this section, I propose a more general framework than the adaptive HM framework I proposed earlier. It can be applied to multiple genomics data analysis tasks like detecting DE genes, quality control, etc. It could even be applied to a broader range of scenarios including the special case where there are no replicates. The basic idea is for any gene in the genome (target gene), we are going to use historical data to define a gene panel capturing certain underlying characteristics for the target gene. A gene panel serves as a secondary or supplementary feature for its target gene in terms of similarity. The genes in any given gene panel are similar enough to their target gene such that they can be treated as “pseudo replicates” for the purpose of inferring the target gene’s summary statistics such as mean and variance. One major key feature for gene panels is that they could remain stable for analogous samples. For example, the gene panels generated from kidney samples of GTEx can be used for kidney samples in TCGA. Therefore, it is

feasible to pre-define a collection of gene panels from historical data and use them directly in data analysis of new experiments even if historical data and new experimental data are generated from different labs, types of samples, or even technologies/platforms. Gene panels can be viewed as a "decompress" protocol to reduce required sample size for

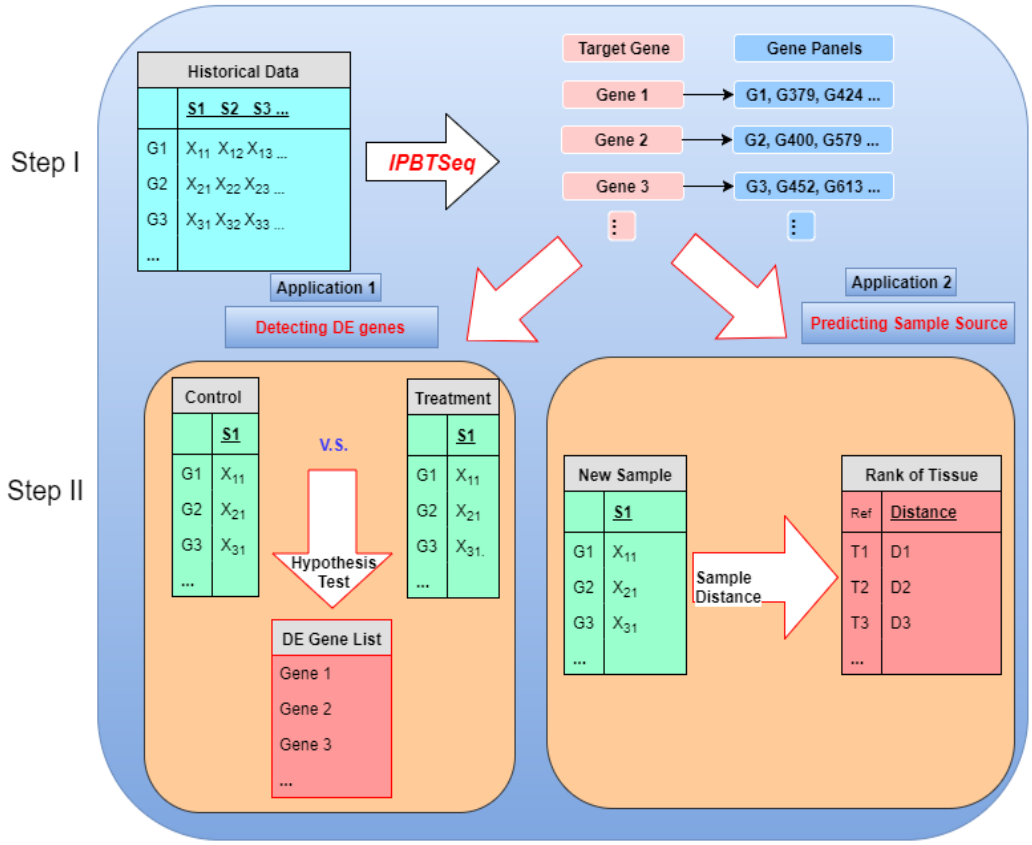

Figure 25 Workflow for IPBTSeq. Step I: Identify gene panels; Step II: Apply gene panels in different tasks

desired accuracy (Cleary, Cong, Lander, & Regev, 2017). We use external data to pre-define underlying structures of specific types of samples for certain characteristics by gene panels. When we have limited sample sizes in our current study, we "decompress" similar genes in the gene panel of the target gene to improve analysis results.

We will concentrate on gene expression specifically to illustrate our gene panel framework in most of our method, simulation and real data analysis sections. As Figure 25 shows, IPBTSeq consists of two steps. First, we identify a gene panel for each target gene based on historical data, then we use the gene panel in current experiment to help with inferences on the target gene.

The essential idea of our method is to impute the variance of a gene (or a feature) using the expression values from genes in a pre-defined gene panel measured in the same experiment. Assuming the gene panel is reasonably large (greater than 20), such variance imputation can be done within just a single sample. And since the variance is estimated intra-sample, normalization across sample is not necessary. The rationale for the gene panel is that genes in the genome are interconnected, which is reflected in the significant correlation observed in their expression measures. We assume such correlation are stable, at least in the same cell or tissue type and under similar condition. Therefore, we hypothesized that statistical properties such as mean and variance can be "imputed" from their "neighbouring" genes found in the gene panel from just a single sample. Once such gene panels are defined, we can use the information to impute the variance for any gene of interest. In order for this strategy to work, we need to test whether a robust gene panel can be identified for all genes in the human genome, and whether the variance imputed by the gene panel is accurate.

4.1.3 Identify gene panels

For a target gene a, the imputation gene panel (referred as "gene panel" hereafter) for gene a is defined as a group of genes $g_1, g_2, \ldots, g_{n_a}$ such that their sample mean and

variance are similar to the actual mean and variance of gene a: $\overline{var}(g_1, g_2, \ \ldots, g_{n_a}) \approx var(g_a)$ the hope is that we can define a gene panel consist of 20-500 genes for every gene in the human genome and for major cell or tissue types under normal conditions. By doing that, when statistical inference tasks such as detecting DE genes are needed on a new dataset, we can use the defined gene panel to impute variance.

Processed data from recount2 are sequencing counts. Similar to Limma-voom (Law et al., 2014), we work with $\log_2$CPM instead of raw counts. Assume $\log_2$CPM for gene g sample j follows normal distribution:

$$X_{gj} \sim Normal\left(\mu_g, \sigma_g^2\right), g = 1, \ldots G, j = 1, \ldots N \tag{47}$$

Here *G* is the number of genes in the human genome, *N* is the sample size. For each gene *i* (target gene), we define a panel index vector $W^{(i)} = \left(w_1^{(i)}, \ w_2^{(i)}, \ldots, w_G^{(i)}\right)^T$:

$$w_g^{(i)} = \begin{cases} 1, gene\ g\ in\ the\ panel\ of\ gene\ i \\ 0, gene\ g\ NOT\ in\ the\ panel\ of\ gene\ i \end{cases} \tag{48}$$

Our goal of panel selection is to identify "similar" genes so that using one target gene's panel genes in one sample can approximate the target gene's certain characteristic, which usually needs multiple samples to obtain a reliable/accurate estimate.

For example, if we would like to focus on detecting DE genes which requires reliable mean and variance estimate for each gene, we select those "similar" genes by means and variances so that using a target gene's panel genes within one sample could approximate the target gene's mean and variance. We first define mean and variance estimate for *i*th gene by its panel genes with only *j*th sample:

$$\text{Panel mean: } \bar{X}_j^{(i)} = \frac{\sum_g x_{gj} w_g^{(i)}}{\sum_g w_g^{(i)}} \tag{49}$$

$$\text{Panel variance: } S_j^{2(i)} = \frac{\sum_g \left(x_{gj} w_g^{(i)} - \bar{X}_j^{(i)} \right)^2}{\sum_g w_g^{(i)} - 1} \tag{50}$$

$$\text{Panel standard deviation (SD): } S_j^{(i)} = \sqrt{S_j^{2(i)}} \tag{51}$$

Then we define two loss functions:

$$L_1(W^{(i)}) = \sum_j \left(\bar{X}_j^{(i)} - \mu_i \right)^2 \tag{52}$$

$$L_2(W^{(i)}) = \sum_j \left(S_j^{(i)} - \sigma_i \right)^2 \tag{53}$$

The $W^{(i)}$ minimize L_1, L_2 is the best panel selection for approximation *i*th gene's mean and SD by its panel information, respectively. Noting that we assume the sample size N is large enough in the historical data so that the sample mean and sample variance for each gene provide accurate estimate of μ_i and σ_i^2. Therefore, we plug in $\widehat{\mu_i} = \frac{\sum_j x_{ij}}{N}$ and $\widehat{\sigma_i^2} = \frac{\sum_j (x_{ij} - \widehat{\mu_i})^2}{N-1}$ for μ_i and σ_i^2 when we minimize L_1 and L_2. In our context, we want to minimize L_1 and L_2 simultaneously. Thus, we work with objective function $L_3 = L_1 + L_2$.

Ideally, an exhaustive search of all possible combinations of $W^{(i)}$ could find its optimal solution. However, the number of genes, which is often more than 10,000, makes an exhaustive search computational infeasible. Even if we use a pre-screening procedure to pick a few hundred (say *k*) candidate panel genes, 2^k combinations of $W^{(i)}$ is still too large for an efficient exhaustive search.

Alternatively, we specify a Bayesian hierarchical model. Instead of fixing $w_g^{(i)}$ for all samples, we assume $w_g^{(i)} \sim Bernoulli(p_g^{(i)})$. In each sample *j*, we have a realization index vector $Z_j^{(i)} = \left(z_{1j}^{(i)},\ z_{2j}^{(i)}, \dots, z_{Gj}^{(i)}\right)^T$ of $W^{(i)}$. The full model is:

$$z_{gj}^{(i)} |\ p_g^{(i)}\ \sim Bernoulli(p_g^{(i)}) \tag{54}$$

$$\begin{aligned} \{X_{ij}, X_{gj}\} |\ z_{gj}^{(i)}, \mu_i, \sigma_i^2, \mu_g, \sigma_g^2 \sim\ & z_{gj}^{(i)} Normal(\mu_i, \sigma_i^2) \\ +\ & (1 - z_{gj}^{(i)}) Normal\left(\mu_g, \sigma_g^2\right) \end{aligned} \tag{55}$$

Our goal is to estimate $p_g^{(i)}$ and select genes with large $p_g^{(i)}$ as gene panels. The problem can be addressed by an expectation–maximization (EM) algorithm. However, when we consider our desired genes to be selected in the panel, the distribution could be extremely close to our target gene's distribution. In that case, the proposed two component mixture model degenerates to a one-component normal distribution. Standard EM fails in these desired scenarios for our gene selection. Hence, we use a regularized EM with conditional entropy $H(z_{gj}^{(i)} \mid X_{ij}, X_{gj};\ \mu_i, \sigma_i^2, \mu_g, \sigma_g^2)$ as the penalty term (H. Li, Zhang, & Jiang, 2005). We define the regularized log likelihood

$$\begin{aligned} &\widetilde{L}\,(p_g^{(i)}, z_{gj}^{(i)}, \mu_i, \sigma_i^2, \mu_g, \sigma_g^2 \mid X_{ij}, X_{gj}) \\ &\quad = L(p_g^{(i)}, z_{gj}^{(i)}, \mu_i, \sigma_i^2, \mu_g, \sigma_g^2 \mid X_{ij}, X_{gj}) \\ &\quad - \eta_0\, H(z_{gj}^{(i)} \mid X_{ij}, X_{gj};\ \mu_i, \sigma_i^2, \mu_g, \sigma_g^2) \end{aligned} \tag{56}$$

The regularized EM enables combining components when two components are similar. In our scenario, when gene g is similar to our target gene *i*, the EM will end up

combining two components. Thus, we apply regularized EM to our target gene and candidate gene and determine whether the candidate is selected by whether the two components have been combined in the end. We list details of regularized EM below.

E step (same as standard EM) :

$$\alpha_{gj}^{(i)} = E(z_{gj}^{(i)} | X_{ij}, X_{gj}; \mu_i, \sigma_i^2, \mu_g, \sigma_g^2)$$

M step (maximize regularized complete log likelihood):

$$p_g^{(i)} = \frac{\sum_j \alpha_{gj}^{(i)}(1 + \eta_0 log\alpha_{gj}^{(i)})}{\sum_j(\alpha_{gj}^{(i)}\left(1 + \eta_0 log\alpha_{gj}^{(i)}\right) + (1 - \alpha_{gj}^{(i)})(1 + \eta_0 log(1 - \alpha_{gj}^{(i)})))}$$

$$\mu_{g1}^{(i)} = \frac{\sum_j X_{gj}\alpha_{gj}^{(i)}(1 + \eta_0 log\alpha_{gj}^{(i)})}{\sum_j \alpha_{gj}^{(i)}\left(1 + \eta_0 log\alpha_{gj}^{(i)}\right)}$$

$$\mu_{g2}^{(i)} = \frac{\sum_j X_{gj}(1 - \alpha_{gj}^{(i)})(1 + \eta_0 \log\left(1 - \alpha_{gj}^{(i)}\right))}{\sum_j(1 - \alpha_{gj}^{(i)})(1 + \eta_0 log(1 - \alpha_{gj}^{(i)}))}$$

$${\sigma^2}_{g1}^{(i)} = \frac{\sum_j\left(X_{gj} - \mu_{g1}^{(i)}\right)^2 \alpha_{gj}^{(i)}(1 + \eta_0 log\alpha_{gj}^{(i)})}{\sum_j \alpha_{gj}^{(i)}\left(1 + \eta_0 log\alpha_{gj}^{(i)}\right)}$$

$${\sigma^2}_{g2}^{(i)} = \frac{\sum_j\left(X_{gj} - \mu_{g1}^{(i)}\right)^2 (1 - \alpha_{gj}^{(i)})(1 + \eta_0 \log\left(1 - \alpha_{gj}^{(i)}\right))}{\sum_j(1 - \alpha_{gj}^{(i)})(1 + \eta_0 log(1 - \alpha_{gj}^{(i)}))}$$

In theory, when we consider target gene *i*, all other *G*-1 genes could be a candidate for gene *i*'s panel. Hence, $(G - 1) * G$ EMs are required. Although a single regularized EM converges in a few seconds in a typical personal laptop, the tremendous number of EM steps required here makes it computationally intensive without

parallelization when implemented on a laptop or PCs. We conduct a pre-scan procedure before formally running EM algorithm since we can rule out significantly "dissimilar" genes easily by a straightforward ranking of gene's means and standard deviations. After the pre-scan procedures, top *k* "similar" genes are chosen for each target genes. The *k* can be customized by users and our default choice are 100/200/500.

Even if we conduct some pre-scan procedures and choose top 100 candidate genes, 10,000 genes require 1,000,000 EMs. The computational burden is acceptable, but may be inconvenient for analysts with limited computational sources. Therefore, we also provide a simpler version to identify gene panels. In the simpler version, two tuning parameters η_1, η_2 are defined. We only need to calculate Bayes Factors:

$$BF_{gj}^{(i)} = \frac{Normal\left(X_{gj} \mid \mu_i, \sigma_i^2\right)}{Normal\left(X_{gj} \mid \mu_g, \sigma_g^2\right)} \tag{57}$$

Then based on $BF_{gj}^{(i)}$, we determine $z_{gj}^{(i)}$ by:

$$\begin{cases} z_{gj}^{(i)} = 1, \text{if } BF_{gj}^{(i)} > \eta_1 \\ z_{gj}^{(i)} = 0, \text{otherwise} \end{cases}$$

After determining all $z_{gj}^{(i)}$, we estimate$p_g^{(i)}$ by $p_g^{(i)} = \frac{\sum_j z_{gj}^{(i)}}{J}$.

Finally, we consider genes in $\left\{g: p_g^{(i)} > \eta_2\right\}$ as gene panels for our target gene *i*.

4.1.4 Distance and Imputation Score

We use mean square error (MSE) to define distance between a reference panel and a RNA-Seq sample. To be more specific, a gene panel P is determined by panel index

vectors $\{W^{(i)}: 1 \leq i \leq G\}$. With a RNA-Seq sample S_j (its values are denoted as X_{gj}), we can calculate all panel means $\{\bar{X}_J^{(i)}: 1 \leq i \leq G\}$ by equation (2). Hence, we can define distance between reference P and sample S_j as:

$$Dist(P, S_j) = \frac{\sum_i \left(\bar{X}_J^{(i)} - X_{ij}\right)^2}{G} \tag{58}$$

It is worth noting that the loss function in (52) used historical data to generate gene panels. It served as "training" object function to obtain an established gene panel. Here, equation (58) uses panel information to get "predicted" panel mean and we calculate a "testing" loss function with "prediction" and true expressions X_{ij}. When we have a new RNA-Seq sample, we can calculate their distance to our pre-defined panels or user customized panels. If some of the panels have a distance smaller than a pre-specified cutoff, we can use the panel with smallest distance for further analysis application. Otherwise, all the panels are not appropriate, a new panel is required for the analysis or more replicates are recommended for reliable analysis.

Similarly, we also define an imputation score *IS* for each gene within a given panel. This score serves as a quality measure for the imputation. A high quality score suggests that the imputed statistics are trustworthy and a low score otherwise. A poor imputation score is a warning that indicates the gene is intrinsically difficult or impossible to impute so that it is better to exclude the gene in our latter analysis. This is similar to the genotype imputation commonly used in statistical genetics. To be specific, for gene *g* and panel *P*, we have:

$$\text{IS}\ (P, g) = \frac{\sum_j \left(\bar{X}_J^{(g)} - X_{gj}\right)^2}{J} \tag{59}$$

where $\bar{X}_J^{(g)}$ is panel mean estimated using panel P on sample S_j and X_{gj} is the observed value for gene g in sample j.

4.2 Simulation Study

I conducted two sets of data based simulation studies to illustrate the consistency of gene panels and its applications to DE gene detection.

4.2.1 Validation of gene panels

To illustrate that our reference panel is capable of imputing and the panel is tissue-specific, we calculated the "distance" between each new "unknown" sample with all of our reference panels. In this study, the panel is constructed using GTEx data on 30 tissues. We then go through each individual TCGA sample, impute the variance for all genes in the human genome using their corresponding panel. We pretend to have no information about the source for TCGA samples and use the distance between each new sample and reference panel to determine the source for TCGA sample.

Figure 26(a) - (d) shows the results for TCGA-LUAD (lung), TCGA-KIRC (kidney), TCGA-LIHC (liver), TCGA-BRCA (breast) samples, respectively. In general, all results indicate that the reference panel is informative to "predict" the unknown sample's source.

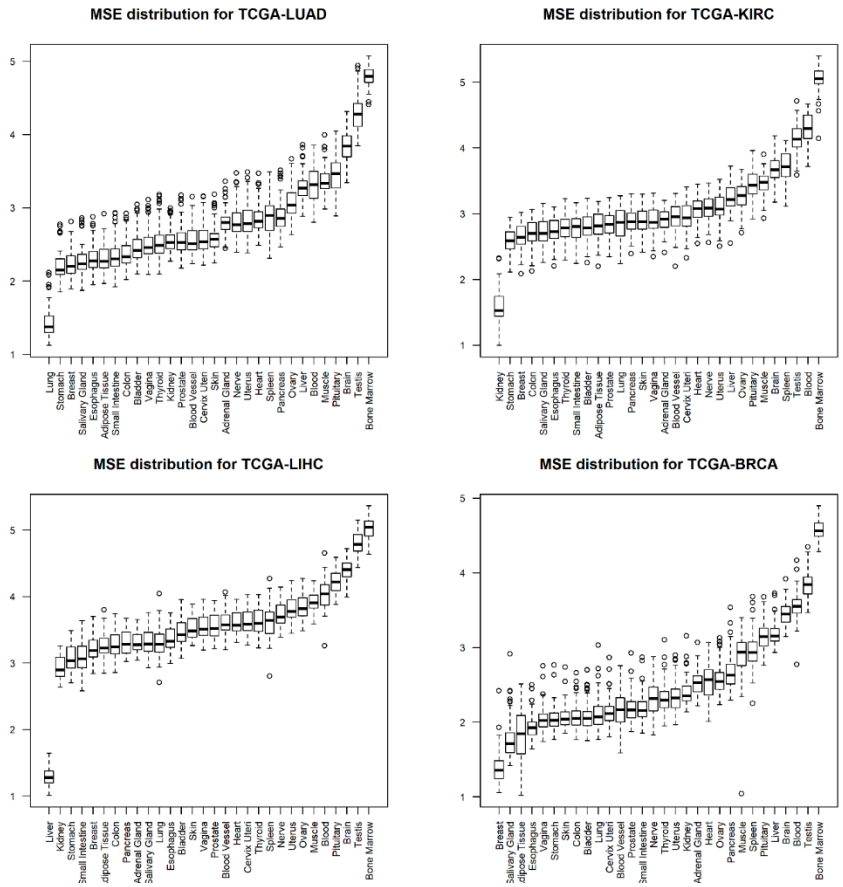

Figure 26 Predict new "unknown" samples. For any new samples, distance between the new sample and predefined panels are calculated. The panel with smallest distance can used to predict the source of the new sample

In Figure 27 (a) – (c), we show the landscape of imputation scores for kidney, liver and lung. The 90% and 95% quantile red lines shows that most genes have well imputed variances. Figure 27 (d) shows a Venn diagram for outliers for different tissues. The Venn diagram demonstrates that the outliers are tissue-specific, which indicates that panels could be defined by tissue sources.

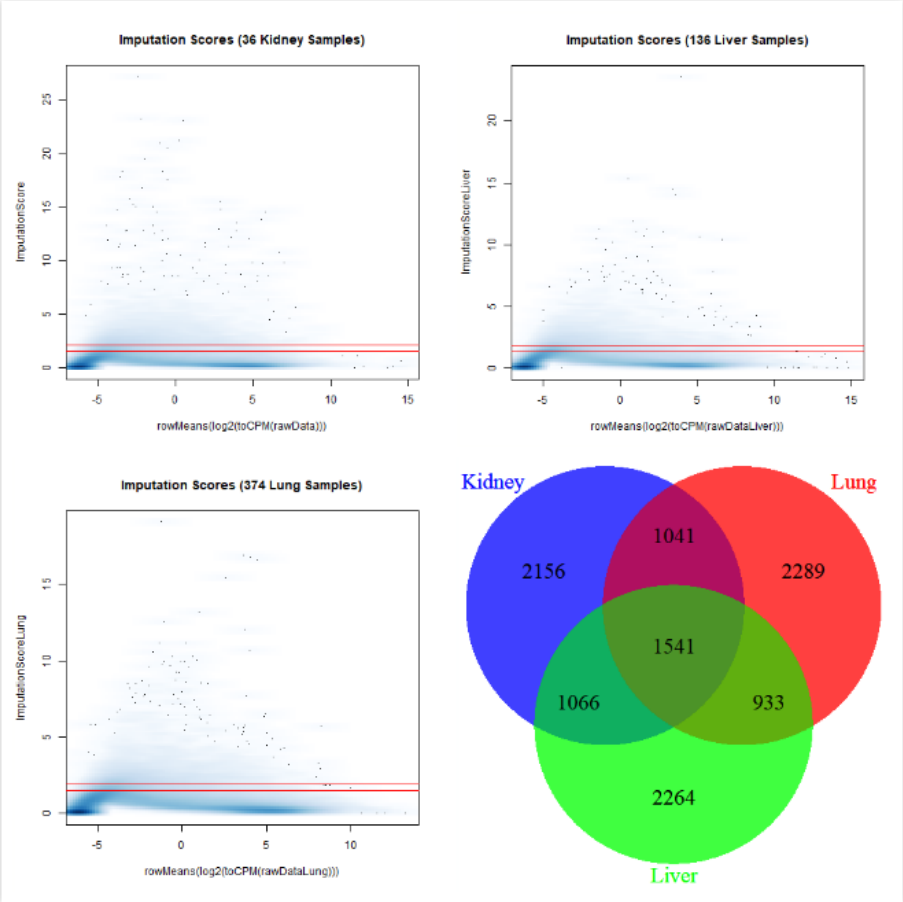

Figure 27 Landscape for Imputation Score. (a)-(c): Imputation score distributions for kidney, liver, lung samples, respectively. (d) Number of overlapped low quality imputed genes for kidney, liver and lung.

4.2.2 Detect DE Genes

In this simulation study, we are going to use gene panels to identify DE genes. Although our major assumption is that $\log_2$CPM follows a normal distribution, we use a negative binomial distribution for counts to generate our simulation data. We follow the procedure of generating simulation data of DESeq2 (Love et al., 2014). Two samples are simulated in control and treatment group, respectively. 20000 genes are generated and 10% of them are DE genes with three-fold change. 50 samples are simulated as external data. Figure 28 shows false discoveries and false discovery rate when identify top genes as DE genes.

We can see that with the help of gene panels, IPBTSeq could achieve a better performance than all existing methods. We use the imputation score in final results to exclude about 5% to 10% genes with unreliable imputation variance.

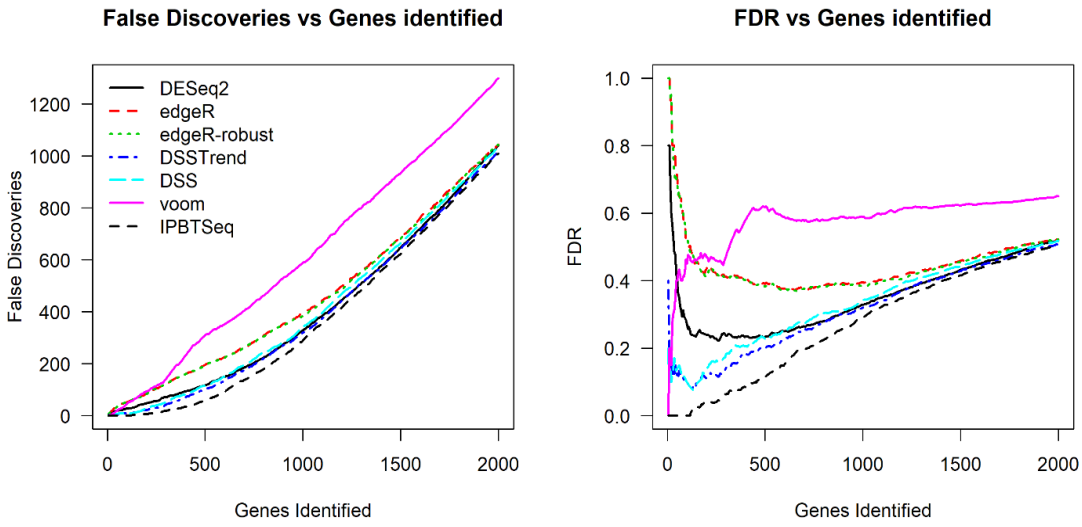

Figure 28 (a) Number of false discoveries and (b) False discovery rates for detecting DE genes in the simulation study.

Another advantage for gene panels is that detecting DE genes becomes possible even without replicates. In Figure 29, we add two more "scenarios" called IPBTSeq1 and IPBTSeq2. IPBTSeq1 only includes the first sample in the first condition and the first sample in the second condition while IPBTSeq2 only includes the second sample in the first condition and the second sample in the second condition. All other methods include two samples in the first condition and two samples in the second condition. We can see that with the help of gene panels, IPBTSeq1 and IPBTSeq2 are better than Limma-voom and similar to edgeR-robust, although not as good as DSS or DESeq2.

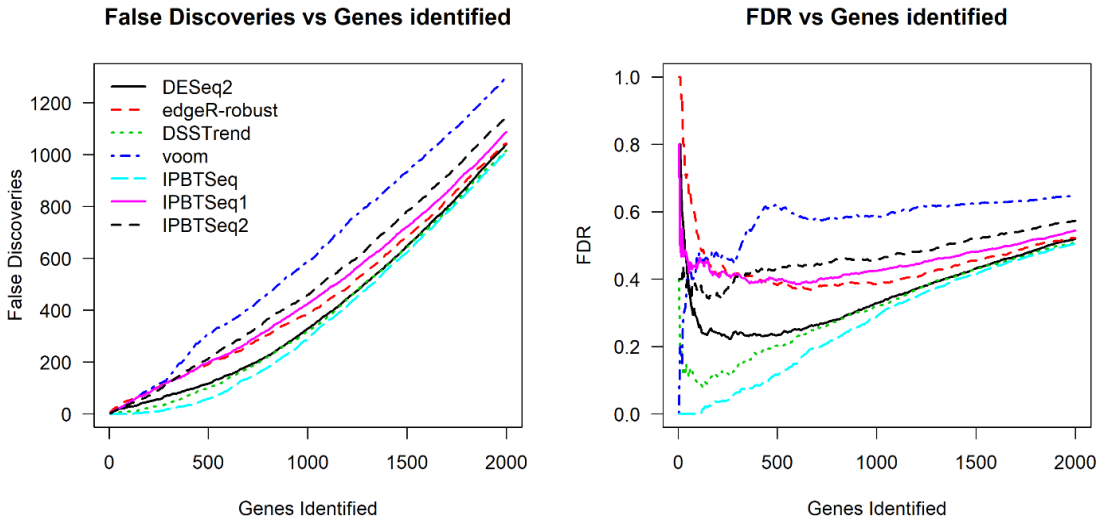

Figure 29 Detecting DE Genes without Replicates (a) Number of false discoveries and (b) False discovery rates

4.3 Real Data Analysis

In this section, we are going to use real data to explore the properties for gene panels and apply it to real data.

4.3.1 Landscape for gene panels

We construct gene panels for 30 tissue types for which at least 10 samples were present in the current GTEx data. Figure 30 depicts distributions of gene panels with top 200 pre-selected candidate genes with default tuning parameters. Figure 30(a) indicates that most of the genes in the human genome have large enough panels to be used in following tasks. Although different tissues have a slightly shifted distributions, they share a common general pattern. We can further investigate the distribution of the number of

genes included in the panel with Figure 30 (b)-(d). Figure 30 (b) – (d) shows the mean and standard deviation plots colored by different numbers of genes in its gene panel for liver, kidney and lung, respectively.

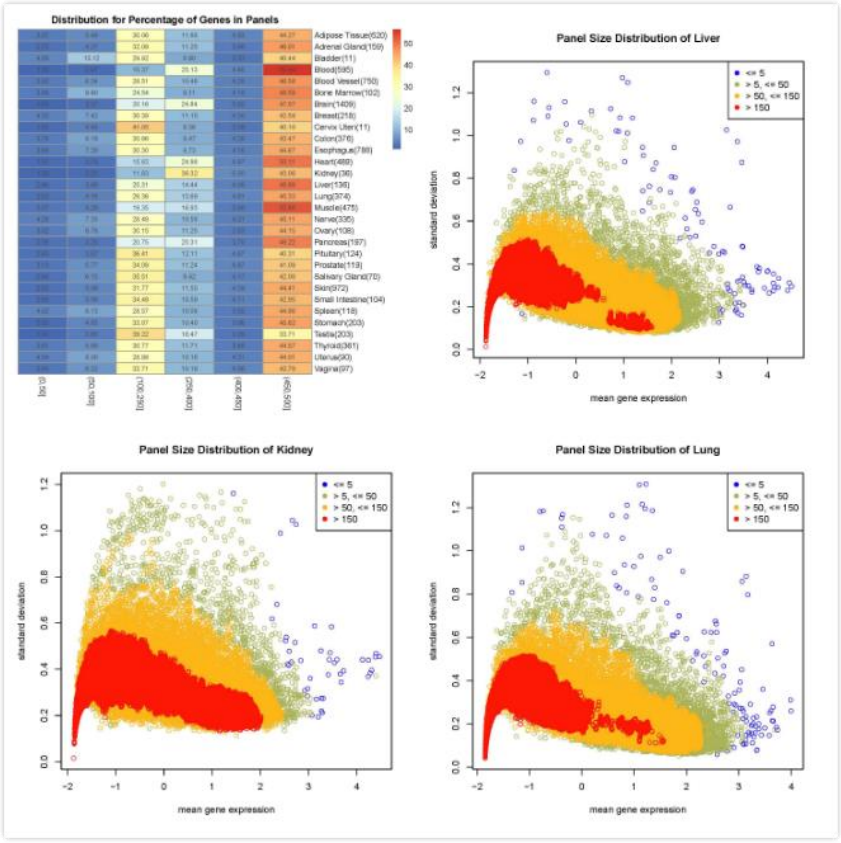

Figure 30 Landscape for gene panels. (a) Distribution of number of genes included in the panel for different tissues. (b)-(d) Detailed distribution for liver, kidney and lung.

The patterns for different tissues are slight different while they share some common trends. In general, low mean and low SD target genes can easily have a large panel while genes with high mean or high SD have much smaller ones. The patterns shown in Figure 30 (b) – (d) is as expected since high mean or high SD genes tend to be more variable and if could be extremely difficult to "impute" their values by similar genes.

4.3.2 Detect DE Genes

Another advantage of our normal based assumption is that it could be easily applied to other data types such as gene expression microarray, 450k array. Similar to Figure 22 and 23, we apply our new methods with gene panels to microarray data. The gene panels are generated from corresponding microarray data.

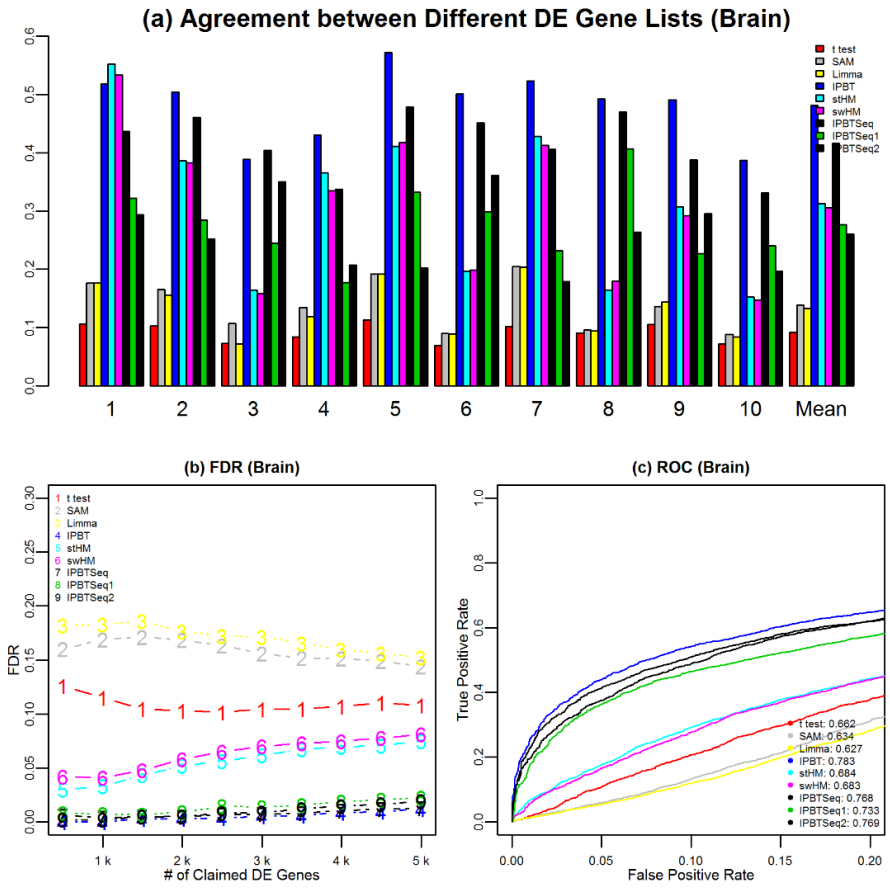

Figure 31 Real Data Analysis. (a) Agreement for different methods. (b) False discovery rates. (c) ROC curves.

Figure 31 (a) shows that IPBTSeq with the same sample size has better performance than existing methods and adaptiveHM while slightly worse performance than IPBT. IPBTSeq without repliseq is slightly worse than adaptive HM but better than other state-of-the-art methods.

4.4 Discussion and Conclusion

In the third topic, I propose a new framework to summarize information from external data by “gene panels” to improve the inference on new experimental data. The core part of the approach is to use a penalized EM to define a gene panel for each target gene in the genome such that genes in the gene panel are similar enough to the target gene and could be treated as “pseudo replicates”. We explore properties for gene panels with data from GTEx and TCGA samples and validate that gene panels are consistent for the same tissues even from different data sources. By simulation studies and real data analysis, we show examples of how gene panels could help in finding DE genes. IPBTSeq with gene panels improves data analytics results appreciably over existing methods. Moreover, gene panels make it possible to conduct analysis when there are no replicates.

The reason why this strategy works is because information contained in the gene expression measures in the genome is redundant. The expressions of many genes are more or less correlated. For most genes, its expression measure can be inferred with reasonable accuracy from other genes. However, most of state-of-the-art methods handle genes independently without considering any further correlation structure between different genes. In our first and second topics, we are also assuming all genes are independent. In our third topic, although our major assumption remains that all genes are independent, we consider an implicit correlation structure between genes with our gene panels. A further rigorous model of gene panels explicitly using gene-gene correlation structures may be more attractive in both theory and application.

Our procedure to define gene panels is data driven. Hence, the panels may not have much biological interpretation and they are different from a group of genes defined

by biological interpretation such as gene pathways. The advantage for the data driven procedure is that it could be more flexible and tailor the demands for different genome analysis tasks. On the other hand, high-quality external data are desired. Using the wrong panels could lead to nonsense results or flawed discoveries. This merits further investigation.

Besides being tissue specific, gene panels are likely technology specific. For example, panels for microarrays may be different from those for RNA-seq due to normalization issues even if the data are measuring the same samples. It is possible to borrow panel information for new technology from old ones such as from microarray to RNA-Seq, from bulk RNA-Seq to single cell RNA-Seq. However, users should be extremely cautious about normalization issues when running IPBTSeq across different technologies.

Chapter 5

Summary and Future Work

In this dissertation, I propose multiple model-based methods for improving genomics data analysis by incorporating existing external datasets.

In the first topic, I utilize informative priors from external data and use those informative priors to improve DE gene detection results with a Bayesian framework (IPBT). To assess the success of IPBT, I use a normal-inv$-\chi 2$ model on gene expression microarray data and Bayes factors (BF) are calculated to rank the top DE genes. To compare with existing methods, I showed that IPBT is equivalent to an adjusted *t*-test in terms of gene ranks. Extensive simulation studies and real data analyses are conducted to demonstrate the advantages of IPBT. These results also illuminate the possibility of utilizing the increasingly available genomics data in statistical inference and provide an alternative practical strategy to deal with the 'large *p*, small *n*' problem. R package IPBT with 96 informative priors is freely available from https://github.com/benliemory/IPBT.

In my second topic, rank-based strategies are proposed to use external information for new datasets (adaptiveHM). I use ranks from external data to define groups for new experiments. A state-of-the-art Bayesian hierarchical model can then be adopted to shrink estimates of standard deviations (SD) within each group. I also propose a group dividing metric (GDM) to decide the optimal number of groups. Massive simulations and real data analysis are conducted to illustrate that adaptiveHM can be applied to different types of data such as gene expression microarray, 450K methylation array, RNA-Seq and BS-Seq.

The results show that adaptiveHM could have similar performances with IPBT. More importantly, adaptiveHM enables borrowing information across different platforms.

In the third topic, a more general framework is proposed to extend the procedure of summarizing information from historical data by "gene panels" (IPBTSeq). IPBTSeq uses a penalized EM to define a gene panel for each target gene in the first step and uses gene panels for further data analysis in the second step. Genes in the gene panel are close enough in terms of certain statistics to the target gene and are regarded as "pseudo replicates". We use normal samples from GTEx and TCGA to demonstrate properties of IPBTSeq and validate the feasibility and effectiveness for gene panels. We conduct simulation studies and real data analysis to show examples of how to utilize gene panels.

In the third topic, we implicitly consider correlation structures between genes using gene panels. A formal model or framework to explore gene-gene correlation structures from external data may be useful in both theory and applications. In addition, all three topics of this dissertation and most of existing methods focus on the problem of testing the null hypotheses $H_0: \beta_{g1} = \beta_{g2}$ and attempt to construct improved test statistics. In many genomics discovery research projects using high-throughput genomics data, investigators are more interested in the ranks of genes by evidence against null hypotheses instead of assigning absolute *p*-values since only a limited number of genes would be followed up for further analysis despite the exact *p*-values are significant or false discovery rates are controlled. Even when the major distributional assumptions fail for certain datasets, the test statistics may perform well from a ranking perspective. This is one reason that most state-of-the-art methods adopt an asymptotic test while only two or three samples

involved. Hence, it could also be useful to explore more about how to assign more meaningful and reliable *p*-value and false discovery rate for results.

Bibliography

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol, 11*(10), R106. doi:10.1186/gb-2010-11-10-r106

Arima, S., Liseo, B., Mariani, F., & Tardella, L. (2011). Exploiting blank spots for model-based background correction in discovering genes with DNA array data. *Statistical Modelling, 11*(2), 89-114. doi:10.1177/1471082x1001100201

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., & Irizarry, R. A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics, 30*(10), 1363-1369. doi:10.1093/bioinformatics/btu049

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., . . . Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet, 25*(1), 25-29. doi:10.1038/75556

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol, 57*(1), 289-300.

Chen, M. H., & Ibrahim, J. G. (2006). The Relationship Between the Power Prior and Hierarchical Models. *Bayesian Analysis, 1*(3), 551-574.

Cleary, B., Cong, L., Lander, E., & Regev, A. (2017). Composite measurements and molecular compressed sensing for highly efficient transcriptomics. *bioRxiv*, 091926.

Collado-Torres, L., Nellore, A., Kammers, K., Ellis, S. E., Taub, M. A., Hansen, K. D., . . . Leek, J. T. (2017). Reproducible RNA-seq analysis using recount2. *Nat Biotechnol, 35*(4), 319-321. doi:10.1038/nbt.3838

Conlon, E. M., Song, J. J., & Liu, A. (2007). Bayesian meta-analysis models for microarray data: a comparative study. *BMC Bioinformatics, 8*, 80. doi:10.1186/1471-2105-8-80

Daigle, B. J., Jr., Deng, A., McLaughlin, T., Cushman, S. W., Cam, M. C., Reaven, G., . . . Altman, R. B. (2010). Using pre-existing microarray datasets to increase experimental power: application to insulin resistance. *PLoS Comput Biol, 6*(3), e1000718. doi:10.1371/journal.pcbi.1000718

Duan, Y. Y., Ye, K. Y., & Smith, E. P. (2006). Evaluating water quality using power priors to incorporate historical information. *Environmetrics, 17*(1), 95-106. doi:10.1002/env.752

Fan, J., Han, F., & Liu, H. (2014). Challenges of Big Data Analysis. *Natl Sci Rev, 1*(2), 293-314. doi:10.1093/nsr/nwt032

Fan, J., & Lv, J. (2010). A Selective Overview of Variable Selection in High Dimensional Feature Space. *Stat Sin, 20*(1), 101-148.

Feng, H., Conneely, K. N., & Wu, H. (2014). A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res, 42*(8), e69. doi:10.1093/nar/gku154

Ganjali, M., Baghfalaki, T., & Berridge, D. (2015). Robust modeling of differential gene expression data using normal/independent distributions: a Bayesian approach. *PLoS One, 10*(4), e0123791. doi:10.1371/journal.pone.0123791

Gelman, A. (2004). *Bayesian Data Analysis* (2nd ed.). Boca Raton, Fla.: Chapman & Hall/CRC.

Ghosh, D., & Qin, Z. S. (2010). Statistical Issues in the Analysis of ChIP-Seq and RNA-Seq Data. *Genes, 1*(2), 317-334. doi:10.3390/genes1020317

Good, I. J. (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. Cambridge, Mass: M.I.T. Press.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning : data mining, inference, and prediction* (2nd ed.). New York: Springer.

Hobbs, B. P., Carlin, B. P., Mandrekar, S. J., & Sargent, D. J. (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics, 67*(3), 1047-1056. doi:10.1111/j.1541-0420.2011.01564.x

Huang da, W., Sherman, B. T., & Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res, 37*(1), 1-13.

Huang da, W., Sherman, B. T., & Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc, 4*(1), 44-57.

Ibrahim, J. G., Chen, M. H., Gwon, Y., & Chen, F. (2015). The power prior: theory and applications. *Stat Med, 34*(28), 3724-3749. doi:10.1002/sim.6728

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., & Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics, 4*(2), 249-264. doi:10.1093/biostatistics/4.2.249

Ji, H., & Liu, X. S. (2010). Analyzing 'omics data using hierarchical models. *Nat Biotechnol, 28*(4), 337-340. doi:10.1038/nbt.1619

Ji, H., & Wong, W. H. (2005). TileMap: create chromosomal map of tiling array hybridizations. *Bioinformatics, 21*(18), 3629-3636.

Kerr, M. K., & Churchill, G. A. (2001). Experimental design for gene expression microarrays. *Biostatistics, 2*(2), 183-201. doi:10.1093/biostatistics/2.2.183

Kim, R. D., & Park, P. J. (2004). Improving identification of differentially expressed genes in microarray studies using information from public databases. *Genome Biol, 5*(9), R70. doi:10.1186/gb-2004-5-9-r70

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol, 15*(2), R29. doi:10.1186/gb-2014-15-2-r29

Lewin, A., Richardson, S., Marshall, C., Glazier, A., & Aitman, T. (2006). Bayesian modeling of differential gene expression. *Biometrics, 62*(1), 1-9. doi:10.1111/j.1541-0420.2005.00394.x

Li, B., Li, Y., & Qin, Z. S. (2016). Improving Hierarchical Models Using Historical Data with Applications in High-Throughput Genomics Data Analysis. *Statistics in Biosciences*, 1-18.

Li, B., Sun, Z., He, Q., Zhu, Y., & Qin, Z. S. (2015). Bayesian inference with historical data-based informative priors improves detection of differentially expressed genes. *Bioinformatics*. doi:10.1093/bioinformatics/btv631

Li, H., Zhang, K., & Jiang, T. (2005). *The regularized EM algorithm.* Paper presented at the AAAI.

Lim, K., Li, Z., Choi, K. P., & Wong, L. (2015). A quantum leap in the reproducibility, precision, and sensitivity of gene expression profile analysis even when sample size is extremely small. *J Bioinform Comput Biol, 13*(4), 1550018. doi:10.1142/S0219720015500183

Lim, K., & Wong, L. (2014). Finding consistent disease subnetworks using PFSNet. *Bioinformatics, 30*(2), 189-196. doi:10.1093/bioinformatics/btt625

Lo, K., & Gottardo, R. (2007). Flexible empirical Bayes models for differential gene expression. *Bioinformatics, 23*(3), 328-335. doi:10.1093/bioinformatics/btl612

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol, 15*(12), 550. doi:10.1186/s13059-014-0550-8

Lukk, M., Kapushesky, M., Nikkila, J., Parkinson, H., Goncalves, A., Huber, W., . . . Brazma, A. (2010). A global map of human gene expression. *Nat Biotechnol, 28*(4), 322-324. doi:10.1038/nbt0410-322

McCall, M. N., Bolstad, B. M., & Irizarry, R. A. (2010). Frozen robust multiarray analysis (fRMA). *Biostatistics, 11*(2), 242-253. doi:10.1093/biostatistics/kxp059

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods, 5*(7), 621-628.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., & Tsui, K. W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J Comput Biol, 8*(1), 37-52. doi:10.1089/106652701300099074

Parmigiani, G., Garett, E. S., Irizarry, R. A., & Zeger, S. L. (2003). *The analysis of gene expression data : methods and software*. New York: Springer.

Qin, Z., Li, B., Conneely, K. N., Wu, H., Hu, M., Ayyala, D., . . . Lin, S. (2016). Statistical Challenges in Analyzing Methylation and Long-Range Chromosomal Interaction Data. *Statistics in Biosciences, 8*(2), 284-309. doi:10.1007/s12561-016-9145-0

Robinson, M. D., Kahraman, A., Law, C. W., Lindsay, H., Nowicka, M., Weber, L. M., & Zhou, X. (2014). Statistical methods for detecting differentially methylated loci and regions. *Front Genet, 5*, 324. doi:10.3389/fgene.2014.00324

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics, 26*(1), 139-140. doi:10.1093/bioinformatics/btp616

Robinson, M. D., & Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics, 23*(21), 2881-2887. doi:10.1093/bioinformatics/btm453

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol, 3*, Article3. doi:10.2202/1544-6115.1027

Soh, D., Dong, D., Guo, Y., & Wong, L. (2011). Finding consistent disease subnetworks across microarray datasets. *BMC Bioinformatics, 12 Suppl 13*, S15. doi:10.1186/1471-2105-12-S13-S15

Sui, Y., Zhao, X., Speed, T. P., & Wu, Z. (2009). Background adjustment for DNA microarrays using a database of microarray experiments. *J Comput Biol, 16*(11), 1501-1515. doi:10.1089/cmb.2009.0063

Tseng, G. C., Ghosh, D., & Feingold, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res, 40*(9), 3785-3799. doi:10.1093/nar/gkr1265

Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A, 98*(9), 5116-5121.

Wu, H., Wang, C., & Wu, Z. (2013). A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics, 14*(2), 232-243. doi:10.1093/biostatistics/kxs033

Wu, H., Xu, T., Feng, H., Chen, L., Li, B., Yao, B., . . . Conneely, K. N. (2015). Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic Acids Res, 43*(21), e141. doi:10.1093/nar/gkv715

Wu, Z., & Irizarry, R. A. (2004). Preprocessing of oligonucleotide array data. *Nat Biotechnol, 22*(6), 656-658; author reply 658. doi:10.1038/nbt0604-656b