1-1-2013

Statistical Methods for Analysis of Multi-Sample Copy Number Variants and ChIP-seq Data

Qian Wu
University of Pennsylvania, wuqian7@gmail.com

Statistical Methods for Analysis of Multi-Sample Copy Number Variants and ChIP-seq Data

Abstract
This dissertation addresses the statistical problems related to multiple-sample copy number variants (CNVs) analysis and analysis of differential enrichment of histone modifications (HMs) between two or more biological conditions based on the Chromatin Immunoprecipitation and sequencing (ChIP-seq) data. The first part of the dissertation develops methods for identifying the copy number variants that are associated with trait values. We develop a novel method, CNVtest, to directly identify the trait-associated CNVs without the need of identifying sample-specific CNVs. Asymptotic theory is developed to show that CNVtest controls the Type I error asymptotically and identifies the true trait-associated CNVs with a high probability. The performance of this method is demonstrated through simulations and an application to identify the CNVs that are associated with population differentiation.

The second part of the dissertation develops methods for detecting genes with differential enrichment of histone modification between two or more experimental conditions based on the ChIP-seq data. We apply several nonparametric methods to identify the genes with differential enrichment. The methods can be applied to the ChIP-seq data of histone modification even without replicates. It is based on nonparametric hypothesis testing in order to capture the spatial differences in protein-enriched profiles. The key of our approaches is to use null genes or input ChIP-seq data to choose the biologically relevant null values of the tests. We demonstrate the method using ChIP-seq data on a comparative epigenomic profiling of adipogenesis of murine adipose stromal cells. Our method detects many genes with differential H3K27ac levels at gene promoter regions between proliferating preadipocytes and mature adipocytes in murine 3T3-L1 cells. The test statistics also correlate well with the gene expression changes and are predictive of gene expression changes, indicating that the identified differential enrichment regions are indeed biologically meaningful.

We further extend these tests to time-course ChIP-seq experiments by evaluating the maximum and mean of the adjacent pair-wise statistics for detecting differentially enriched genes across several time points. We compare and evaluate different nonparametric tests for differential enrichment analysis and observe that the kernel-smoothing methods perform better in controlling the Type I errors, although the ranking of genes with differentially enriched regions are comparable using different test statistics.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Epidemiology & Biostatistics

First Advisor
Hongzhe Li

Keywords
ChIP-seq, CNV, Histone modification, Kernel-smoothing, Multi-sample, Nonparametric test

Subject Categories
Biostatistics

STATISTICAL METHODS FOR ANALYSIS OF MULTI-SAMPLE COPY NUMBER VARIANTS AND CHIP-SEQ DATA

Qian Wu

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2013

Supervisor of Dissertation

_____

Hongzhe Li, Professor of Biostatistics

Graduate Group Chairperson

_____

Daniel F. Heitjan, Professor of Biostatistics

Dissertation Committee
Mingyao Li, Associate Professor of Biostatistics
Sarah Ratcliffe, Associate Professor of Biostatistics
Kyoung-Jae Won, Research Assistant Professor of Genetics
Nancy R. Zhang, Associate Professor of Statistics

STATISTICAL METHODS FOR ANALYSIS OF MULTI-SAMPLE COPY NUMBER VARIANTS AND CHIP-SEQ DATA

ACKNOWLEDGEMENT

I have a lot of good memories and experiences at Penn. During the completion of my Ph.D. dissertation, many professors, colleagues and friends have given me valuable support and encouragement. I would like to express my sincerely appreciation to all the people who helped me during my doctoral years.

First and foremost, I would like to thank my advisor, Dr. Hongzhe Li. I have been working with Dr. Li since my first year at Penn. Under his guidance, I developed a keen interest in statistical genetics. He led me into the projects quickly, taught me how to think and grow as a statistician and provided me the opportunity to collaborate with others. He is always there for giving me suggestions and guidance. I felt very lucky to have Dr. Li as my advisor. My father was an oncologist, though he passed away when I were young, and Dr. Li has given me support like my father, not only as an excellent mentor but also guiding me with sincere and selfless advice for career and life.

I would also like to thank Dr. Jonas Ellenberg for his patience and encouragement during my Ph.D. study. He was my academic advisor for the first two years and was extremely supportive of both my academic and career development. I would also like to thank Dr. James Dignam and Dr. Ed Zhang for supervising my research work at Radiation Therapy Oncology Group (RTOG). I also enjoyed my time as a summer intern for Dr. Xiaohua Douglas Zhang at Merck. He opened the door to statistical genetics and gave me tremendous encouragement and provided continuous opportunities for collaboration. These working experiences have helped me to grow quickly as a biostatistician.

I am deeply grateful to the members of my dissertation committee, Dr. Nancy Zhang, Dr. Mingyao Li, Dr. Sarah Ratcliffe and Dr. Kyoung-Jae Won, for their effort and inspiring suggestions throughout this process. Special thanks to Dr. Won for his help in teaching me ChIP-seq data and related biological questions for my dissertation. I would like to thank Dr. Jessie Jeng for her continuously support and help in my first CNV project.

Last but not least, I want to thank my fiance Lin Chai, my dear mother, and my grandmother for their love and support. Their support has helped me to continuously pursue my dream without worrying about long distance or any difficulties. I also appreciate all of the people I have met at Penn. Without their help, I would not have grown as quickly as I did. Thanks!

ABSTRACT

STATISTICAL METHODS FOR ANALYSIS OF MULTI-SAMPLE COPY NUMBER VARIANTS AND CHIP-SEQ DATA

Qian Wu

Hongzhe Li

This dissertation addresses the statistical problems related to multiple-sample copy number variants (CNVs) analysis and analysis of differential enrichment of histone modifications (HMs) between two or more biological conditions based on the Chromatin Immunoprecipitation and sequencing (ChIP-seq) data. The first part of the dissertation develops methods for identifying the copy number variants that are associated with trait values. We develop a novel method, CNVtest, to directly identify the trait-associated CNVs without the need of identifying sample-specific CNVs. Asymptotic theory is developed to show that CNVtest controls the Type I error asymptotically and identifies the true trait-associated CNVs with a high probability. The performance of this method is demonstrated through simulations and an application to identify the CNVs that are associated with population differentiation.

The second part of the dissertation develops methods for detecting genes with differential enrichment of histone modification between two or more experimental conditions based on the ChIP-seq data. We apply several nonparametric methods to identify the genes with differential enrichment. The methods can be applied to the ChIP-seq data of histone modification even without replicates. It is based on nonparametric hypothesis testing in order to capture the spatial differences in protein-enriched profiles. The key of our approaches is to use null genes or input ChIP-seq data to choose

the biologically relevant null values of the tests. We demonstrate the method using ChIP-seq data on a comparative epigenomic profiling of adipogenesis of murine adipose stromal cells. Our method detects many genes with differential H3K27ac levels at gene promoter regions between proliferating preadipocytes and mature adipocytes in murine 3T3-L1 cells. The test statistics also correlate well with the gene expression changes and are predictive of gene expression changes, indicating that the identified differential enrichment regions are indeed biologically meaningful.

We further extend these tests to time-course ChIP-seq experiments by evaluating the maximum and mean of the adjacent pair-wise statistics for detecting differentially enriched genes across several time points. We compare and evaluate different nonparametric tests for differential enrichment analysis and observe that the kernel-smoothing methods perform better in controlling the Type I errors, although the ranking of genes with differentially enriched regions are comparable using different test statistics.

TABLE OF CONTENTS

ACKNOWLEDGEMENT iii

ABSTRACT v

LIST OF TABLES ix

LIST OF ILLUSTRATIONS xi

CHAPTER 1 : Introduction 1

1.1 Copy Number Variants 2

1.2 ChIP-seq Experiments 4

CHAPTER 2 : A Statistical Method for detecting trait-associated Copy Number Variants 7

2.1 Introduction 7

2.2 Statistical Model and CNV Association Test 9

2.3 A Procedure for Identifying the Trait-associated CNVs and Its Theoretical Properties 11

2.4 Simulation Studies 15

2.5 Application to Population Differentiation CNV study 17

2.6 Conclusion and Discussion 23

CHAPTER 3 : Kernel-based Tests for Two-sample Differential Enrichment Analysis Using ChIP-seq data 24

3.1 Introduction 24

3.2 A Motivating Comparative ChIP-seq Study, Data Transformation and Statistical Model . . . 27
3.3 Kernel-smoothing-based Nonparametric Tests . . . 28
3.4 Application to a Comparative ChIP-seq Study During Mouse Adipogenesis . . . 32
3.5 Effects of Bandwidth Selection on Identifying the Genes with Differential Enrichment . . . 44
3.6 Application to an ENCODE ChIP-seq Data with Two Replicates . . . 48
3.7 Extension to Multiple Experimental Conditions and ANOVA-type Test Statistics . . . 50
3.8 Conclusions and Discussion . . . 53

CHAPTER 4 : Two Alternative Nonparametric Tests for Differential ChIP-seq Data Analysis . . . 55
4.1 Introduction . . . 55
4.2 Two-sample Non-parametric Tests . . . 57
4.3 Application to ChIP-seq Study During Mouse Adipogenesis . . . 61
4.4 Extension to Time-Course ChIP-seq Data . . . 73
4.5 Application to a Comparative Time Course ChIP-seq Study During Mouse Adiopogenesis . . . 80
4.6 Conclusions and Discussion . . . 87

CHAPTER 5 : Conclusions and Future Work . . . 88

APPENDICES . . . 92

BIBLIOGRAPHY . . . 101

LIST OF TABLES

TABLE 2.1 : CNVs identified by CNVtest that show different frequencies between Europe and Asian populations 20

TABLE 3.1 : Comparison of model fit R^2 and prediction (PE) 42

TABLE 4.1 : Numbers of genes with DE regions identified by different tests 65

TABLE 4.2 : Numbers of genes with DE regions identified for the ENCODE data sets 72

TABLE 4.3 : Simulation to evaluate the type 1 errors 73

LIST OF ILLUSTRATIONS

FIGURE 2.1 : Simulation results on power comparisons of CNVtest 18
FIGURE 2.2 : Length-standardized sum of the clone intensities 21
FIGURE 2.3 : The clone intensities around the 6 CNVs identified by CNVtest 22

FIGURE 3.1 : Histograms of two test statistics for the null genes 34
FIGURE 3.2 : Observed ChIP-seq bin-counts for top twelve genes ranked by the test statistics $Z_{0\lambda,WH}$ 35
FIGURE 3.3 : Comparison of the proposed statistics and the fold-changes statistics and DBChIP statistics 36
FIGURE 3.4 : Observed mouse adipogenesis ChIP-seq bin-counts over the promoter region 37
FIGURE 3.5 : Plots of gene expression fold changes as a function of two different test statistics 39
FIGURE 3.6 : Plots of proportions of up/down-regulated genes in different intervals of the test statistics 41
FIGURE 3.7 : Model-fitting and prediction for log of the gene expression fold changes 43
FIGURE 3.8 : Histograms of the test statistics $Z_{\lambda_t,WH}$ with the different bandwidths 46
FIGURE 3.9 : ROC curves for identifying differentially expressed genes . . 47
FIGURE 3.10 :Histograms of differential enrichment test statistics Z_{new} for ENCODE data 49
FIGURE 3.11 :F-distributions of null genes for simulation and real data sets 52

FIGURE 4.1 : Histograms of the two test statistics, (a)$Z_{diff,\ eqlvar}$ and (b) $Z_{diff,\ unvar}$ for the null genes 62
FIGURE 4.2 : Comparison of different statistics 64
FIGURE 4.3 : Plots of ROC curves of four test statistics 66
FIGURE 4.4 : Plots of true positive rate curves of four test statistics . . . 68
FIGURE 4.5 : Comparison between two replicated ENCODE input data sets 70
FIGURE 4.6 : Histogram of test statistics $Z_{all,uneql}$ for all 23807 genes in the ENCODE data set. 71
FIGURE 4.7 : Histogram of the test statistics T_{max} for 10,000 samples simulated under the null multivariate normal distribution . . . 78
FIGURE 4.8 : Histograms of the test statistics for the null genes 81
FIGURE 4.9 : Observed ChIP-seq bin-counts for top twelve genes ranked by TS_{max} statistics 83
FIGURE 4.10 :Observed ChIP-seq bin-counts for top twelve genes ranked by TS_{mean} statistics 84
FIGURE 4.11 :Plots of the ROC curves for four different test statistics . . 85
FIGURE 4.12 :Plots of the TPR curves for four different test statistics . . 86

CHAPTER 1

Introduction

Many problems in genomics can be formulated as signal detection problems in statistics. They involve identification of genomic regions that show different characteristics than the background regions. High-throughput technologies have been widely used to generate data for detecting these important local genomic signals. This dissertation focuses on statistical methods for analysis of multiple-sample genomic data, including development of a statistical procedure to identify the copy number variants (CNVs) that are associated with phenotypes and nonparametric tests for differential enrichment based on ChIP-seq data. Different from available methods that often only consider one sample, the focus of our research is on multiple sample analysis in order to detect differential signals, which include the CNVs that are associated with outcomes and the genes that show differential enrichment of histone modifications between two or more conditions.

Most available methods involve a two-step procedure to identify these genomic regions of interest, where the local genomic signal such as CNVs or histone modification regions are first identified for each of the samples. The frequencies of these local signals are then compared and associated with trait values or experimental conditions. Such approaches have two limitations: (1) the local genomic regions identified for different samples may not have exactly the same boundaries, which makes the cross-sample analysis difficult; (2) the local regions identified often strongly depend on certain threshold values on the statistics such as p-value. Different thresholds can lead to very different sets of signals, which also complicate the second stage analysis. We aim to develop multi-sample approaches to both problems.

1.1. Copy Number Variants

Structural variants in the human genome (Sebat et al., 2004; Feuk et al., 2006), including copy number variants (CNVs) and balanced rearrangements such as inversions and translocations, play an important role in the genetics of complex diseases. CNVs are alternations of DNA of a genome that results in the cell having less or more than two copies of segments of the DNA. CNVs correspond to relatively large regions of the genome, ranging from about one kilobase to several megabases, that are deleted or duplicated. CNVs represent an important type of genetic variants observed in human genomes. Recent studies have shown that CNVs are associated with developmental and neuropsychiatric disorders (Feuk et al., 2006; Walsh et al., 2008; Stefansson et al., 2008; Stone et al., 2008) and cancer (Diskin et al., 2009). These findings have led to the identification of novel disease-causing mutations other than single nucleotide polymorphisms, thus contributing important new insights into the genetics of these complex diseases. Changes in DNA copy number have also been highly implicated in tumor genomes. The copy number changes in tumor genomes are often referred to as copy number aberrations (CNAs). Compared to germline CNVs, these CNAs are often longer, sometime involve the whole chromosome arms. In this dissertation, we focus on the CNVs from the germline constitutional genome where most of the CNVs are sparse and short (Zhang et al., 2009; Cai et al., 2012).

CNVs can be discovered by cytogenetic techniques, array comparative genomic hybridization (Urban et al., 2006) and by single nucleotide polymorphism (SNP) arrays (Redon et al., 2006). The emerging technologies of DNA sequencing have further enabled the identification of CNVs by next-generation sequencing (NGS) in high resolution (Cai et al., 2012). NGS can generate millions of short sequence reads along the whole human genome. When these short reads are mapped to the reference genome,

both distances of paired-end data and read-depth (RD) data can reveal the possible structure variations of the target genome (for reviews, see Medvedev et al. (2009) and Alkan et al. (2011)). Novel statistical methods for CNVs analysis based on the NGS data have been developed (Cai et al., 2012). We focus on CNV analysis based on clone-based arrays or the SNP arrays, where the data can be approximately modeled by sequences of ordered Gaussian random variables.

In Chapter 2, we consider the problem of identifying the CNVs that are associated with the trait value such as disease status or quantitative traits. CNVs represent one important type of genetic variants that are associated with many complex diseases. Statistical methods have been developed for identifying the CNVs both at the individual and at the population levels (Wang et al., 2007; Jeng et al., 2010; Zhang et al., 2008a). However, methods for testing the CNV association are limited. Most available methods employ a two-step approach, where the CNVs carried by the samples are identified first and then tested for association (Diskin et al., 2009). Because the identified CNVs vary from sample to sample in their exact boundaries, one has to first determine the shared CNV regions and then prepare a candidate CNV pool for the second step testing. The results of such tests depend on the threshold used for CNV identification and also the choice of the number of CNVs to be tested.

We develop a method, CNVtest, to directly identify the trait-associated CNVs without the need of identifying sample-specific CNVs. The procedure scans the genome with intervals of variable lengths and identifies the trait associated intervals based on examining the score statistics. The procedure is computationally faster than the two-step approaches and does not require the specification of the CNVs to be tested. We show that CNVtest asymptotically controls the Type I error and identifies the true trait-associated CNVs with a high probability. We demonstrate the methods

using simulations and an application to identify the CNVs that are associated with population differentiation between Europeans and Asians (Redon et al., 2006).

1.2. ChIP-seq Experiments

ChIP-sequencing, also known as ChIP-seq, is a method used to analyze protein interactions with DNA. ChIP-seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins. The technologies have been widely applied in biomedical research to identify the binding sites of important transcription factors (TFs) and genomic landscape of histone modifications in living cells (Landt et al., 2012). In ChIP assays, a transcription factor, cofactor, or other chromatin protein of interest is enriched by immunoprecipitation from cross-linked cells, along with its associated DNA. Genomic DNA sites enriched in this manner were initially identified by array-based data and more recently by DNA sequencing (ChIP-seq) (Barski et al., 2007; Johnson et al., 2007; Robertson et al., 2007). Often, it is also important in a ChIP-seq experiment to run a control using "input DNA", i.e. non-ChIP genomic DNA in the same cell types being studied, so that sequencing biases can be identified and adjusted for (Landt et al., 2012).

Previous research has largely focused on developing peak-calling procedures to detect the binding sites for TFs (Zhang et al., 2008b; Kuan et al., 2011; Ji et al., 2008; Schwartzman et al., 2013; Spyrou et al., 2009). However, these procedures may fail when applied to ChIP-seq data of histone modifications, which have diffuse signals and multiple local peaks (O'Geen et al., 2011). Histone marks are sometimes diffusely enriched over several nucleosomes of hundreds of base pairs or in some cases thousands or tens of thousands of base pairs. This often leads to peaks being over-

called in a histone-modification-enriched region, where several peaks might be called but a human would prefer to view the whole region as an enriched unit. The peak calling algorithm can also fail to detect an enriched region where there is a subtle but consistent enrichment but where no single locus is enriched enough to count as a "peak" according to the algorithm's criteria. There may also be apparent gaps in regions that are actually enriched, as a result of insufficiently deep sequencing (Liu et al., 2010).

Besides peaking finding, it is often very important to identify genomic regions or genes with differential enrichment of histone modifications between two or more experimental conditions or cell types (Mikkelsen et al., 2010). In Chapters 3 and 4, we formulate the differential enrichment problem as a hypothesis testing problem and investigate several nonparametric tests for identifying genes with differentially enriched regions based on ChIP-seq data. Parametric methods based on Poisson/Negative Binomial distribution have been proposed to address this differential enrichment problem and most of these methods require biological replications (Mikkelsen et al., 2010; Liang and Keleş, 2012). However, many ChIP-seq data usually have a few or even no replicates.

In Chapter 3, we apply a kernel smoothing-based nonparametric test to identify the genes with differentially enriched regions that can be applied to the ChIP-seq data even without any replicates. Our method is based on nonparametric hypothesis testing and kernel smoothing in order to capture the spatial differences in histone-enriched profiles. Using a large bandwidth, our method can smooth out potential systematic biases that have been described in next-generation sequencing in general and ChIP-seq in particular. Such biases can be due to a preference for sequencing GC rich regions and mapping bias from the frequency of occurrence of particular short ho-

mologous sequences in the genome and from genomic amplifications and repeats. We demonstrate the method using a ChIP-seq data on comparative epigenomic profiling of adipogenesis of adipose stromal cells. Our method detects many genes with differential H3K27ac levels at gene promoter regions between proliferating preadipocytes and mature adipocytes. The test statistics also correlate well with the gene expression changes and are predictive of gene expression changes, indicating that the identified differential enrichment regions are indeed biologically meaningful. Extension to ChIP-seq data from multiple experimental conditions is also presented.

In Chapter 4, we apply two other nonparametric tests that do not require smoothing the data first. In the literature, there are few methods available to detect genes with differentially enriched regions among more than two conditions, such as multiple time-course ChIP-seq data. We investigate the time-course histone modification enrichment changes of the genes across four time points. Multivariate test statistics are derived as the mean (TSmean) or maximum (TSmax) of three adjacent pair-wise test statistics. Methods for variance estimation under homoscedasticity and heteroscedasticity in error variances are discussed. Comparing the performance of different test statistics is conducted via ROC curves and True Positive Rate (TPR) curves in both two-sample and multi-sample cases. Both real data and simulation results shows the TSmax with kernel smoothing tends to outperform other methods.

Finally, in Chapter 5, we present conclusions and outline possible future research.

CHAPTER 2

A Statistical Method for detecting trait-associated Copy Number Variants

2.1. Introduction

Structural variants in the human genome (Sebat et al., 2004; Feuk et al., 2006), including copy number variants (CNVs) and balanced rearrangements such as inversions and translocations, play an important role in the genetics of complex disease. CNVs, ranging from about one kilobase to several megabases, are alternations of DNA of a genome that result in the cell having less or more than two copies of segments of the DNA. CNVs represent an important type of genetic variants observed in human genomes. Recent studies have shown that CNVs are associated with developmental and neuropsychiatric disorders (Feuk et al., 2006; Walsh et al., 2008; Stefansson et al., 2008; Stone et al., 2008) and cancer (Diskin et al., 2009). Identification of these novel disease-causing CNV mutations has contributed important new insights into the genetics of these complex diseases. Thus, identifying the CNVs that are associated with complex traits is an important problem in human genetic research.

Many novel and powerful statistical methods have been developed recently for identifying the CNVs in a given sample based on array data, SNP chip intensity data, and next generation sequencing data. Important examples include the optimal likelihood ratio selection method (Jeng et al., 2010), the hidden Markov model-based method (Wang et al., 2007), and change-point based methods (Olshen et al., 2004). To identify the recurrent copy number variants that appears in multiple samples, Zhang et al. (2008a) introduced a method for detecting simultaneous change-points

in multiple sequences that is only effective for detecting the common variants. Siegmund et al. (2010) extended their method by introducing a prior variant frequency that needs to be specified. Jeng et al. (2013) proposed a proportion adaptive sparse segment identification procedure that is adaptive to the unknown CNV frequencies.

Despite these novel methods for CNV detection and identification, methods for testing the CNV association are very limited. Current methods for CNV testing fall into two categories. One is to assume that a set of CNVs are known and to test association of these CNVs with complex phenotypes. Barnes et al. (2008) developed an approach for testing CNV association using a latent variable framework. However, the current databases of all CNVs are still very incomplete and testing only the known CNVs can miss the new CNVs that are associated with the phenotype of interest. Another common approach for CNV testing is a two-step approach, where CNVs are first identified for each sample and the CNVs that appear in multiple samples are then tested using chi-square or Fisher's exact test (Diskin et al., 2009). One limitation of such approaches is that the uncertainty associated with the inferred CNVs is not accounted for in the testing and the CNVs identified depend on the threshold used. In addition, since the CNVs identified may not have exactly the same boundaries, one has to decide which CNV regions to test. Finally, it is not clear how one should control for the genome-wide error rate since the number of CNVs to be tested is not known before performing the single sample CNV analysis.

In this section, we propose a new statistical method for identifying trait-associated CNVs. Instead of assuming a known set of CNVs or first identifying the CNVs carried by the samples, the proposed method directly identifies the CNVs that are associated with the trait of interest. The procedure scans the genome with intervals of variable lengths and identifies the trait associated intervals based on examining the

score statistics. The procedure is computationally faster than the two-step approaches and does not require the specification of the CNVs to be tested. We show that the procedure can control the genome-wide error rate and also has a high probability of identifying the trait-associated CNVs.

Chapter 2 is organized as follows. We present the statistical model representing the relationship between CNVs and a phenotype in Section 2.2. In Section 2.3, we present a scanning procedure for identifying trait-associated CNVs and give the theoretical properties. The performance of our method is evaluated using simulations in Section 2.4. In Section 2.5, we demonstrate our method in identifying the CNVs that are associated with population differentiation. Finally, a brief discussion is given in Section 2.6.

2.2. Statistical Model and CNV Association Test

Suppose that we have data on n independent individuals. Let Y_i be the phenotype value for the ith individual, X_{ij} be the observed marker intensity (e.g., the log R Ratio from the SNP chip data) for the ith individual and jth marker, $i = 1, \cdots, n$ and $j = 1, \cdots, m$, where $m = m_n$ possibly increases with n. Here Y_i can be a binary variable as in case-control studies or continuous variable, e.g., in eQTL studies, Y_i can be the expression level of a gene. For the SNP chip data, the observed marker intensity data is log R-Ratio, $X_{ij} = \log_2(R_{obs}/R_{ref})$, where R_{obs} represents the total intensity of two alleles at the jth SNP for the ith sample and R_{ref} the corresponding quantity for a reference sample. When there is no copy number change in a genomic region for individual i, we expect that the X_{ij}'s in that region are realizations of a baseline distribution. In the following, for each sample, we normalize the intensity data to have variance of 1 by dividing by the median absolute deviation. Suppose

there is a total of $q = q_{m,n}$ CNVs in all n individuals with q possibly increasing with m and n and is unknown. Let $\mathbb{I} = \{I_1, \ldots, I_q\}$ be the collection of the corresponding CNV segments/intervals. The value X_{ij} in a CNV segment deviates from 0 to the negative or positive side depending on whether the segment is deleted or duplicated.

Since only a certain proportion of the samples carry a given CNV, we denote the carriers' proportion for CNV at I_k as π_k, $1 \leq k \leq q$. We assume

$$X_{ij} \sim \begin{cases} (1-\pi_k)N(0,1) + \pi_k N(\mu_k, \sigma_k^2), & j \in I_k \text{ for some } I_k \in \mathbb{I} \\ N(0,1), & \text{otherwise,} \end{cases} \tag{2.1}$$

where $\mu_k \neq 0$ represents the mean value of the jump sizes in the k-th CNV segment and σ_k may or may not equal 1, which reflects the fact that different variation may be introduced by the CNV carriers. Here π_k, μ_k and σ_k are unknown for each $I_k \in \mathbb{I}$.

For a given candidate interval τ and individual i, we summarize the marker intensity data in this interval by the length-standardized sum

$$\bar{X}_{i\tau} = (\sum_{j \in \tau} X_{ij})/\sqrt{|\tau|}. \tag{2.2}$$

Further, define

$$Z_{i\tau} = 1(|\bar{X}_{i\tau}| > \nu) \tag{2.3}$$

for some $\nu > 0$ to indicate whether or not the ith individual carries some copy number changes in interval τ. The threshold ν will be specified in the next section. To link carrier status at interval τ to the phenotype, we assume the following generalized linear model (GLM) for the phenotype Y_i with the likelihood function

$$\exp\{Y_i\psi - b(\psi)/\gamma + c(Y_i, \gamma)\}, \tag{2.4}$$

where $\psi = g(\alpha + \beta_\tau Z_{i\tau})$ is the link function for $Z_{i\tau}$ and Y_i and γ is the dispersion parameter. In this model, α is the intercept and β_τ is the regression coefficients that associates the possible CNV at τ to the mean value of the phenotype. Our goal is to identify the elements in $\mathbb{I}$ that have non-zero β coefficient. The identified elements indicate the locations of the trait-associated CNVs.

2.3. A Procedure for Identifying the Trait-associated CNVs and Its Theoretical Properties

In this section, we present a scanning procedure for identifying the trait-associated CNVs followed by the theoretical analysis of its Type I error controls and power.

2.3.1. A scanning procedure for identifying the trait-associated CNVs

Since most CNVs are short, we only consider short intervals with length $\leq L$ in the sequences of the observed genome-wide data. The L is chosen to satisfy the following condition:

$$\bar{s} \leq L < \underline{d}, \qquad \text{and} \qquad \log L = o(\log m), \tag{2.5}$$

where $\bar{s} = \max_{1\leq k\leq q} |I_k|$ and $\underline{d} = \min_{1\leq k\leq q-1}\{\text{distance between } I_k \text{ and } I_{k+1}\}$. This condition guarantees that all the CNV segments can be covered by some intervals considered in the algorithm and, at the same time, none of the intervals is long enough to reach more than one CNV segment. In the applications we consider, most CNVs are very short and sparse, so condition (2.5) is easy to be satisfied. We usually choose $L = 20$ for SNP chip data, because most of the CNVs are shorter than 20 SNPs. Let $\mathcal{I}$ be the collection of all mL intervals of length $\leq L$. The threshold in (2.3) is set at

$$\nu = \sqrt{2\log(mL)}. \tag{2.6}$$

This is the same threshold used in Jeng et al. (2010) for detecting CNVs in a long sequence of m genome-wide observations for one individual. A threshold at this level optimally controls false positive CNV identification for each individual asymptotically and greatly reduces the number of intervals that need to be considered for association tests.

We first select the intervals in $\mathcal{I}$ that have $Z_{i\tau} = 1$ for at least one individual and denote the collection of such intervals as

$$\mathcal{R} = \{\tau \in \mathcal{I} : 0 < \sum_{i=1}^{n} Z_{i\tau} < n\}. \tag{2.7}$$

Let $\hat{r} = |\mathcal{R}|$ be the total number of such intervals. Note that the collection $\mathcal{R}$ is much smaller than $\mathcal{I}$ and only includes intervals where copy number changes are observed in the samples. However, $\mathcal{R}$ is not simply the collection of identified sample-specific CNVs as it includes all the intervals that may overlap with the true CNVs. Since the CNV boundaries may vary from individual to individual, including the whole collection $\mathcal{R}$ into the testing step below avoids identifying the sample-specific CNVs and the shared CNV regions across the samples.

As a next step, based on the GLM model (2.4), we test

$$H_{\tau 0} : \beta_\tau = 0 \qquad \text{v.s.} \qquad H_{\tau 1} : \beta_\tau \neq 0$$

for any $\tau \in \mathcal{R}$ using the score statistic

$$S_{n,\tau} = n^{-1/2} \sum_{i=1}^{n} Z_{i\tau}(Y_i - \bar{Y})/S_{Z_\tau} S_Y, \tag{2.8}$$

where $S_{\bar{Z}_\tau}$ and S_Y are the sample standard deviations of Z_τ and Y. The score

statistic $S_{n,\tau}$ has an asymptotic standard normal distribution under $H_{\tau 0}$ for $\tau \in \mathcal{R}$. Therefore, we reject $H_{\tau 0}$ if $|S_{n,\tau}| > \lambda$, where λ is a threshold determined by the limiting distribution of $S_{n,\tau}$ under $H_{\tau 0}$ and the number of score tests performed. We set

$$\lambda = \sqrt{2 \log(\hat{r})} \tag{2.9}$$

in order to control the genome-wide errors.

Our scanning procedure, called CNVtest, identifies the elements in $\mathbb{I}$ that are significantly associated with the trait value Y by selecting the intervals in $\mathcal{R}$ with their absolute score statistics above λ and achieving local maximums. Specifically, CNVtest involves the following steps:

1. Pick an L. Select $\mathcal{R}$ as in (2.7).
2. Calculate $S_{n,\tau}$ as in (2.8) for all $\tau \in \mathcal{R}$.
3. Let $\mathbb{I}^{(1)} = \{\tau \in \mathcal{R} : |S_{n,\tau}| > \lambda\}$, where λ is defined in (2.9). Let $l = 1$.
4. Let $\hat{I}_l = \arg\max_{\tau \in \mathbb{I}^{(l)}} |S_{n,\tau}|$, and update $\mathbb{I}^{(l+1)} = \mathbb{I}^{(l)} \backslash \{\tau \in \mathbb{I}^{(l)} : \tau \cap \hat{I}_l \neq \emptyset\}$.
5. Repeat Step 4-5 with $l = l + 1$ until $\mathbb{I}^{(l)}$ is empty.

Finally, we denote the trait-associated CNVs by $\hat{\mathbb{I}} = \{\hat{I}_1, \hat{I}_2, \ldots\}$. If this set is empty, then we conclude that there is no trait-associated CNV.

2.3.2. Theoretical results on error control and power analysis

Recall that $q = q_{m,n}$ is the total number of true CNVs in n individuals. We assume

$$\log q = o(\log m) \qquad \text{and} \qquad q \to \infty \quad \text{as } n \to \infty, \tag{2.10}$$

which means that the CNVs are sparse and their number increases with the number of individuals. Further, for each CNV, we assume

$$\mu_k \sqrt{|I_k|} \geq \sqrt{2(1+\epsilon)\log m}, \qquad 1 \leq k \leq q. \tag{2.11}$$

for some $\epsilon > 0$. Condition (2.11) is a necessary condition for CNVs to be detectable in a sequence of m genome-wide observations (Jeng et al., 2010).

The following theorem states that with a large probability, CNVtest controls the genome-wide error rate. In other words, the CNVtest does not select the null intervals in $\mathcal{I}$.

Theorem 2.3.1 *Assume (2.1), (2.4), (2.10), (2.11), and (2.5). Let* $\mathcal{I}_0 = \{\tau \in \mathcal{I} : \tau \cap I_k = \emptyset$ *for any* $I_k \in \mathbb{I}\}$ *be the set of intervals that do not overlap with any of the CNVs in the true CNV set* $\mathbb{I}$*. Then*

$$P(\exists \tau \in \mathcal{I}_0 : \tau \in \hat{\mathbb{I}}) \to 0 \qquad \text{as } n \to \infty.$$

This theorem implies that the probability of CNVtest identifying wrong trait-associated CNVs goes to zero when the sample size is large enough.

We next study the power of CNVtest in identifying the trait-associated CNVs. For a given interval τ, define

$$D(\tau) = g'(\alpha)\sqrt{\mathrm{Var}(Z_\tau) b''\{g(\alpha)\}/\gamma}, \tag{2.12}$$

where $g(\cdot)$, $b(\cdot)$, α, and γ are defined in the GLM model (2.4). Note that $\mathrm{Var}(Z_\tau)$ depends on the length of the interval $|\tau|$ and the corresponding CNV mean value μ_τ.

Theorem 2.3.2 *Assume the same conditions as in Theorem 2.3.1. Suppose there exists an element* $I_k \in \mathbb{I}$ *such that*

$$\beta_{I_k} \geq \frac{\sqrt{2(1+\eta)\log m}}{D(I_k)\sqrt{n}} \tag{2.13}$$

for some $\eta > 0$. *Then,* $H_{I_k 0}$ *is rejected by the CNVtest with probability going to* 1 *as* $n \to \infty$. *Further, suppose* $\pi_k < 1/2$ *and* $\beta_{I_k} > \beta_\tau$ *for any* τ *such that* $\tau \cap I_k \neq \emptyset$ *and* $\tau \neq I_k$. *Then,* $P(S_{n,I_k} > S_{n,\tau}) \to 1$ *as* $n \to \infty$.

Theorem 2.3.2 shows that when β_{I_k} is large enough, I_k is selected to enter the candidate set $\mathbb{I}^{(1)}$ in the algorithm with a high probability. The additional conditions in the second part of the theorem imply the monotonicity of the mean value of the score statistics $S_{n,\tau}$ with respect to how much τ overlaps with I_k, so that the score statistic of the true segment I_k dominates the score statistics of other intervals overlapping with I_k and the true segment I_k is selected by the algorithm.

2.4. Simulation Studies

In this section, Monte Carlo simulations are presented to evaluate the performance of CNVtest. We simulate data sets with $n = 1,000$ individuals, of whom 500 are cases and 500 are controls. For each individual, the log-R intensity values are generated at $m = 5,000$ markers. We simulate three CNVs with their lengths set at $s = 10$. One of them is a null CNV with the same frequency of 0.15 in both case and control groups. Another is a disease-associated CNV with a frequency of 0.10 in the control group and a frequency of $p = 0.15, 0.20, 0.25$, and 0.30 in the case group. We also consider the case when the locations of a CNV are not exactly the same across individuals and simulate the third CNV as a disease-associated CNV with locations varying randomly within an interval of length 15. Therefore, the carriers for the third CNV

have overlapping but not exactly the same CNV segments. We set the shifted mean at $\mu = 1.5, 1.75, 2, 2.25$ and 2.5. Each observation $X_{ij}, i = 1, ..., n, j = 1, ..., m$ is generated from $N(A_{ij}, 1)$. If marker j is located in a CNV segment and the i^{th} individual is a carrier of the variant, $A_{ij} = \mu$; otherwise, $A_{ij} = 0$. The phenotype $Y_i, i = 1, ..., n$ takes value of 1 and 0 for case and control individual, respectively.

We apply CNVtest with $L = 15$ and $\nu = \sqrt{2\log(mL)} = 4.74$ to select the disease-associated CNVs. The simulations are repeated 50 times. To evaluate the performance of CNVtest, we show three summary statistics: the score statistic as in (2.8), the empirical power, which equals the proportion of times that a disease associated segment is selected in the 50 replications, and the empirical over-selection, which equals the proportion of times that an interval not overlapping with the disease-associated CNV is selected. The estimated standard errors of the means of these statistics are derived from calculating the standard deviation of 500 bootstrap means of the 50 results from 50 replications.

We first examine the effects of CNV jump size μ on the CNVtest performances where the CNV carrier frequency is fixed at 20% in cases and 10% in controls for the disease-associated CNVs, and at 15% in both cases and controls for the null CNV. Figure 2.1 (a) shows the score statistics calculated for the null CNV and also the disease-associated CNVs with the jump size changing from 1.5 to 2.5, together with the threshold level determined by (2.9). We observe that the score statistics for the null CNV is constant and is always much smaller than the threshold. On the other hand, the score statistics for the disease associated CNVs increases as μ increases. In addition, shifts in exact CNV boundaries lead to smaller score statistics, especially when μ is small. Figure 2.1 (b) shows the empirical power of CNVtest for identifying the disease-associated CNVs. As expected, larger μ leads to a higher power of identifying

the true CNVs. Again, shifts in exact CNV boundaries lead to a slight loss of power, especially when μ is small. We observed that the empirical over-selections are always zero for all data sets simulated, and they are not affected by the values of μ.

We then fix $\mu = 2.0$ and examine how the carrier proportion in cases affects the power of identifying the disease-associated CNVs. Figure 2.1 (c) shows the score statistics evaluated for the null CNV and the disease-associated CNVs with carrier proportion in cases changing from 15% to 30%, together with the threshold level determined by (2.9). We observe that the score statistics for the null CNV are constant and always much smaller than the threshold. On the other hand, the score statistics for the disease associated CNVs increase as the carrier proportion in the cases increases. Again, for all simulations, we did not observe any false identification.

2.5. Application to Population Differentiation CNV study

Redon et al. (2006) presented the first genome-wide global variation analysis of DNA copy number in the human genome where DNA EBV-transformed lymphoblastoid cell lines of the 270 HapMap samples was screened for CNVs using clone-based comparative genomic hybridization (Whole Genome TilePath, WGTP) array consisting of 26,463 large-insert clones. To demonstrate our method, we consider data from two populations: 89 of European descent from Utah (CEU), 45 unrelated Japanese from Tokyo (JPT) and 45 unrelated Han Chinese from Beijing (CHB). Our goal is to identify the genomic regions that show difference in copy number between CEU and Asian populations (JPT+CHB). Such population differentiation in CNV can provide important insights into genetic diversity and evolution.

For each individual, we first standardize the clone intensity data by mean and variance calculated for this individual. Since one clone covers a longer region than the SNP

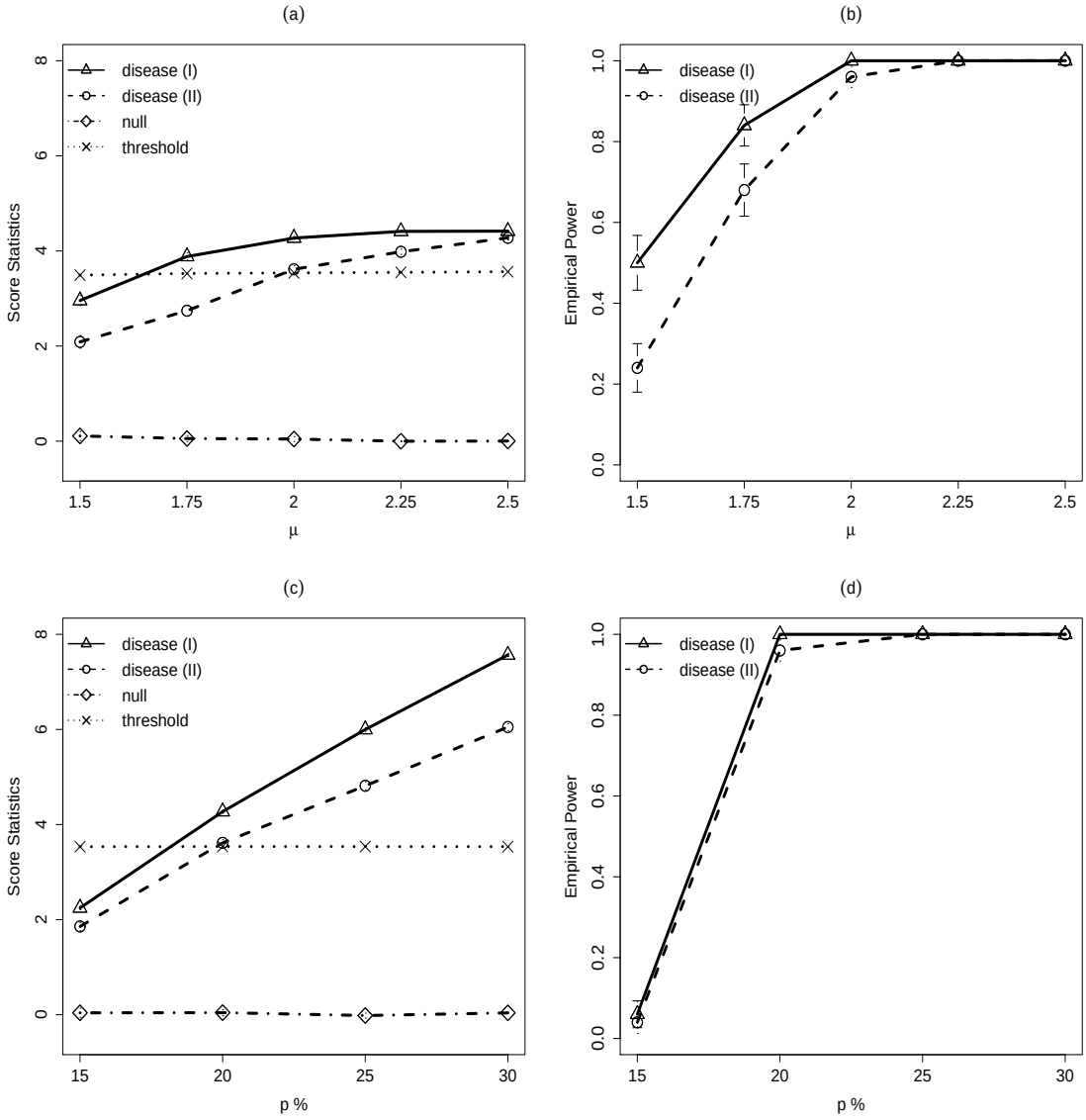


Figure 2.1: Simulation results. (a)-(b): Effect of the CNV jump size μ from 1.5 to 2.25 on (a) score statistics for CNVs with carrier probability of 20% in case and 10% in control and (b) power of detecting the associated CNV. (c)-(d): Effect of the CNV frequency in case p from 15% to 30% on (c) score statistics for CNVs with carrier probability of 20% in case and 10% in control and (d) power of detecting the associated CNV.

data, we choose $L = 10$ in our CNVtest so that the largest CNV covers at most 10 clones. Here we consider both duplication and deletion copy number variants and modify (2.3) by $Z_{i\tau}^{dup} = 1(\bar{X}_{i\tau} > \nu)$ for duplication and $Z_{i\tau}^{del} = 1(\bar{X}_{i\tau} < -\nu)$ for

deletion, where $\nu = \sqrt{2\log(mL)} \approx 4.997$. The resulting $\hat{r}^{dup}(= |\mathcal{R}^{dup}|) = 26,496$ and $\hat{r}^{del}(= |\mathcal{R}^{del}|) = 13,585$. Note that both $\hat{r}^{dup}$ and $\hat{r}^{del}$ are much smaller than the number of possible intervals in the whole genome, which is at the order of m^2. Consequently, the threshold $\lambda^{dup} = \sqrt{2\log(\hat{r}^{dup})} \approx 4.513$ and $\lambda^{del} = \sqrt{2\log(\hat{r}^{del})} \approx 4.363$, respectively.

CNVtest identified five duplication CNVs and one deletion CNV that showed different frequencies between the European and Asian populations. Table 2.1 shows their clone locations, size, overlapping genes and their score statistics defined in (2.8). Figure 2.2 shows the scatter-plots of the length-adjusted sum of clone intensity statistics defined in (2.2) for each of the samples for each of the six identified CNV regions, clearly indicating the differences of the carrier frequencies. To show that the clones in the identified CNV regions indeed have different intensities for samples in these two different populations, we present in Figure 2.3 the observed clone intensities for the clones within and outside the identified CNV regions respectively for each of the samples. Again, the identified CNV regions indeed show some differences in clone intensities from their neighboring clones. Note that the two CNVs on chromosome 9 are very close to each other and have similar intensity patterns in the samples. It is likely that they form a large CNV. This is due to the fact that we chose $L = 10$ in CNVtest. However, as in any CNV analysis, a post-processing step may simply combine these two CNVs into one.

Redon et al. (2006) reported two CNVs that exhibit the highest population differentiation between CEU and JPT+CHB, one of which, the duplication CNV on chromosomes 17 that includes gene MAPT, is also identified by CNVtest. CNVtest did not identify the CNV on chromosome 3 reported by Redon et al. (2006). However, this CNV only includes one clone and does not have any known genes in it. The

deletion CNV identified by CNVtest, which includes gene DCTN4, was presented to have the highest population differentiation between CEU and Yoruban samples. The intensity plot in Figure 2.3 for this region shows clear a difference between the CEU and JPT+CHB samples.

Besides samples from CEU and JPT+HCB, Redon et al. (2006) also obtained the clone data for 90 Yoruban (YRI) samples. When comparing CEU and YRI, CNVtest identified 4 deletion CNVs and 11 duplication CNVs that showed very different frequencies. These CNVs include all 6 CNVs that were reported in Redon et al. (2006) to have the highest population differentiation. When comparing YRI and JPT+HCB, CNVtest identified 2 deletion CNVs and 12 duplication CNVs, including 2 CNVs that were reported in Redon et al. (2006) to have the highest population differentiation.

2.6. Conclusion and Discussion

We have developed a new statistical method, CNVtest, for genome-wide CNV association studies. Compared with the commonly used two-step approaches, CNVtest is computationally much faster because the genome is only scanned once. The computational complexity of this method is the same as the likelihood ratio selector of Jeng et al. (2010) and the multiple sample CNV analysis procedure of Jeng et al. (2013), all in the order of $O(mL)$. In addition, it avoids the often troublesome task of determining which CNV regions one should test for association and how to adjust for multiple comparisons. The method is particularly effective when the CNV regions from the different carriers do not exactly cover the same intervals. The CNVtest is also flexible and can be applied to identify CNVs associated with different phenotypes through the use of the generalized linear models.

CNVtest can also be applied to CNV association study using the read depth data from

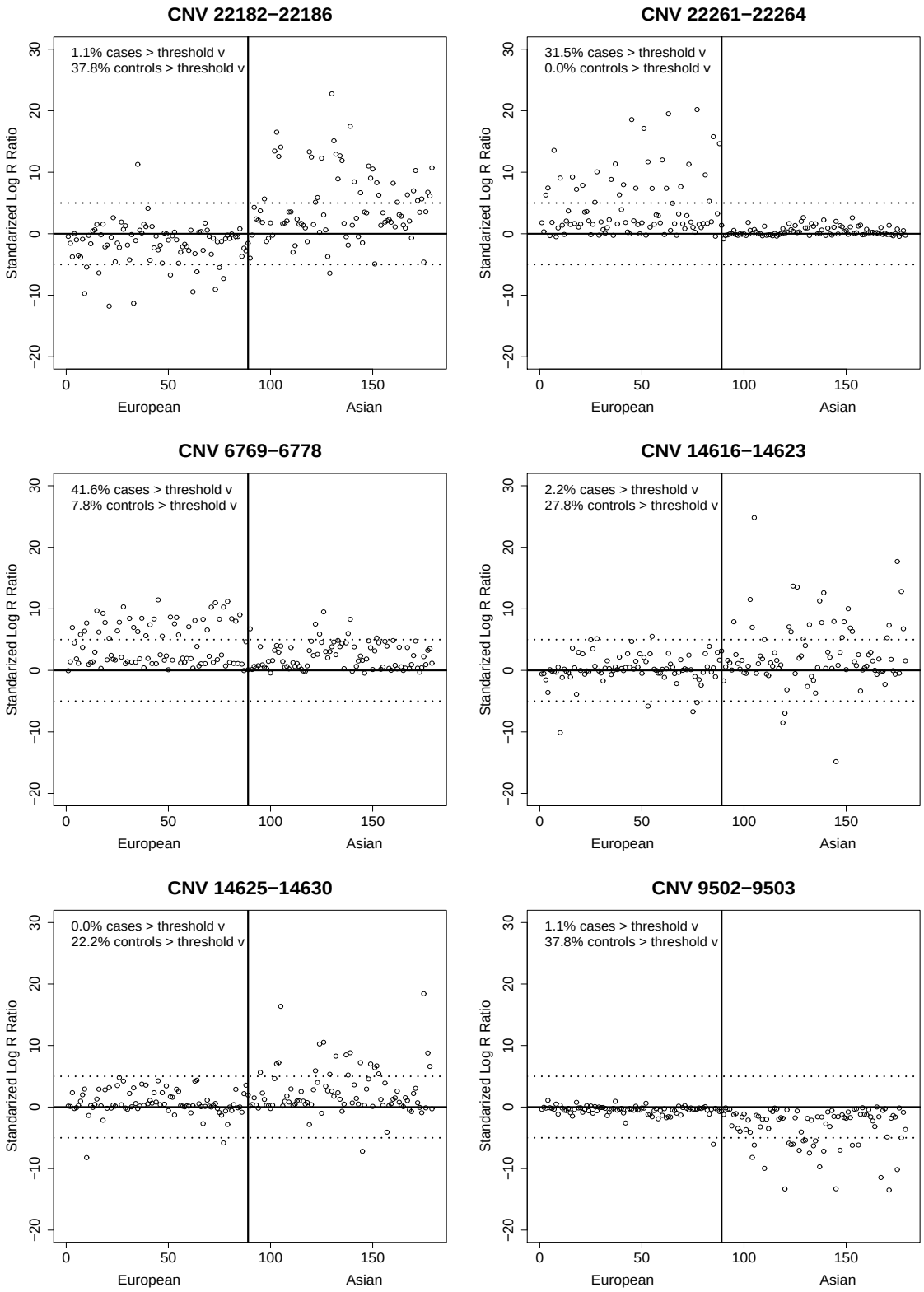


Figure 2.2: Length-standardized sum of the clone intensities of the European (CEU) and Asian samples (JPT+HCB) for the 6 CNVs identified by CNVtest. The estimated CNV carrier proportions are also shown. 21

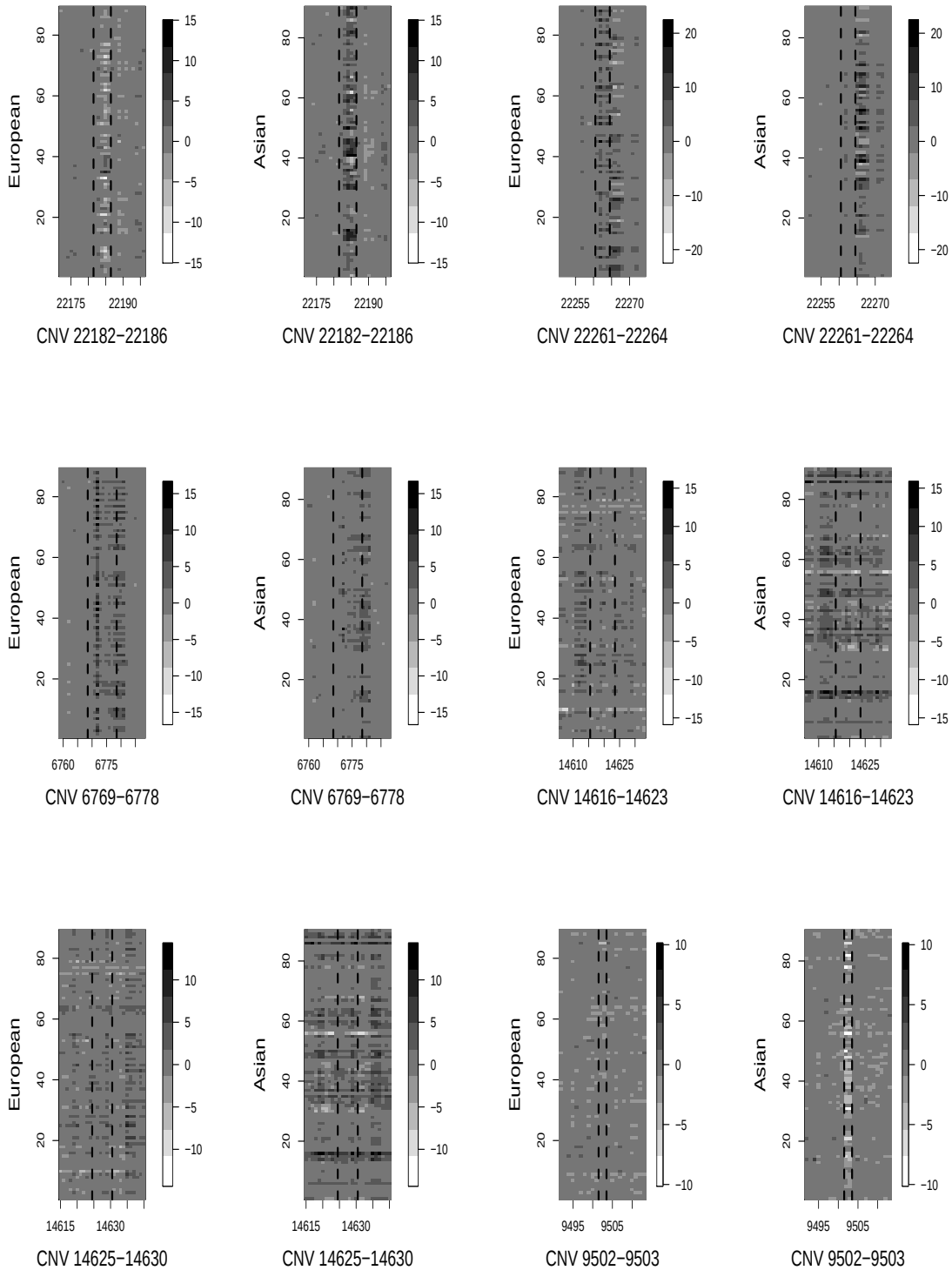


Figure 2.3: The clone intensities around the 6 CNVs identified by CNVtest (marked by dashed vertical lines) for each of the European (CEU) and Asian samples (JPT+HCB).

Table 2.1: The CNVs identified by CNVtest that show different frequencies between Europe and Asian populations. Clone locations, chromosome, CNV size, overlapping genes (based on NCBI36, March 2006, Build 19) and the corresponding score statistics (Score) are shown.

Clone	Start - End	Chrom	Size	Genes	Score
		Duplication CNV			
22182-22186	31,239,836 - 31,981,395	17	741 Kb	RDM1,CCL1/L2/L3/L4, TBC1D3G/C,PRC17, AK125932,LYZL6, ZNHIT3,MY019,etc	-6.15
22261-22264	41,439,751 - 41,722,491	17	282 Kb	MAPT,KANSL1, LOC284058	5.76
6769-6778	68,858,466 - 70,272,807	4	1414 Kb	UGT2B, YTHDC1, TMPRSS11E	5.22
14616-14623	44,819,176 - 45,798,788	9	979 Kb	LOC100132167, CR615666	-4.75
14625-14630	64,368,148 - 65,433,585	9	1065 Kb	LOC401507, AL953854.2-002	-4.69
		Deletion CNV			
9502-9503	150,080,197 - 150,265,935	5	186 Kb	DCTN4, MST150, ZNF300	-4.77

the next generation sequencing. One can use the local median transformation procedure proposed in Cai et al. (2012) to transform the read-depth data to approximately normally distributed data and directly apply the CNVtest to the transformed data. We expect to have similar power and genome-wide error control as the intensity-based data.

CHAPTER 3

Kernel-based Tests for Two-sample Differential Enrichment Analysis Using ChIP-seq data

3.1. Introduction

Chromatin immunoprecipitation sequencing (ChIP-seq) technology is a powerful tool for analyzing protein interactions with DNA (Park, 2009). ChIP-seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins. It can be used to map global binding sites of transcription factors (TFs) and genomic landscape of histone modification marks (HMs). This high-throughput technology can create millions of short parallel sequencing reads and provide more accurate mapping information for the binding regions in the whole genome with lower cost (Johnson et al., 2007; Mikkelsen et al., 2010; Mortazavi et al., 2008; Barski et al., 2007) than array-based methods. Both TF binding and histone modification play important roles in gene regulation, where TFs bind to DNA at a promoter region to promote or block gene transcription. The signal of TFs usually shows one sharp peak at binding sites. Multiple histone modification marks have been reported to be associated with transcription initialization, open chromatin and repression of transcription (Mikkelsen et al., 2010; Hon et al., 2009).

Most previous work in analysis of ChIP-seq data has focused on developing peak-calling procedures to find the binding sites for TFs (Zhang et al., 2008b; Kuan et al., 2011; Ji et al., 2008; Schwartzman et al., 2013; Spyrou et al., 2009). Identifying the enriched region of histone modification marks is difficult since their signals are more

spread out (O'Geen et al., 2011). The signals of HMs are diffuse and usually have multiple local peaks, which are hard to identify by directly applying peak-calling algorithms.

Another important question is to identify the genomic regions that show differential enrichment of histone modification between two experimental conditions, such as different cellular states or different time points (Mikkelsen et al., 2010; Liang and Keleş, 2012). Indeed, different types of differential enrichment have been observed, including shift of nucleosome positions, peak height differences and presence/absence of HM marks (Chen et al., 2011; He et al., 2010). Chen et al. (2011) further demonstrated that the spatial distributions of histone marks are predictive for promoter locations and promoter usage. Angel et al. (2011) show that during cold, the H3K27me3 levels progressively increase at a tightly localized nucleation region in *Arabidopsis*, indicating the importance of studying the peak height, not just the presence/absence of peaks.

One common approach to identifying differentially enriched regions is to apply a peak-calling algorithm to identify the enriched regions for each of the two conditions. The regions with peaks in one condition but without peaks in the other condition are then selected. However, selection of enriched regions often depends on the thresholds used in the peak-calling algorithm. Small differences in the calculated p-values or the FDR threshold used by the peak-finding program can lead to very different sets of peaks. Furthermore, this simple procedure has limitations in detecting the differential enrichment of different peak heights or different peak locations.

Several parametric methods based on Poisson/negative binomial distribution have been proposed to address this differential enrichment problem in ChIP-seq data such as DiffBind and DBChIP (Stark and Brown, 2011; Liang and Keleş, 2012). Most of

these methods require biological replications to estimate the parameters, especially the dispersion parameter in the negative binomial model (Kuan et al., 2011). However, many ChIP-seq data usually have a few or even no replicates. Taslim et al. (2009) proposed a nonlinear method that uses locally weighted regression (Lowess) for ChIP-seq data normalization. Shao et al. (2012) developed a method to quantitatively compare ChIP-seq data sets. To circumvent the issue of differences in signal-to-noise ratios between samples, they focused on ChIP-enriched regions and introduced the idea that ChIP-seq common peaks could serve as a reference to build the rescaling model for normalization. The inputs of all the methods mentioned rely on first identifying the enriched regions and then obtaining the total tag or read counts in these regions. Such approaches have two limitations. First, one has to identify the regions using peak-finding algorithms. Second, by summarizing the number of tags into one single number of the region, one can potentially lose important spatial profile differences such as shifts of the signal region or shapes of signals.

In this Chapter , we propose a nonparametric method to identify the genes with differentially enriched regions based on the ChIP-seq data. Instead of first identifying the enriched regions or peaks as most of the existing methods do, we consider the regions close to genes that may contain important regulatory elements such as the promoter regions, the gene body and downstream regions of the genes. For each of the regions, we summarize the data as counts of sequencing reads in each of the bins of a given length (e.g., 25 bps). The counts in these candidate regions provide important information about different HM levels between two cellular states. After transforming the count data to approximately normal, we apply kernel smoothing to the differences of the data and develop a nonparametric hypothesis testing based on the kernel smoothing. Applying smoothing to the data helps to eliminate the small

local differences that are unlikely to be biologically relevant.

We demonstrate the method using ChIP-seq data on a comparative epigenomic profiling of adipogenesis of murine 3T3-L1 cells reported in Mikkelsen et al. (2010). Our method detects genes with differential H3K27ac levels at gene promoter regions between proliferating preadipocytes and mature adipocytes, which agree with what were observed in Mikkelsen et al. (2010) based on fold-change analysis. The test statistics also correlate with the gene expression changes well, indicating that the identified differences are indeed biologically meaningful. Our results also indicate that the combination of different histone modification profiles can predict the fold changes of gene expressions very well.

3.2. A Motivating Comparative ChIP-seq Study, Data Transformation and Statistical Model

We consider the ChIP-seq experiments reported in Mikkelsen et al. (2010) on murine 3T3-L1 cells undergoing adipogenesis. Specifically, they generated genome-wide chromatin state maps using ChIP-seq profiling, where they mapped six HMs and two TFs at four time points, including proliferating (day -2) and confluent (day 0) preadipocytes, immature adpipocytes (day 2) and mature adipocytes (day 7). We focus our analysis on H3K27ac mark, which is expected to be enriched at active promoters or enhancers. In order to identify the genes that show differential H3K27ac levels between the preadipocytes (day -2) and mature adipocytes (day 7), we consider the upstream 5000 bp region and downstream 2000 bp regions around transcription start site (TSS) for each gene and divide the regions into 280 bins of 25bps. We map the raw data using Bowtie (Langmead et al., 2009), extend reads to the fragment size and then obtain the genome wide coverage data with a fixed bin size of 25 bp. Since

the two ChIP-seq samples usually are sequenced at different depths (total number of reads). We scale the counts according to the sequencing depth ratio. Suppose that there are m genes and for each gene i, there are n observed read counts X_{ikj} in bin k under condition j, for $i = 1, \cdots, m$, $k = 1, \cdots, n$ and $j = 1, 2$. Our goal is to find the genes with differential H3K27ac levels at their promotor regions between mature adipocytes and preadipocytes.

For each gene i and each condition j, we assume the data $X_{ikj}, k = 1, \cdots, n$ are approximately Poisson with means μ_{ikj}. We first apply variance-stabilizing transformation (VST) procedure to transform the variables to the variables $X^*_{ikj} = 2\sqrt{X_{ikj} + 0.25}$, as recommended by Brown et al. (2010, 2005). Thus, we can treat X^*_{ikj}'s as approximate normal variables with mean $2\sqrt{\lambda_{ikj}}$ and variance of 1. For the ith gene, in order to test for differential enrichment between two conditions, we calculate the difference between the two conditions as $Y_{ik} = X^*_{ik1} - X^*_{ik2}$. If there is no differential enrichment, $Y_i^{\mathrm{T}} = (Y_{ik}, ..., Y_{in})$ should have a mean value of zero.

We further denote $Y_i(t_k) = Y_{ik}$, for $t_k = k/n \in (0, 1]$, $k = 1, ..., n$. We assume the following "signal+white noise" model for the normalized differences,

$$Y_i(t_k) = f_i(t_k) + \sigma_i W_i(t_k), \tag{3.1}$$

where $f_i(t)$ is a smooth function that characterizes the difference of the ChIP-seq enrichment profiles and $W_i(t_k)$ is Gaussian noise with mean 0 and variance 1. For the ith gene, the null hypothesis that there is no differential enrichment between two conditions is equivalent to testing

$$H_0 : f_i(t) = 0. \tag{3.2}$$

3.3. Kernel-smoothing-based Nonparametric Tests

For a given gene i, we propose a kernel-smoothing based nonparametric test (Lepski and Spokoiny, 1999) to test the null hypothesis (3.2). For notational simplicity, we omit the subscript i in the following. Let K be a proper kernel, which is a symmetric, continuous density function with expectation zero. We use a normal kernel function, which satisfies all these regularity conditions and fits the real data well. For a fixed bandwidth value $\lambda \in [0,1]$, we consider the kernel estimator $\tilde{Y}_\lambda(t)$ with $t \in [0,1]$,$s \in [0,1]$ and its standard decomposition as

$$\begin{aligned} \tilde{Y}_\lambda(t) &= \frac{1}{\lambda}\int K\left(\frac{t-s}{\lambda}\right)Y(s)ds \\ &= \frac{1}{\lambda}\int K\left(\frac{t-s}{\lambda}\right)f(s)ds + \frac{\sigma}{\lambda}\int K\left(\frac{t-s}{\lambda}\right)W(s)ds \\ &= f_\lambda(t) + \sigma\xi_\lambda(t) \end{aligned} \tag{3.3}$$

where $f_\lambda(t) = \frac{1}{\lambda}\int K(\frac{t-s}{\lambda})f(s)ds$ and $\xi_\lambda(t) = \frac{1}{\lambda}\int K(\frac{t-s}{\lambda})W(s)ds$.

Based on Lepski and Spokoiny (1999), we use the integral of the squared kernel estimator T_λ defined as

$$T_\lambda = \frac{||\tilde{Y}_\lambda||^2}{\hat{\sigma}^2} = \frac{\int_0^1 \tilde{Y}_\lambda^2(t)dt}{\hat{\sigma}^2} \tag{3.4}$$

to test the null hypothesis $H_0 : ||f(t)|| = 0$, where $\hat{\sigma}^2$ is some estimate of the error variance, which we discuss in Section 3.3.2. Under the null H_0, one has

$$\tilde{Y}_{0\lambda}(t) = \sigma\xi_\lambda(t) \tag{3.5}$$

and the test statistic becomes $T_{0\lambda} = \int_0^1 \xi_\lambda^2(t)dt$. Since $W(t_i)$ follows $N(0,1)$, we have

$$\xi_\lambda(t) = \frac{1}{\lambda}\int_0^1 K\left(\frac{t-s}{\lambda}\right)W(s)ds$$

For the Gaussian kernel, the expectation of $T_{0\lambda}$ is given by

$$\mathrm{E}(T_{0\lambda}) = \frac{1}{n\lambda}||K||^2 = \frac{1}{n\lambda}\frac{1}{2\sqrt{\pi}}.$$

We derived the closed-form variance as

$$\mathrm{Var}(T_{0\lambda}) = \frac{1}{n^2\lambda}\frac{1}{\sqrt{2\pi}}.$$

(see Appendix B.1 for derivation). We can then define the test statistic as

$$Z_{0\lambda} = \frac{T_\lambda - \mathrm{E}(T_{0\lambda})}{\sqrt{\mathrm{Var}(T_{0\lambda})}}, \tag{3.6}$$

which follows $N(0,1)$ as $n \to \infty$ under the null hypothesis.

3.3.1. An alternative derivation of the test statistic

We present in this section an alternative derivation of the test statistic that has better finite sample performance than the statistic (3.6) when n is not too large (see Section 3.4 for an illustration). Note that the kernel smoother $\tilde{Y}_\lambda(t)$ can be written as a linear combination of $Y^{\mathrm{T}} = (Y_1, ..., Y_n)$,

$$\tilde{Y}_\lambda(t) = S_\lambda Y, \tag{3.7}$$

where S_λ is considered as the hat matrix,

$$S_\lambda = \frac{1}{n\lambda}\begin{pmatrix} K(\frac{t_1-s_1}{\lambda}) & \dots & K(\frac{t_1-s_n}{\lambda}) \\ \vdots & \ddots & \vdots \\ K(\frac{t_n-s_1}{\lambda}) & \dots & K(\frac{t_n-s_n}{\lambda}) \end{pmatrix}.$$

and the trace of S_λ is the degrees of freedom (df) of the kernel smoother (Hastie and Tibshirani, 1990).

Based on (3.3), (3.4) and (3.7), the statistic T_λ can be approximated by

$$T_\lambda = \frac{1}{n\sigma^2}\sum_{k=1}^{n}\tilde{Y}_{k\lambda}^{\ 2} = \frac{1}{n\sigma^2}Y^{\mathrm{T}}S_\lambda^{\mathrm{T}}S_\lambda Y \tag{3.8}$$

where the $n \times n$ matrix S_λ^{T} is the transpose of S_λ. Let $M = S_\lambda^{\mathrm{T}}S_\lambda$ with the following eigen-decomposition, $V^{\mathrm{T}}M\ V = D$, where $D = diag(d_1, ..., d_n)$, $d_1 \geq ... \geq d_n$, are the eigenvalues and V is the orthogonal matrix of the eigenvectors. Under the null, based on (3.5), Y/σ follows a multivariate normal distribution $N_n(0, I_n)$. Let $U^{\mathrm{T}} = (U_1, ..., U_n) = V^{\mathrm{T}}Y/\sigma$, we can rewrite T_λ as

$$T_\lambda \ = \ \frac{1}{n}U^{\mathrm{T}}DU = \frac{1}{n}\sum_{k=1}^{n} d_k U_k^2.$$

Since V is an orthogonal matrix, under the null hypothesis, the vector U follows $N_n(0, VV^{\mathrm{T}}) = N_n(0, I_n)$ and therefore U_k^2 are $i.i.d$ random variables following χ_1^2 and T_λ follows a mixture of n χ^2 distributions with weights d_k/n. Furthermore, based on Bentler and Xie (2000), under the null, T_λ can be approximated by a weighted χ^2 distribution, $\delta\chi_d^2$, where

$$d = \lceil (\sum_{k=1}^{n} d_k)^2 / \sum_{k=1}^{n} d_k^2 \rceil,\ \delta = \left(\sum_{k=1}^{n} d_k/n\right)/d.$$

Alternatively, using the Wilson-Hilferty transformation (Wilson and Hilferty, 1931), we have

$$Z_{0\lambda,WH} = \frac{\sqrt[3]{\frac{T_\lambda}{\delta d}} - \left(1 - \frac{2}{9d}\right)}{\sqrt{\frac{2}{9d}}}, \tag{3.9}$$

which follows a $N(0,1)$ under the null hypothesis (see Appendix B.2 for details). We use this statistic in our analysis.

3.3.2. Estimate σ for each gene

In order to calculate the test statistic specified as (3.4) or (3.8), we need the variance estimate ${\hat{\sigma}_i}^2$ for each gene i. After the transformation steps in Section 3.2, for each gene i, we assume that the observations Y_{ik} have the same variance σ_i^2. We consider the Nadaraya-Watson nonparametric regression with kernel smoothers as (3.3),

$$\tilde{Y}_\lambda(t) = S_\lambda Y$$

where $df = tr(S_\lambda)$ is the degrees of freedom of the kernel smoother (Hastie and Tibshirani, 1990). We can estimate the variance σ_i^2 by calculating the residual sum of squares

$$\hat{\sigma}^2 = \frac{[\tilde{Y}_\lambda(t) - Y(t)]^{\mathrm{T}}[\tilde{Y}_\lambda(t) - Y(t)]}{n - df} = \frac{\sum_{k=1}^{n}[Y_k - \tilde{Y}_\lambda(t_k)]^2}{n - df}. \tag{3.10}$$

Since we consider the ChIP-seq data with very few or no replications, the estimates $\hat{\sigma}_i^2$ can be too small for very small counts. To improve precision, we use an approach similar to Efron et al. (2001) and Tusher et al. (2001): we add a constant $a_0 = 90th$ percentile of the standard deviations to make the standard deviation of each gene

bigger to avoid false identification of genes with differential enrichment. The final modified estimator of the variance is $\tilde{\sigma}_i^2 = (\hat{\sigma}_i + a_0)^2$.

Finally, we choose the bandwidth in the kernel smoothing λ relatively large to avoid fitting the very small local changes. In our analysis of the real data sets with $n = 280$ observations, we choose $\lambda = 20/280$. The details of bandwidth selection are discussed in Chapter 3.5.

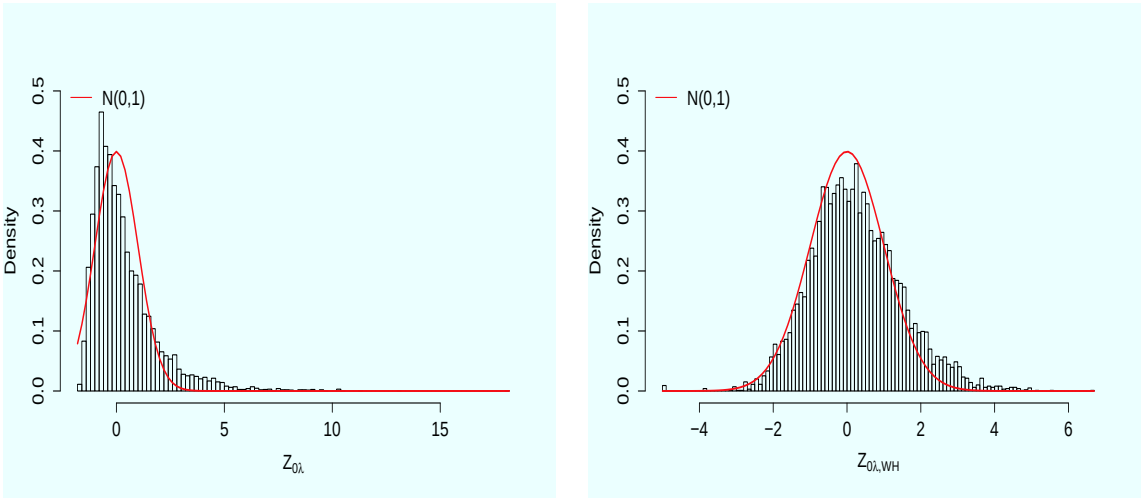
3.4. Application to a Comparative ChIP-seq Study During Mouse Adipogenesis

We present results of our analysis of the comparative ChIP-seq data described in Section 3.2. Our initial analysis focused on H3K27ac at gene promoter regions since it is known that H3K27ac is positively associated with gene expression (Mikkelsen et al., 2010). We divided the DNA region around the transcription starting site (-5000 to 2000 bp) into n=280 bins, where the length of each bin is 25 bps. The data set includes m=29,716 genes. Our goal is to identify the genes with differential H3K27ac levels at the promoter regions between proliferating preadipocytes (day -2) and mature adipocytes (day 7).

3.4.1. Comparison of the $Z_{0\lambda,WH}$ statistics and fold-change statistics

For each gene, after the normal-transformation as in Section 3.2, we fit a kernel-smoothing function to the difference data using a bandwidth of $\lambda = 20/280$, which over-smooth the very small signals that are likely due to noise. We calculate the test statistic for each of the 29,716 genes. To compare different test statistics $Z_{0\lambda}$, and $Z_{0\lambda,WH}$, we plot the histograms of these two test statistics in Figure 3.1 for 9,874 genes with the maximum number of read counts in both days fewer than 5.

Due to the very small read counts in these genes, these genes are most likely not differentially enriched and therefore the test statistics should follow the standard normal distribution. Clearly, $Z_{0\lambda,WH}$ follows $N(0,1)$ closer than $Z_{0\lambda}$. We therefore use this statistic in all the following analyses.



(a) (b)

Figure 3.1: Histograms of two test statistics for the mouse adipogenesis ChIP-seq data, (a) $Z_{0\lambda}$ and (b) $Z_{0\lambda,HW}$, for 9,874 genes with the maximum number of read counts in both day -2 and day 7 fewer than 5. The red curve in each plot represents the standard normal density.

Using the test statistic $Z_{0\lambda,HW}$, we observed that about one-third of the genes showed differential enrichment between preadipocytes and mature adipocytes using a Bonferroni-adjusted p-value of 0.05. This is expected since the cells are very different between these two days. Large-scale differential enrichment was also observed in Mikkelsen et al. (2010). We observe different patterns of differential enrichment. Figure 3.2 shows the observed data for 12 genes with the largest test statistics. Clearly, for some genes, H3K27ac is only present in one condition. Genes that were enriched at both time points showed clearly different H3K27ac levels.

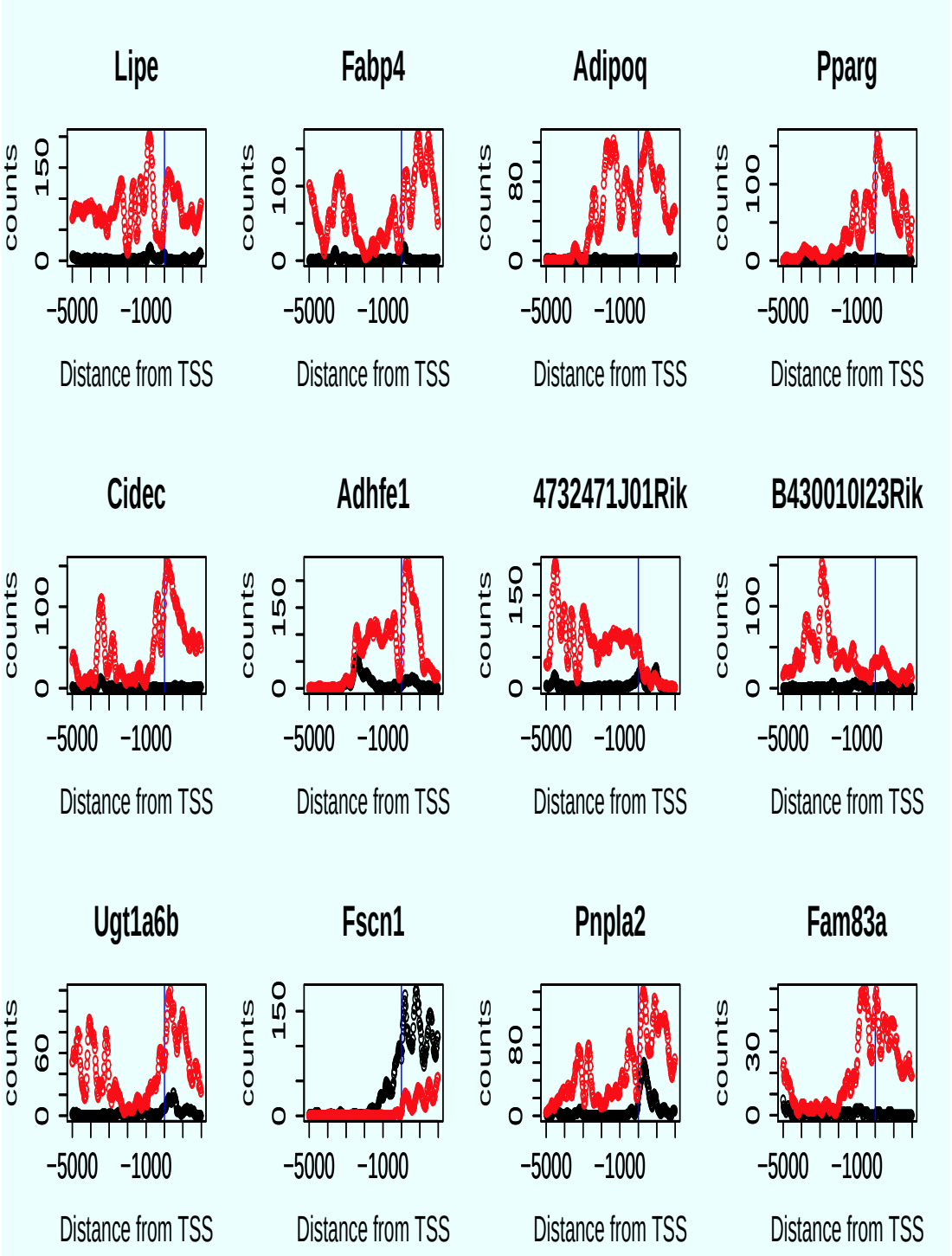


Figure 3.2: Observed mouse adipogenesis ChIP-seq bin-counts for top twelve genes ranked by the test statistics $Z_{0\lambda,WH}$ over the promoter region for day -2 (red) and day 7 (black). Vertical line represents the transcription starting site.

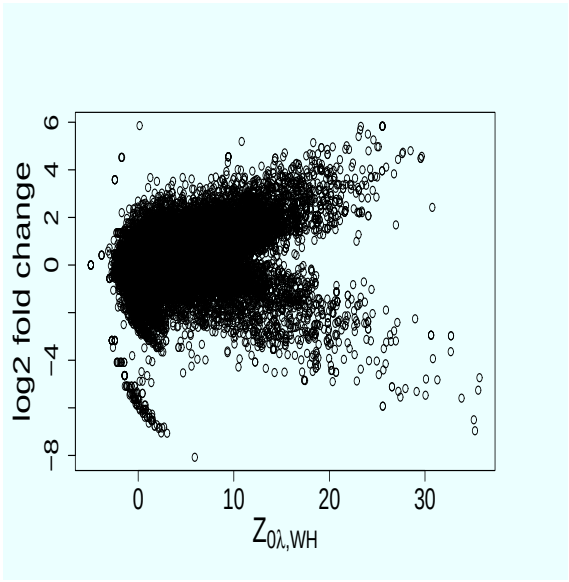
As a comparison, for each of the genes, we also calculate the simple fold-change statistics and the statistics used in DBChIP(Liang and Keleş, 2012). Figure 3.3 (a) and (b) shows the plots of our proposed statistics versus the fold-change statistics and the DBChIP statistics. Since the DBChIP statistics are almost identical to the fold-change statistics in (c), we only compare results with the fold-change statistics in the following. In general, we observe that large $Z_{0\lambda,HW}$ statistics correspond to large fold-changes or large DBChIP statistics. We observed a small set of genes that have very small $Z_{0\lambda,WH}$-statistics, but with very large fold changes or DBChIP statistics. These genes tend to have very small read counts. We also observe that some genes have very small fold-changes, but with large $Z_{0\lambda,HW}$-statistics. Figure 3.4 shows the plots of 12 such genes. Many of such genes show a clear shift of peaks between two different cell states, which cannot be captured simply using total read counts as in fold-changes and the DBChIP statistics. This indicates the importance of modeling the spatial ChIP-seq enrichment profiles.

3.4.2. Differential enrichment statistics and gene expression changes

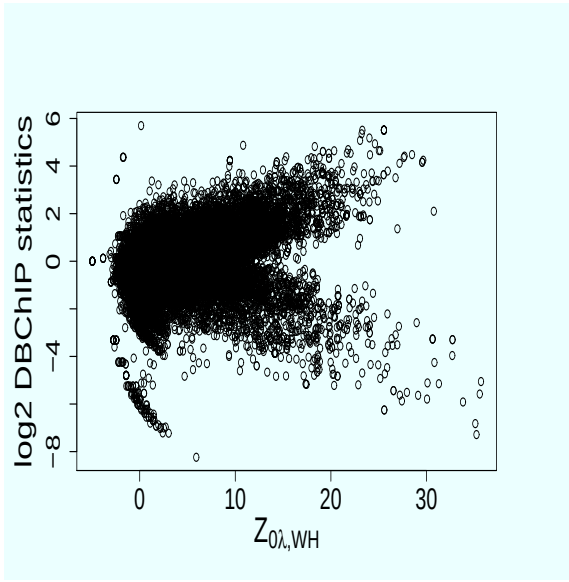
We next investigate the relationship between our test statistics $Z_{0\lambda,WH}$ and changes in expressions of the genes between the two time points. The gene expression data contains two replicates for each condition, and we take the average of two replicates as the mean value W_{ij} for each gene $i = 1, \cdots, m$ and condition $j = 1, 2$. We define the $\log_2$ of the fold-change of the expression levels as

$$\Delta W_i = \log_2 \frac{W_{i2}}{W_{i1}}$$

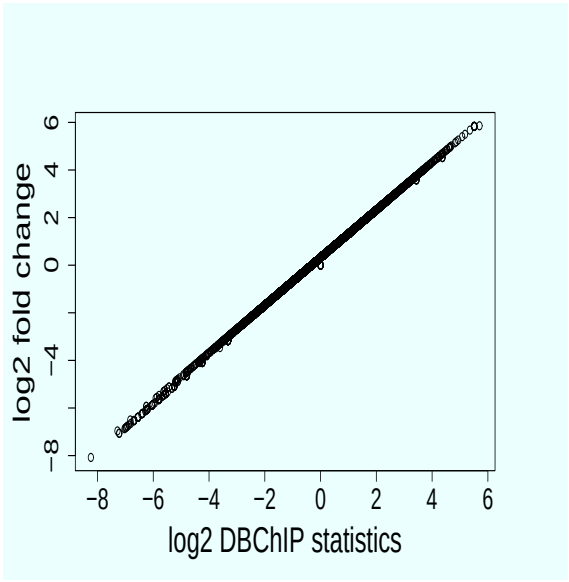
for the ith gene. We then divided genes into two groups depending on whether higher enrichment was observed at day 7 or day -2. Specifically, we fit the kernel smoothing



(a)



(b)



(c)

Figure 3.3: Comparison of (a) the proposed statistics and the fold-changes statistics, (b) the proposed statistics and the DBChIP statistics, and (c) the fold-change statistics and the DBChIP statistics for the mouse adipogenesis ChIP-seq data.

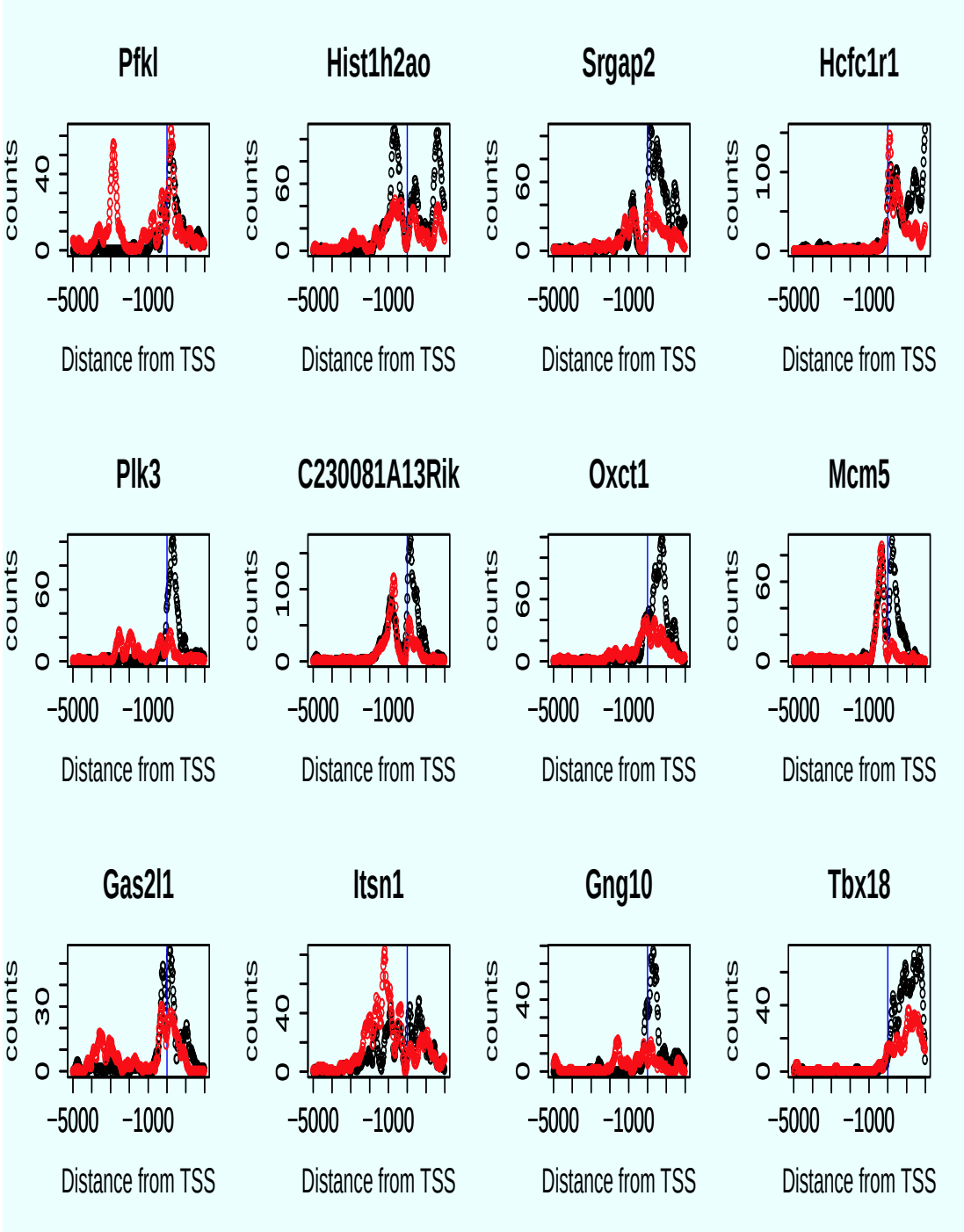


Figure 3.4: Observed mouse adipogenesis ChIP-seq bin-counts over the promoter region for day -2 (red) and day 7 (black) for twelve genes with large $Z_{0\lambda,WH}$ but small fold changes. Vertical line represents the transcription starting site.

curve to data for each gene under day 7 and day -2 and obtain the maximum of the curves. The genes are classified as being enriched at day 7 (or day -2) if the maximum height is higher at day 7 (or day -2). Figure 3.5 shows the gene expression fold changes against the test statistics $Z_{0\lambda,WH}$ together with the Lowess fit for genes that are enriched at day -2. We observe that larger enrichment statistics correspond to down-regulation of these genes. Similarly, Figure 3.5 also shows the gene expression fold changes against the test statistics $Z_{0\lambda,WH}$ together with Lowess fit for genes that are enriched at day 7. We observe that larger statistics correspond to up-regulation of these genes. Both plots make biological sense since enrichment of H3K27ac is known to activate gene expression. As comparisons, similar plots are given in Figure 3.5 for the fold-change statistics. The patterns from fold-change statistics are not as clear as using our proposed statistics $Z_{0\lambda,WH}$.

To demonstrate this further, we define gene i as being up-regulated if $\Delta W_i > 1$ and down-regulated if $\Delta W_i < -1$. In Figure 3.6 (a), we divide our test statistics $Z_{0\lambda,WH}$ into equal-length intervals $(< 0, 0-5, 5-10, 10-15, 15-20, > 20)$ for the genes that have higher enrichment at day -2. We observe that the proportion of down-regulated genes increases as test statistics increase. One the other hand, the proportions remain almost constant and close to zero for up-regulated genes. On the other hand, in Figure 3.6 (b), for the genes that have higher enrichment at day 7, we observe exactly the opposite. This indicates that our statistics correspond to gene expression changes very well. As a comparison, we present similar plots based on dividing the genes based on fold changes of the total reads counts (See Figure 3.6 (c) and (d)). We observed that the separations are not as clear as using our proposed statistics.

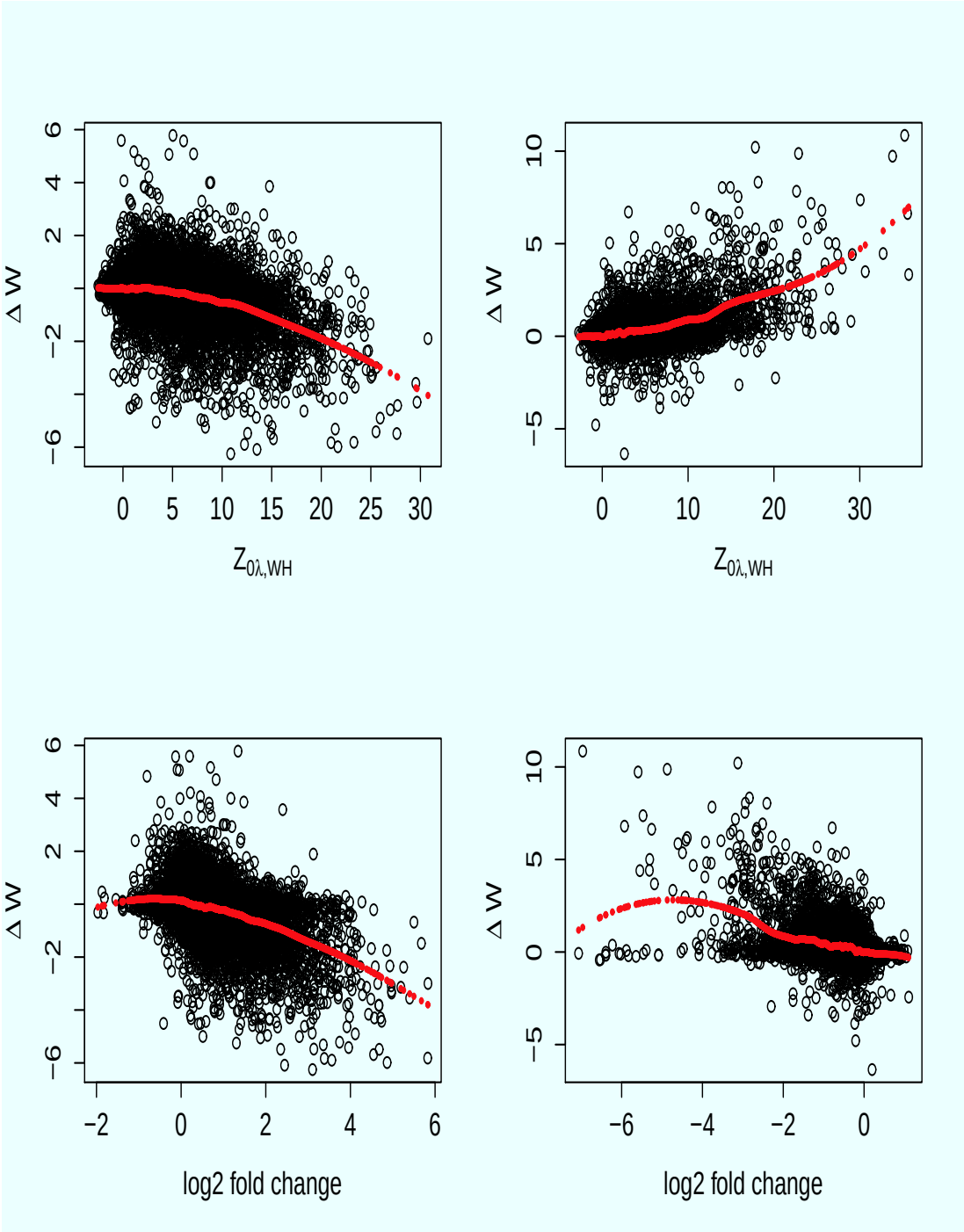


Figure 3.5: Plots of gene expression fold changes as a function of two different test statistics for the mouse adipogenesis ChIP-seq. Top: proposed smoothing-kernel test statistics; bottom: fold changes. Left panel: genes with enriched H3K27ac at day -2; right panel: genes with enriched H3K27ac at day 7.

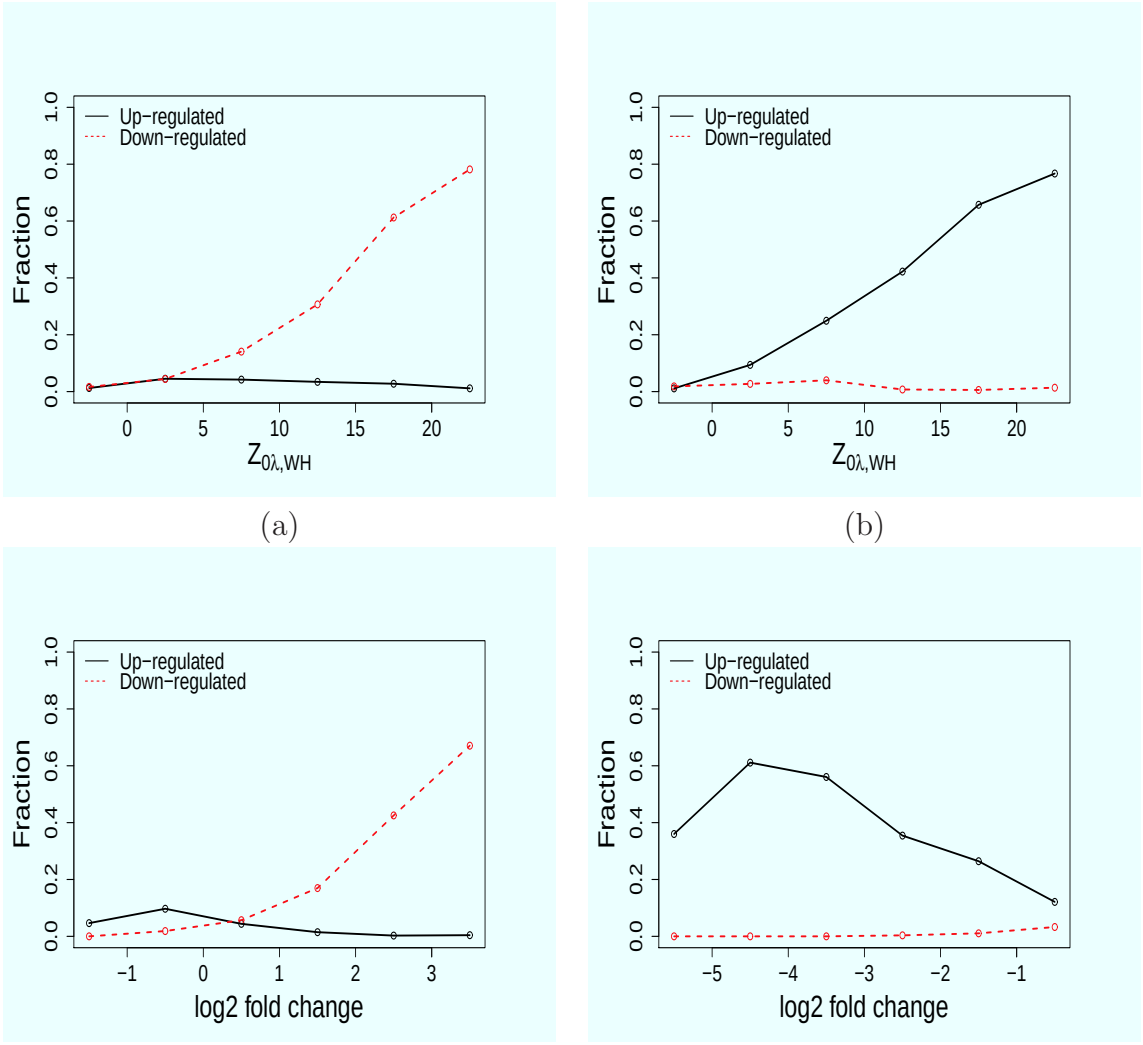


Figure 3.6: Plots of proportions of up/down-regulated genes in different intervals of the test statistics for the mouse adipogenesis ChIP-seq data. (a)-(b): proposed smoothing-kernel test statistics; (c)-(d): fold change statistics. (a), (c): genes with enriched H3K27ac at day -2; (b), (d): genes with enriched H3K27ac at day 7.

3.4.3. Prediction of gene expression fold changes using histone modification profiles

We next evaluate how well our proposed statistics can be used for predicting the fold changes of gene expression using ChIP-seq data. Besides the H3K27ac ChIP-seq data, we also have data from another five histone modification marks, including H3K4me1, H3K4me2, H3K4me3, H3K27me3 and H3K36me3. In addition, for each gene, besides the promoter region, we also consider the histone modifications in gene body and downstream regions. We evaluate the prediction for fold changes of gene expression by randomly selecting half of the genes as the training set and fit a linear regression model,

$$\Delta W_i = \beta_0 + \sum_{h=1}^{6} \sum_{l=1}^{3} \beta_{hl} TS_{i,hl}, \tag{3.11}$$

where h indexes the six histone modification marks and h indexes promoter region, gene body and downstream region, $TS_{i,hl}$ is the differential histone enrichment statistics for HM h for the ith gene at the lth location. Using the fitted model, we then predict the gene expression for the left-out genes. We repeated this 100 times and calculated the average R^2 for model fits for the training genes and the prediction error for genes in the testing sets. As a comparison, we also considered the same model as (3.11) using the simple fold change statistics as the predictors. Figure 3.7 shows the model fit for training genes and prediction results for testing genes using our proposed statistics $Z_{0\lambda,WH}$ and the fold change statistics as predictors. Clearly we observe that our proposed statistics give a much better model fit and better prediction results. The average R^2 over 100 random splitting of the genes is 0.57 using our statistics and 0.46 using simple fold changes, and the average prediction error is 0.47 using our statistics and 0.59 using simple fold changes.

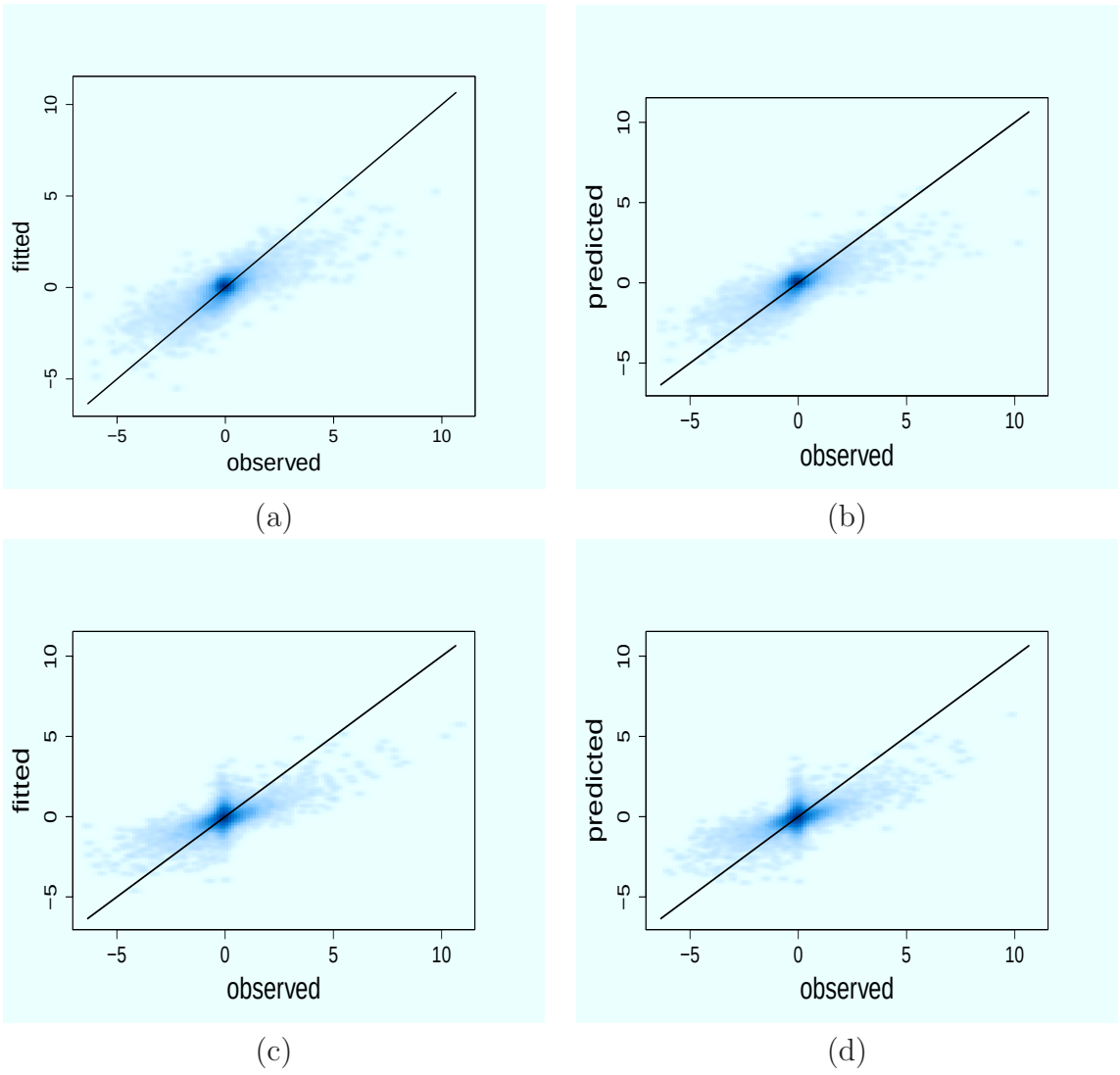
We also observed that histone modification dynamics at the promoter and gene body are more predictive than the signals in the downstream regions for predicting the gene expression changes (see Table 3.1 for details). This is expected since the histone modification marks we used are associated with transcription initiation (H3K4me3), open chromatin (H3K4me1/me2 and H3K27ac), transcription elongation (H3K36me3) and Polycomb-mediated repression (H3K27me3).

Table 3.1: Comparison of model fit R^2 and prediction (PE) of gene expression fold changes using the proposed statistic $Z_{0\lambda,WH}$ and fold change based on ChIP-seq data of promoter, gene body and downstream regions of all six histone modification marks as predictors and models using all three regions. The results are based on 100 runs of randomly selecting half of the genes as training set and another half as testing set. Numbers in parentheses are standard errors.

	$Z_{0\lambda,WH}$		Fold change	
	R^2	PE	R^2	PE
Promotor	0.45 (0.009)	0.60 (0.012)	0.35 (0.009)	0.72 (0.015)
Gene body	0.49 (0.008)	0.57 (0.015)	0.40 (0.011)	0.66 (0.014)
Downstream	0.30 (0.009)	0.78 (0.018)	0.18 (0.007)	0.90 (0.023)
All regions	0.57 (0.008)	0.47 (0.013)	0.46 (0.009)	0.59 (0.012)

3.5. Effects of Bandwidth Selection on Identifying the Genes with Differential Enrichment

In applying our kernel-based test in analyzing the mouse ChIP-seq data, we used a global bandwidth of $\lambda = 20/280$ for all the genes. Since the algorithm performs around 30,000 tests to find a list of genes with significant differentially enriched regions, the bandwidth used in the tests should be fixed to the same value. In addition, any reasonable test should capture the spatial profiles of signals in the gene regions of interest. On the other hand, the test should also smooth out the small local noises, which are not biologically interesting. We suggest using a relatively large bandwidth to reduce possible false positives. Alternatively, the standard method is



(a) (b)

(c) (d)

Figure 3.7: Model fit (left panel) and prediction (right panel) for log of the gene expression fold changes using the proposed statistics $Z_{0\lambda,HW}$ and fold changes of six histone-modification ChIP-seq data at promoter, gene body and downstream region.

using cross-validation to find the optimal rate $c(1/n)^{1/5}$ (Gasser et al., 1991). In addition, Neumeyer and Dette (2003) suggests to obtain the nonparametric variance estimator $\hat{\sigma}_i^2$ (Rice, 1984) for each gene. We can then summarize these variance estimates using the median value and to estimate the bandwidth by

$$\lambda = \left\{ \frac{\text{median}(\hat{\sigma}_i^2, i = 1, ..., n)}{n} \right\}^{1/5}.$$

We further check the sensitivity of bandwidth selection on the performance of our proposed kernel-based test by considering a set of different bandwidth values, $\lambda_1 = 5/280$, $\lambda_2 = 20/280$, $\lambda_3 = 60/280$, and $\lambda_4 = 90/280$. Here, λ_3 and λ_4 correspond to the bandwidths chosen by the nonparametric variance estimation method (Neumeyer and Dette, 2003) and the optimal rate $(1/n)^{1/5}$ (Gasser et al., 1991), respectively. We calculate the kernel-based test statistics and denote these statistics as $Z_{\lambda_l,WH}$, $l = 1, 2, 3, 4$. We present in Figure 3.8 the histograms of $Z_{\lambda_l,WH}$, $l = 1, 2, 3, 4$ for the 9,874 genes with the maximum number of read count in both days fewer than 5, which are analogues to the plot (b) in Figure 3.1. Clearly, the statistics $Z_{\lambda_1,WH}$ with a relatively small bandwidth lead to false positive detection where the distribution of null genes clearly deviates to the right side of $N(0, 1)$. On the other hand, when a large bandwidth is used, as in statistics $Z_{\lambda_3,WH}$ and $Z_{\lambda_4,WH}$, the tests are conservative, although they still fit the standard normal density curves (red line) reasonably well.

We also examine how different bandwidths affect the ability of identifying differentially expressed genes, where a gene is defined as a true differentially expressed gene if $|\Delta W_i| > 1$. The ROC curves in Figure 3.9 show that in general, larger bandwidth gives better results than smaller one. Overall, we observe that it is essential to smooth out the small local signals in order to reduce false positive identification of genes with

differential enrichment.

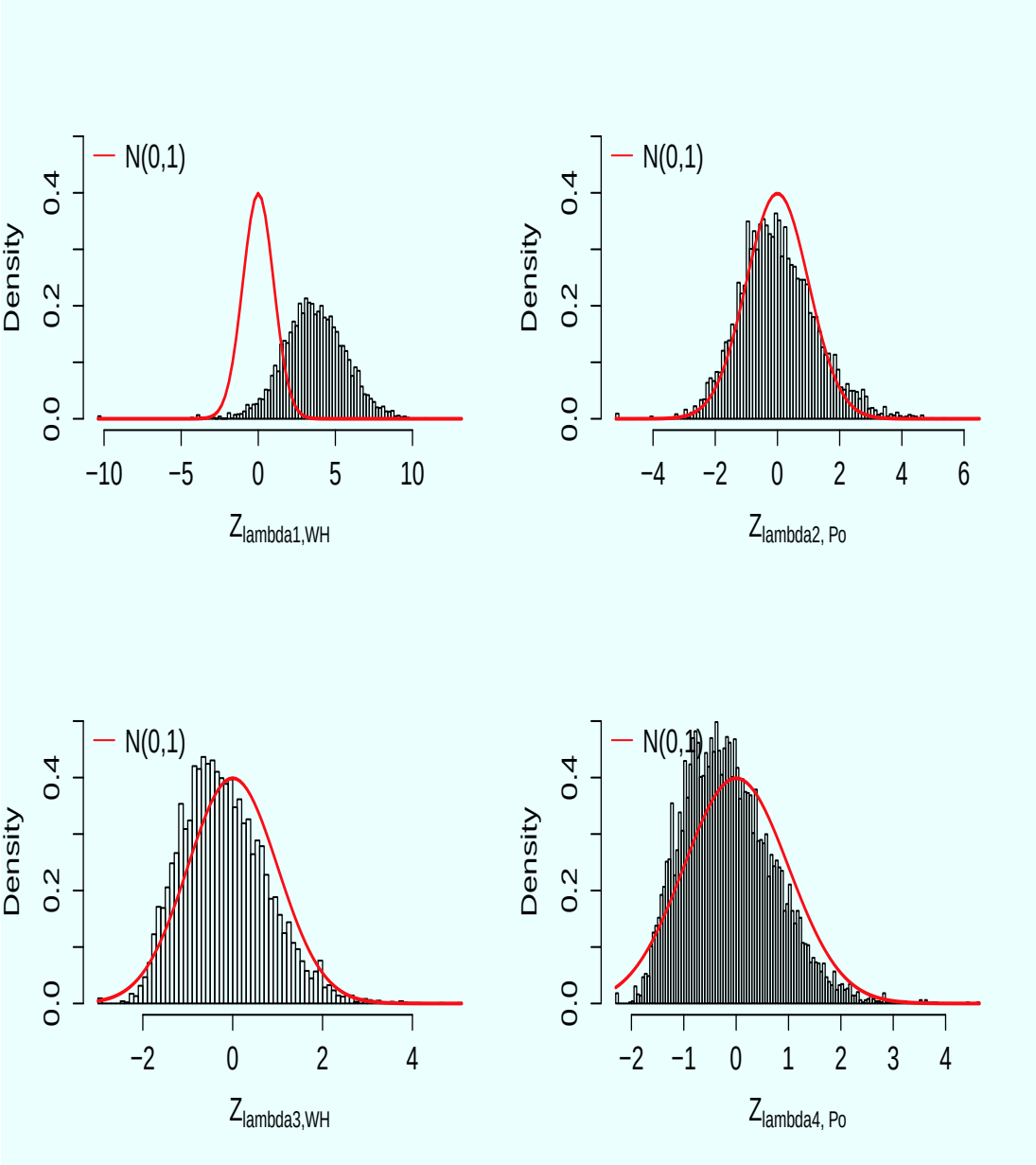


Figure 3.8: Histogram of the test statistics $Z_{\lambda_t,WH}$ with the different bandwidths: (a) $\lambda_1 = 5/280$ (b) $\lambda_2 = 20/280$ (c) $\lambda_3 = 60/280$ (d) $\lambda_4 = 90/280$ for 9,874 genes with the maximum number of read count in both day -2 and day 7 fewer than 5 in mouse adipogenesis ChIP-seq data.

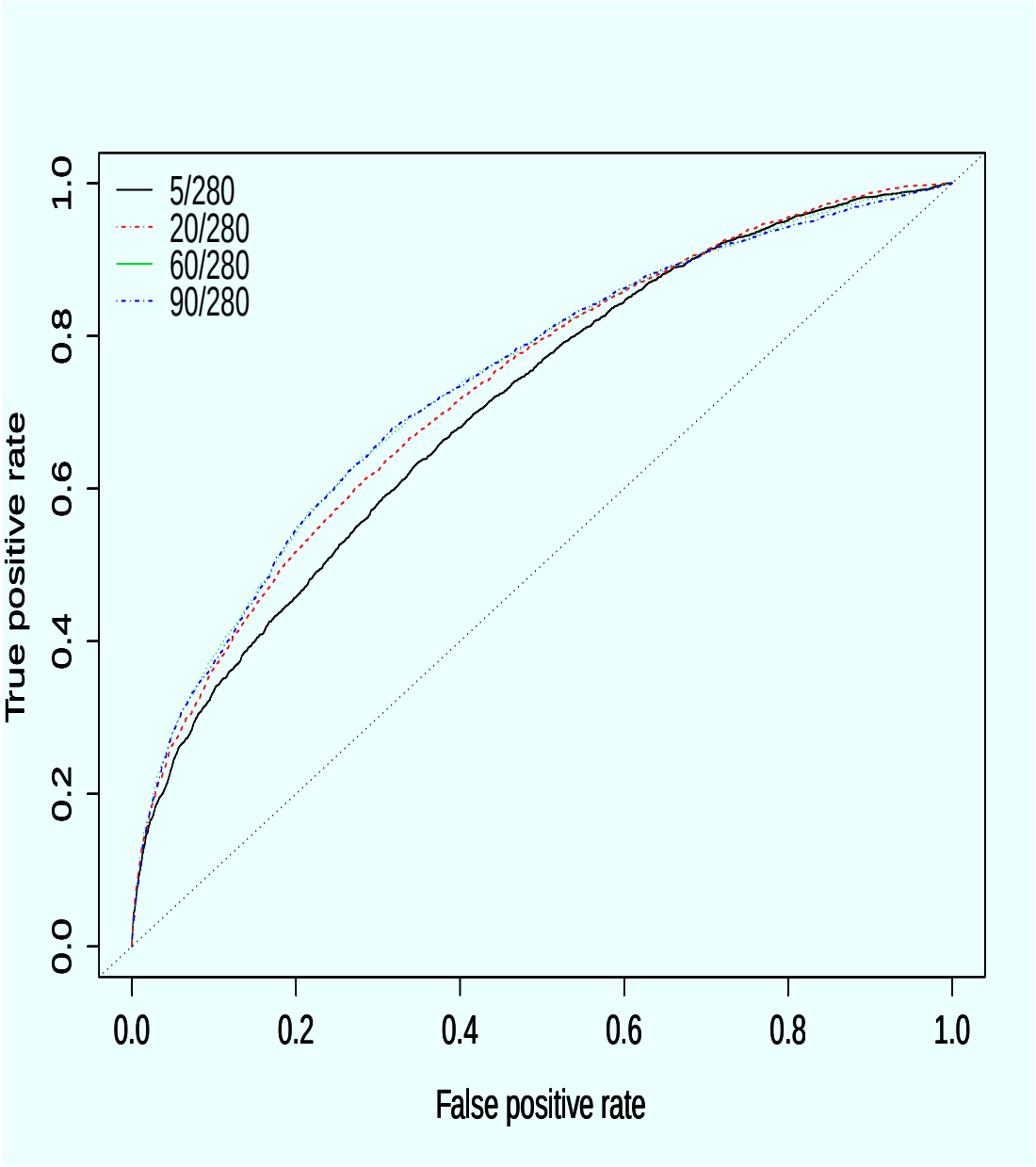


Figure 3.9: ROC curves for identifying differentially expressed genes between day -2 and day 7 using test statistics $Z_{\lambda_t,WH}$ with the different bandwidths: (a) $\lambda_1 = 5/280$ (b) $\lambda_2 = 20/280$ (c) $\lambda_3 = 60/280$ (d) $\lambda_4 = 90/280$ for all the genes in mouse adipogenesis ChIP-seq data.

3.6. Application to an ENCODE ChIP-seq Data with Two Replicates

To further evaluate the possible false positives in identifying genes with differential enrichment of histone modification, we analyze the ChIP-seq data sets reported in the ENCODE project (ENCODE Project Consortium et al., 2012) for a B-lymphoblastoid cell line of human GM12878, which is also part of the 1000 Genomes project, and HeLa-S3 cervical carcinoma cells. Our analysis still focuses on the H3K27ac mark at the promoter regions of the genes with count data available in $n = 280$ bins for each gene. In this experiment, there are a total of $m^* = 23807$ genes. Besides the ChIP-seq data for two biological replicates, two input data are also available. Ideally, we should not expect any genes with differential enrichment between the two replicates. We apply the same procedure as in our analysis of the mouse data in Section 3.4 to the data between two ChIP-seq replicates and calculate test statistics $Z_{new,i}$ for each gene i, $i = 1, ..., m = 23807$. The histogram of Z_{new} for all the genes in Figure 3.10 (top plot) shows that the majority of the test statistics follow the standard normal distribution. In addition, using a Bonferroni adjusted p-value of 0.05, our procedure identifies only 263 genes that show differential enrichment between two replicates, which results in a less than 1.5 % false discovery rate. This analysis further demonstrates that our proposed kernel-based nonparametric testing procedure is not only powerful enough to detect the true differentially enriched regions but also makes fewer false detections.

Finally, we also perform an analysis to identify the genes with differential enrichment of histone modification between a B-lymphoblastoid cells and HeLa-S3 cervical carcinoma cells. Figure 3.10 (bottom plot) shows the histogram of the test statistics for all 23807 genes. Using a Bonferroni threshold for genome-wide level of 0.05, we identify 6647 genes that show differential H3K27ac levels at their promoter regions.

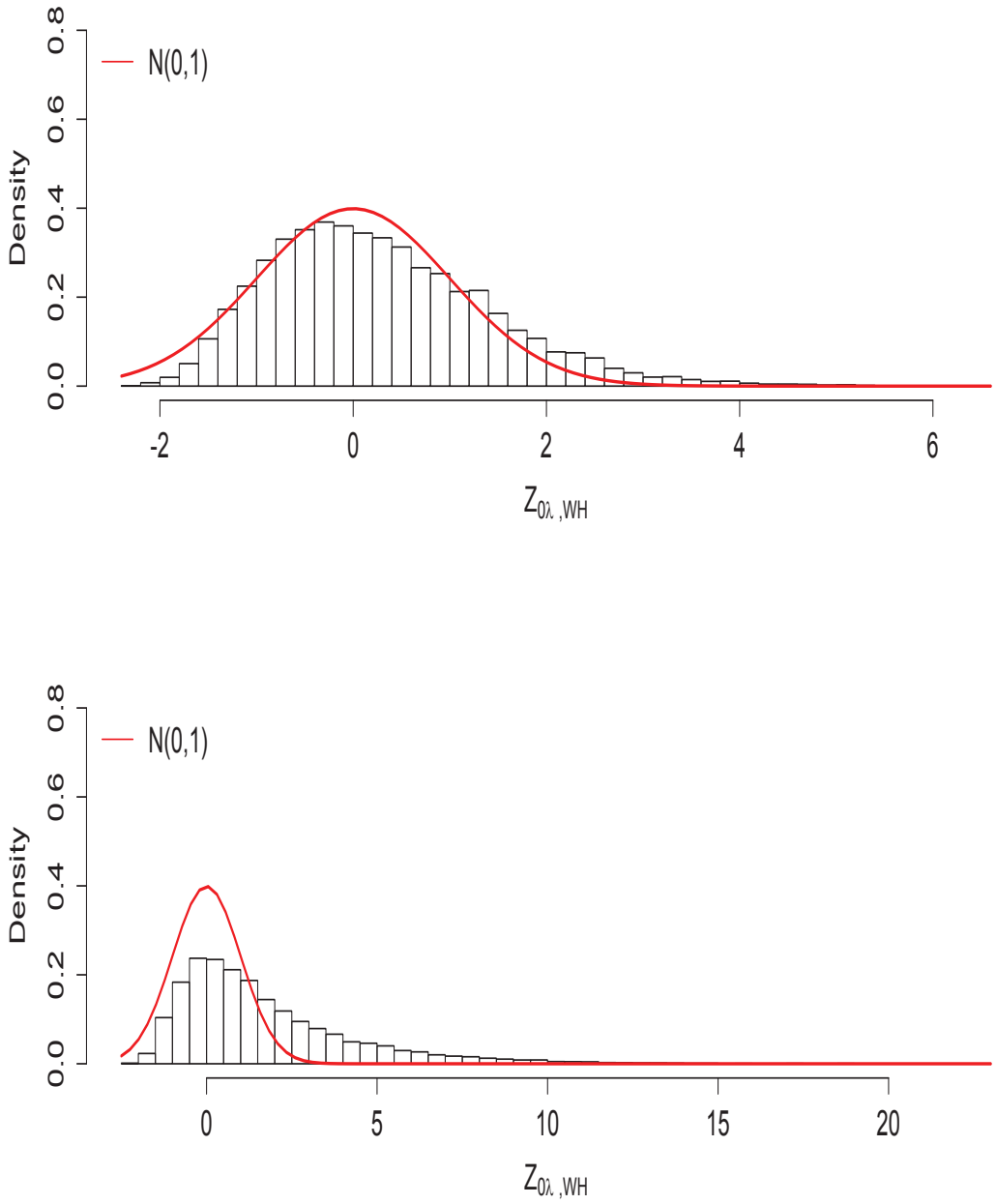


Figure 3.10: Top: Histogram of differential enrichment test statistics Z_{new} between two biological replicates of the ENCODE data for all 23807 genes. Bottom: Histogram of differential enrichment test statistics Z_{new} between two cell types (B-lymphoblastoid cell vs HeLa-S3 cervical carcinoma cells) of the ENCODE data for all 23807 genes. The red curve represents the standard normal density.

3.7. Extension to Multiple Experimental Conditions and ANOVA-type Test Statistics

Our proposed method can also be extended to identify differential enrichment in multiple conditions. Motivated by the same ChIP-seq data of Mikkelsen et al. (2010) with H3K27ac at four time points, we are interested in identifying genes that show any changes of H3K27ac levels at promoter regions during the four time points. Instead of fitting kernel smoothing curves on one-sample difference $Y(t) = X_1^*(t) - X_2^*(t)$, we fit the kernel on $X_j(t)$ for each condition j. For each gene k and condition j, $k = 1, ..., m = 29716$ and $j = 1, 2, 3, J = 4$, we assume the data follow a "signal + noise" model (omitting k),

$$X_j(t) = f_j(t) + W_j(t).$$

For each gene, the null hypothesis of interest is

$$H_0 : f_1(t) = f_2(t) = f_3(t) = f_4(t) = f(t). \tag{3.12}$$

Motivated by the ANOVA statistics to test the equality of the means in multiple-sample cases (Young and Bowman, 1995; Dette and Neumeyer, 2001), we propose the following statistic for testing the null hypothesis (3.12),

$$\begin{aligned} TS^{anova} &= \frac{\sum_j^J n(\frac{1}{n}\sum_{i=1}^n \hat{f}_j(t_i) - \bar{f}(t_i))^2}{\sum_j^J \sum_{i=1}^n (X_j(t_i) - \hat{f}_j(t_i))^2} \\ &= \frac{\sum_j^J n(\frac{1}{n}\sum_{i=1}^n \hat{f}_j(t_i) - \bar{f}(t_i))^2}{\sum_j^J \tilde{\sigma}_j(n - df)} \end{aligned} \tag{3.13}$$

where $\hat{f}_j(t) = \tilde{Y}_\lambda(t)$ and $\tilde{\sigma}_j$ is defined similar as (3.7) (3.10) for each condition j, and

$\bar{f}(t) = \sum_{j=1}^{J} \hat{f}_j(t)/J$. The statistic should follow an F-distribution,

$$\begin{aligned} F &= TS^{anova} \times \frac{J(n-df)}{J-1} \\ &= \frac{\sum_j^J n(\frac{1}{n}\sum_{i=1}^n \hat{f}_j(t_i) - \bar{f}(t_i))^2/(J-1)}{\sum_j^J \tilde{\sigma}_j/J} \xrightarrow{H_0} F(J-1, J(n-df)) \end{aligned} \quad (3.14)$$

To demonstrate the F distribution of ANOVA-type statistics, we apply the methods on a simple simulated data and the ChIP-seq data H3K27ac measured in four time points. In the simulation, we define the null genes that satisfy (3.12), which means $X_j(t)$ are just $i.i.d$ white noises. We simulate $m = 10000$ null genes with 4 observations over $n = 280$ bins. In Figure 3.11, the left panel shows the histogram and p-value of ANOVA-type test statistics for all the simulated genes. Clearly, the test statistics for these null genes follow an F distribution and the p-values follow a uniform distribution. This demonstrates that the null distribution of our test statistics indeed follows an F distribution.

Furthermore, we calculate the ANOVA-type statistics for each gene on the real data. We use a similar idea as in Chapter 3.3.2 to add a small constant $a_0 = 80\%$ percentile of the denominator of TS^{anova}. We observe about one-forth of the genes showing differential enrichment during the four time points. In Figure 3.11, the right panel shows the histogram and p-values of ANOVA-type test statistics for the genes with maximum number of read counts in all four conditions fewer than 5. We observe that the distribution of test statistics is close to an F distribution but with a slightly long tail, and the corresponding p-values slightly deviate from a uniform $U(0,1)$ distribution.

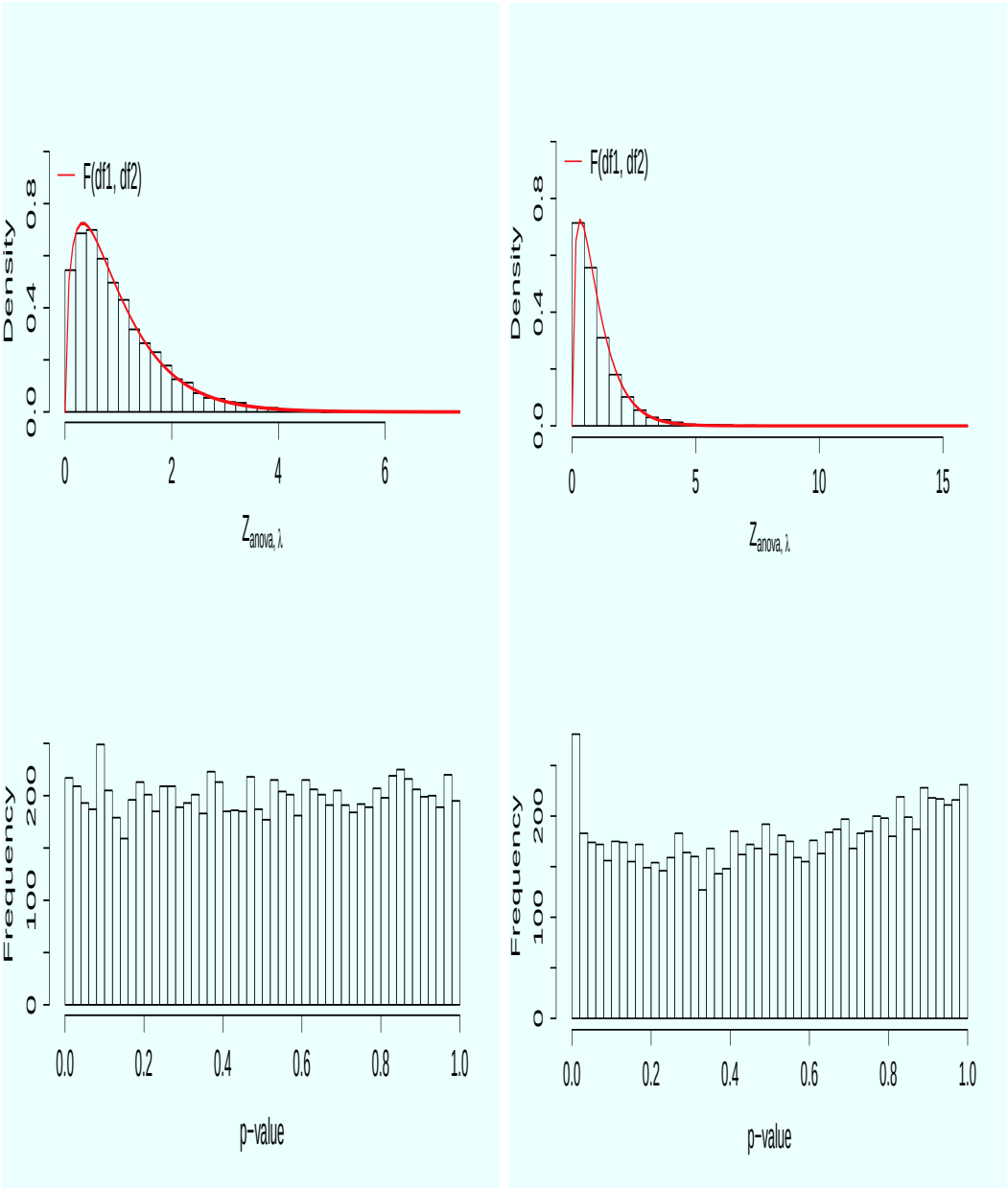


Figure 3.11: F-distributions of null genes for simulated data set (left panel) and mouse adipogenesis ChIP-seq data set (right panel). Top: histogram of the test statistics; Bottom: histogram of the p-values.

3.8. Conclusions and Discussion

We have proposed a kernel-smoothing based nonparametric test to identify genes with differential enrichment for ChIP-seq data. Different from all the currently available methods, our method models the spatial histone enrichment profiles at the promoter regions of the genes, rather than simply modeling the total read counts in a given window. The method can therefore capture different types of differences in histone-enriched profiles between two experimental conditions. To detect differences in enrichment profiles, we constructed a nonparametric statistic based on kernel-smoothing on the differences of the profiles after approximate normal transformation of the data. We have shown that the proposed statistics corresponds to the gene expression changes better than other statistics and the models based on a combination of different histone modification marks can effectively predict the gene expression fold changes. Although prediction of gene expression using the ChIP-seq data has been studied in many published works (Karlic et al., 2010; Dong et al., 2012), these papers focused only on prediction of gene expression at a static state. Our results further demonstrate that change of histone modifications and the dynamic chromatin signatures can also be very predictive for the fold-changes of gene expression between two different cellular states.

We considered only the problem of identifying the differential enrichment regions between two or more conditions, where we fit the kernel-smoothing to the differences of the normal transformed data in order to further smooth out the small local changes that might be due to differences in GC contents or mappability of the sequencing reads. By smoothing, we expect that our procedure is robust to such small changes due to genomic features. To identify differential enrichment in multiple conditions, we propose an ANOVA-type statistics (3.14). In Chapter 4, we will introduce new

statistics for multiple-sample enrichment analysis by taking the mean or maximum of the pair-wise statistics defined in this Chapter.

CHAPTER 4

Two Alternative Nonparametric Tests for Differential ChIP-seq Data Analysis

4.1. Introduction

In ChIP-seq studies, one important biological problem is to identify the genomic regions that show differential enrichment of the same histone modification mark (HM) between two or more experimental conditions (Mikkelsen et al., 2010). It is also important to detect the change of bivalent states between two or more HM marks (Xie et al., 2013). The ChIP-seq data can be summarized as counts of short reads in non-overlapping bins in the genomic regions of interest, e.g., promoter region or gene body. After an appropriate transformation, statistically, this problem can be formulated as testing the equality of L ($L \geq 2$) mean functions. The observed data at the kth bin on each condition j $X_j(t_k)$ could be modeled as

$$X_j(t_k) = f_j(t_k) + \sigma_j(t_k)W_j(t_k), \tag{4.1}$$

where $t_k = k/n \in [0, 1]$, $k = 1, ..., n$, $W_{jk} = W_j(t_k)$ are $i.i.d$ errors, $f_j(t)$ is a smooth function that characterizes the spatial enrichment profile of ChIP-seq data and $\sigma_j(t)$ is the variance function for condition j. We are interested in testing the null hypothesis,

$$H_0 : f_1(t) = \cdots = f_L(t) \tag{4.2}$$

with $L \geq 2$.

Nonparametric tests of the equality of functions in the two-sample case have been extensively studied (Hall and Hart, 1990; King et al., 1991; Munk and Dette, 1998; Lepski and Spokoiny, 1999; Neumeyer and Dette, 2003). In Chapter 3, this dissertation discussed a two-sample kernel based nonparametric testing procedure based on Lepski and Spokoiny (1999) and King et al. (1991). We applied the method to a ChIP-seq data from Mikkelsen et al. (2010) and identified genes with differential H3K27ac levels at promoter regions between two cellular states.

However, this method requires the following assumptions on error terms: 1) Gaussian error, $W_j(t_k)$ follows standard normal distribution; 2) variance of error term is a constant across bins (homoscedastic errors, $\sigma_j(t_k) = \sigma_j$). Furthermore, kernel smoothing requires one to choose a proper kernel function K and bandwidth λ for each gene. In many cases, the results would be sensitive to the choice of bandwidth or kernel functions (discussed in Section 3.5).

In this Chapter, we explore the nonparametric tests developed by Munk and Dette (1998) without using smoothing. In addition, the nonparametric tests relax the assumption of Gaussian errors with constant variances across n bins. Section 4.2 discusses the application of nonparametric tests to two-sample ChIP-seq comparisons. Two new test statistics proposed by Munk and Dette (1998) under homoscedasticity and heteroscedasticity assumptions are computed and discussed in Section 4.2.2 and 4.2.3, respectively. In Section 4.3.2, the performance of the new statistics is compared with the kernel-smoothing test statistic and fold changes statistic. We again analyze the ChIP-seq data obtained from Mikkelsen et al. (2010) and analyzed in Chapter 3. It includes histone modification data in four cellular states: proliferating (day -2), confluent preadipocytes (day 0), immature adipocytes (day 2) and mature adipocytes (day 7).

Besides differential enrichment analysis between two time points, it is also of interest to identify genes that show differential enrichment in any of these four cellular states. This motivates us to consider the problem of multi-sample ChIP-seq comparisons in Section 4.4 and to propose tests for testing the equality of L ($L > 2$) functions. Current approaches on multiple sample tests (Hardle and Marron, 1990; Young and Bowman, 1995; Munk and Dette, 1998; Neumeyer and Dette, 2003; Dette and Neumeyer, 2001; Cuevas et al., 2004) are mostly based on ANOVA-type statistics (discussed in Section 3.7) or sum of all pairwise two-sample statistics. In Section 4.4, we consider test statistics that are based on the maximum (or mean) of the pairwise statistics in order to test the equality of L functions.

We apply the methods to the same ChIP-seq comparative epigenomic profiling of adipogenesis of murine 3T3-L1 cells data as in Chapter 3. Our method detects many genes with differential H3K27ac levels at gene promoter regions across day -2, 0, 2, and day 7, which agree with what were observed in Mikkelsen et al. (2010). Furthermore, we compare these nonparametric test statistics with the kernel-based statistics in differential enrichment analysis and in associating the differential enrichment statistics to gene expression changes.

4.2. Two-sample Non-parametric Tests

In Chapter 3, we developed a kernel-based nonparametric procedure to identify the genes with differential enrichment regions between two conditions. To eliminate possible false positives due to small local changes, we used a relative large bandwidth $\lambda = 20/280$ in the analysis to over-smooth the data. Alternatively, we propose to apply non-parametric tests that do not require kernel smoothing. We consider two such tests: one assumes homoscedastic error variances, another allows heteroscedasticity

in error variances.

We use a similar pre-processing normalization method as in Chapter 3. For each gene i and each condition j, after square root transformation, the data $X^*_{ij}(t_k), k = 1, \cdots, n$ are approximately normal variables with mean $2\sqrt{\lambda_{ij}(t_k)}$ and variance of 1. For each condition j, $j = 1, 2$, we assume the data follow "signal + noise" model as (4.1) (omitting index i),

$$X_1(t_k) = f_1(t_k) + \sigma_1(t_k)W_1(t_k),$$
$$X_2(t_k) = f_2(t_k) + \sigma_2(t_k)W_2(t_k).$$

This model does not make parametric distributional assumptions on the noises. It only requires the errors to be symmetric around 0 and to have finite, twice-differentiable variance functions, $\sigma_1(t_k)$ and $\sigma_2(t_k)$.

For a given gene, the null hypothesis of interest is

$$H_0 : f_1(t) = f_2(t),$$

which can also be written as

$$H_0 : TS = ||f_1(t) - f_2(t)||^2 = 0. \tag{4.3}$$

However, due to issues related to data normalization and noises, in real ChIP-seq applications, we are more interested in testing the null hypothesis

$$H_0 : ||f_1(t) - f_2(t)||^2 = c \text{ vs } H_a : ||f_1(t) - f_2(t)||^2 > c,$$

for some biologically meaningful null value c, which represents the minimal difference between the two functions, $||f_1(t) - f_2(t)||^2$. For example, if input ChIP-seq data are available, we can estimate c for each gene based on input data. Alternatively, we can treat the genes with only very small tag counts as the null genes and use these null genes to estimate the c value.

4.2.1. Non-parametric tests of Munk and Dette (1998)

Munk and Dette (1998) proposed the following test statistic for the null hypothesis (4.3),

$$\hat{TS} = T_{diff} = \frac{\sum_{k=1}^{n-1}(X_1(t_k) - X_2(t_k)) \times (X_1(t_{k+1}) - X_2(t_{k+1}))}{(n-1)}, \tag{4.4}$$

where the expectation of T_{diff} is given by

$$TS = E(T_{diff}) = ||f_1 - f_2||^2 = \int (f_1(t) - f_2(t))^2 dt, \tag{4.5}$$

and the variance of T_{diff} is given by

$$\begin{aligned} Var(T_{diff}) &= \frac{||\sigma_1^2||^2 + ||\sigma_2^2||^2 + 2||\sigma_1\sigma_2||^2 + 4||(f_1 - f_2)\sigma_1||^2 + 4||(f_1 - f_2)\sigma_2||^2}{n-1} \\ &= \frac{1}{n-1} \int (\sigma_1^2(t) + \sigma_2^2(t))^2 + 4(f_1 - f_2)^2(\sigma_1^2(t) + \sigma_2^2(t))dt. \end{aligned} \tag{4.6}$$

We discuss the variance estimation in detail in Sections 4.2.2 and 4.2.3. Following the central limit theorem, we can define the new test statistics as

$$Z_{diff} = \frac{T_{diff} - E(T_{diff})}{\sqrt{Var(T_{diff})}} \tag{4.7}$$

which follows $N(0, 1)$ under the null hypothesis as $n \to \infty$. We then reject the null

hypothesis (4.3) if $Z_{diff} > z_{1-\alpha}$, where $z_{1-\alpha}$ is the $(1-\alpha)$ percentile of $N(0,1)$, which should be adjusted to account for multiple testing in applications.

4.2.2. Variance estimation under homogeneous variance assumption

Under the homoscedastic error variance assumption,

$$\sigma_1(t) = \sigma_1 \ , \ \sigma_2(t) = \sigma_2$$

σ_1 and σ_2 are not always the same and need to be estimated separately. For notational simplicity, we omit the subscript j in the following discussion and only show how to estimate the variance σ^2. If $f(t) = 0$, the most common estimator would be the sample standard deviation. We use the estimator proposed by Rice (1984) to estimate the noise variance for nonparametric regression,

$$\hat{\sigma}_j^2 = \frac{1}{2(n-1)} \sum_{k=2}^{n} (X_j(t_k) - X_j(t_{k-1}))^2. \tag{4.8}$$

Plugging (4.8) into (4.6), and based on (4.4) and (4.5), the estimation of variance of the test statistic under the equal variance assumption is

$$\begin{aligned} \hat{\sigma}_{eql}^2 &= \frac{(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)^2 + 4||f_1 - f_2||^2(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)}{n-1} \\ &\xrightarrow{Slutsky} \frac{(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)^2 + 4\hat{TS}(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)}{n-1}. \end{aligned} \tag{4.9}$$

4.2.3. Variance estimation under heterogeneous variance assumption

In the real applications, the variance may not always satisfy homoscedasticity assumption and may change as a function of the mean values. Munk and Dette (1998)

proposed the following variance estimate under unequal variance assumption,

$$||\hat{\sigma}_j^2||^2 = \frac{1}{4(n-3)} \sum_{k=2}^{n-2} (X_j(t_k) - X_j(t_{k-1}))^2 (X_j(t_{k+2}) - X_j(t_{k+1}))^2,$$

$$||(f_1 - f_2)\hat{\sigma}_j||^2 = \frac{1}{2(n-3)} \sum_{k=2}^{n-2} (X_1(t_{k-1}) - X_2(t_{k-1}))(X_1(t_k) - X_2(t_k))(X_j(t_{k+2}) - X_j(t_{k+1}))^2,$$

$$||\sigma_1 \hat{\sigma}_2||^2 = \frac{1}{4(n-1)} \sum_{k=2}^{n} (X_1(t_k) - X_1(t_{k-1}))^2 (X_2(t_k) - X_2(t_{k-1}))^2.$$

We then obtain the following variance estimate of the test statistic,

$$\begin{aligned} \hat{\sigma}_{unv}^2 &= \frac{1}{4(n-3)} \sum_{k=2}^{n-2} \sum_{j=1}^{2} (X_j(t_k) - X_j(t_{k-1}))^2 (X_j(t_{k+2}) - X_j(t_{k+1}))^2 \\ &+ \frac{1}{2(n-1)} \sum_{k=2}^{n} (X_1(t_k) - X_1(t_{k-1}))^2 (X_2(t_k) - X_2(t_{k-1}))^2 \\ &+ \frac{2}{(n-3)} \sum_{k=2}^{n-2} \sum_{j=1}^{2} (X_1(t_{k-1}) - X_2(t_{k-1}))(X_1(t_k) - X_2(t_k))(X_j(t_{k+2}) - X_j(t_{k+1}))^2. \end{aligned} \tag{4.10}$$

In real applications, for the genes that are not enriched by the histone under the study, we often observe data with very small or even zero counts, in which case the variance estimation $\hat{\sigma}^2$ can be too small, which can lead to identifying biologically uninteresting genes. To modulate this effect, we add a small constant $a_0 = 90\%$ percentile of the estimated standard standard deviations to each of the estimated standard deviations. This variance modulation has also been used in the variance estimation of kernel-smoothing based method.

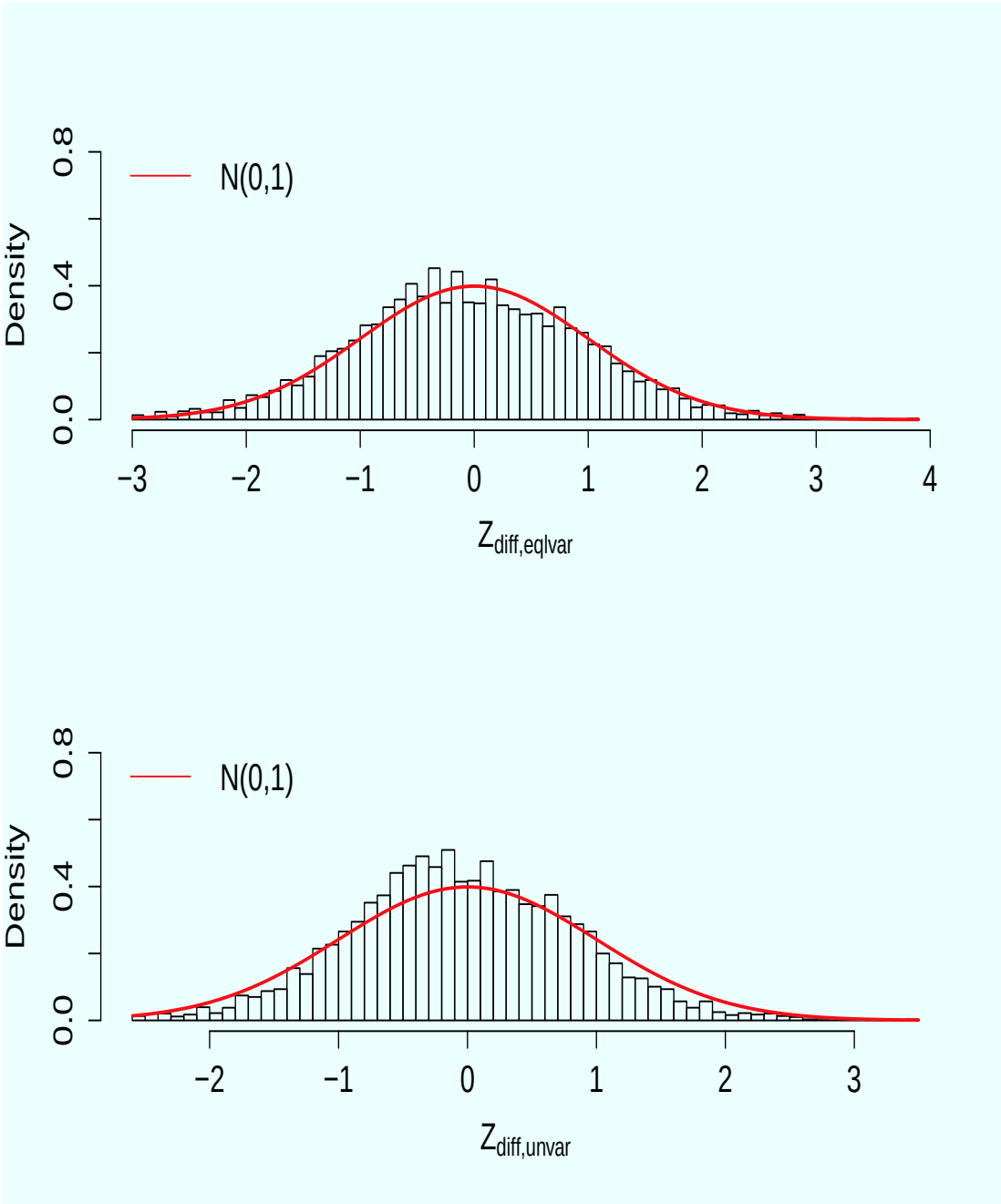


Figure 4.1: Histograms of the two test statistics, (a)$Z_{diff,\ eqlvar}$ and (b) $Z_{diff,\ unvar}$, for 9874 genes with maximum number of read count in both day -2 and day 7 fewer than 5 in the mouse adipogenesis ChIP-seq data. The red curve in each plot represents the standard normal density.

4.3. Application to ChIP-seq Study During Mouse Adipogenesis

4.3.1. Null distribution of the test statistics

We apply these two nonparametric tests to the same comparative ChIP-seq data of H3K27ac mark between day -2 and day 7 as in Section 3.4. In order to determine the biologically relevant value c in the null hypothesis $H_0 : TS = c$, we treat the same set of 9874 genes with read counts fewer than 5 as the null genes. We calculate $\hat{T}_{i,diff}$ for the ith null gene, $i = 1, ..., C = 9874$ and then take the mean of $\hat{T}_{i,diff}$ to obtain the value c,

$$c = \mathrm{E}(\hat{T}_{diff} | i \in \mathrm{NULL}) = \frac{1}{C} \sum_{i=1}^{C} \hat{T}_{i,diff}. \tag{4.11}$$

Here, $c \approx 0.78$ is the minimal distance of two signal functions between day 7 and day -2 for H3K27ac. In Figure 4.1, we present the histograms of two new statistics $Z_{diff,\ eqlvar}$ and $Z_{diff,\ unvar}$ by testing $H_0 : TS = c$ for these null genes. Clearly, both of the statistics are close to standard normal distribution, which is very similar to Figure 3.1 for $Z_{0\lambda,WH}$. Therefore, the null distributions of these two test statistics are reasonable.

4.3.2. Comparison of different test statistics

Figure 4.2 (a) shows that the two test statistics $Z_{diff,\ eqlvar}$ and $Z_{diff,\ unvar}$ are almost identical. Since $Z_{diff,\ unvar}$ requires fewer assumptions, we use this to represent the Z_{diff} in the following discussion unless otherwise noted. Thus we only check the plots of $Z_{diff,\ unvar}$ versus $Z_{\lambda,WH}$ and fold-change statistics. Figure 4.2 (b) shows that $Z_{diff,\ unvar}$ is positively correlated with $Z_{\lambda,WH}$ and plots (c) show that between $Z_{diff,\ unvar}$ and fold-change statistics have very similar pattern as in Figure 3.3. In

general, large Z_{diff} values correspond to large fold-changes or large $Z_{\lambda,WH}$. The top 12 genes with largest test statistics identified by $Z_{\lambda,WH}$ and Z_{diff} are almost the same set of genes with differentially enriched regions, which show either peak shift, enrichment intensity difference or peak/no peak (results not shown).

Table 4.1 compares the numbers of genes with differential enrichment (DE) regions identified using different statistics at the Bonferonni adjusted p-value of 0.05. For the 9874 null genes with the maximum number of read counts in both day -2 and day 7 fewer than 5, DBChIP identifies many of these genes as differentially enriched genes. The number of genes with DE regions identified by DBChIP also heavily depends on the threshold c used.

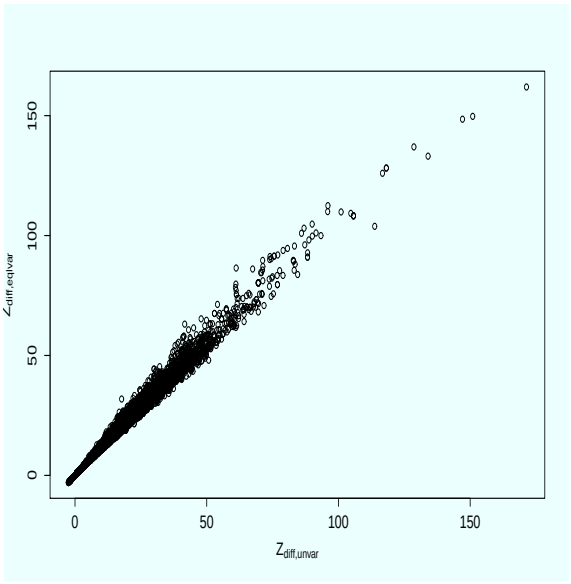
Table 4.1: Numbers of genes with DE regions identified by different tests in the mouse adipogenesis ChIP-seq data, including $Z_{0\lambda,WH}$, $Z_{diff,unequal}$ and DBChIP test with fold change value c=1.5 (default), c=1 and c=2 (max). 9874 Null genes: genes with the maximum number of read counts in both day -2 and day 7 fewer than 5.

			DBChIP		
	$Z_{0\lambda,WH}$	$Z_{diff,unequal}$	$c = 1.5$	$c = 1$	$c = 2$
Null genes	3	0	888	2707	399
DB genes	10,467	13,677	6918	17206	3597

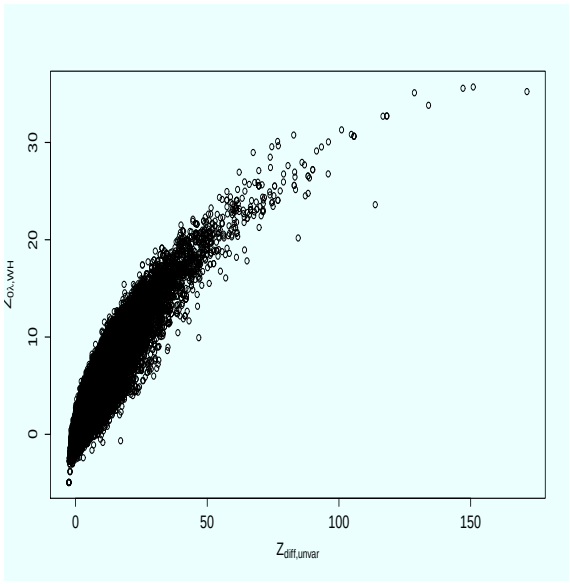
4.3.3. Correlation between ChIP-seq differential enrichment statistics and gene expression fold-changes

We next compare how different test statistics for differential enrichment are correlated with gene expression fold changes between the two time points. We define for gene k, $\Delta_k = 1$ if the kth gene has a more than 2^δ fold change between the two time points,

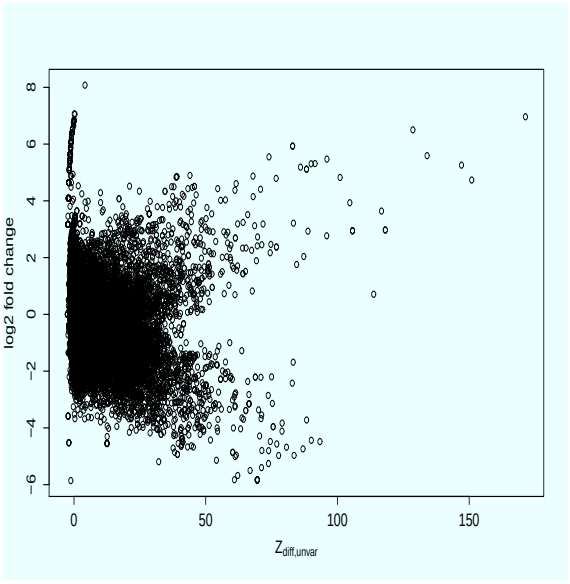
$$\Delta_k = I\{|\log_2 \frac{W_{k1}}{W_{k2}}| > \delta\}, \tag{4.12}$$



(a)



(b)



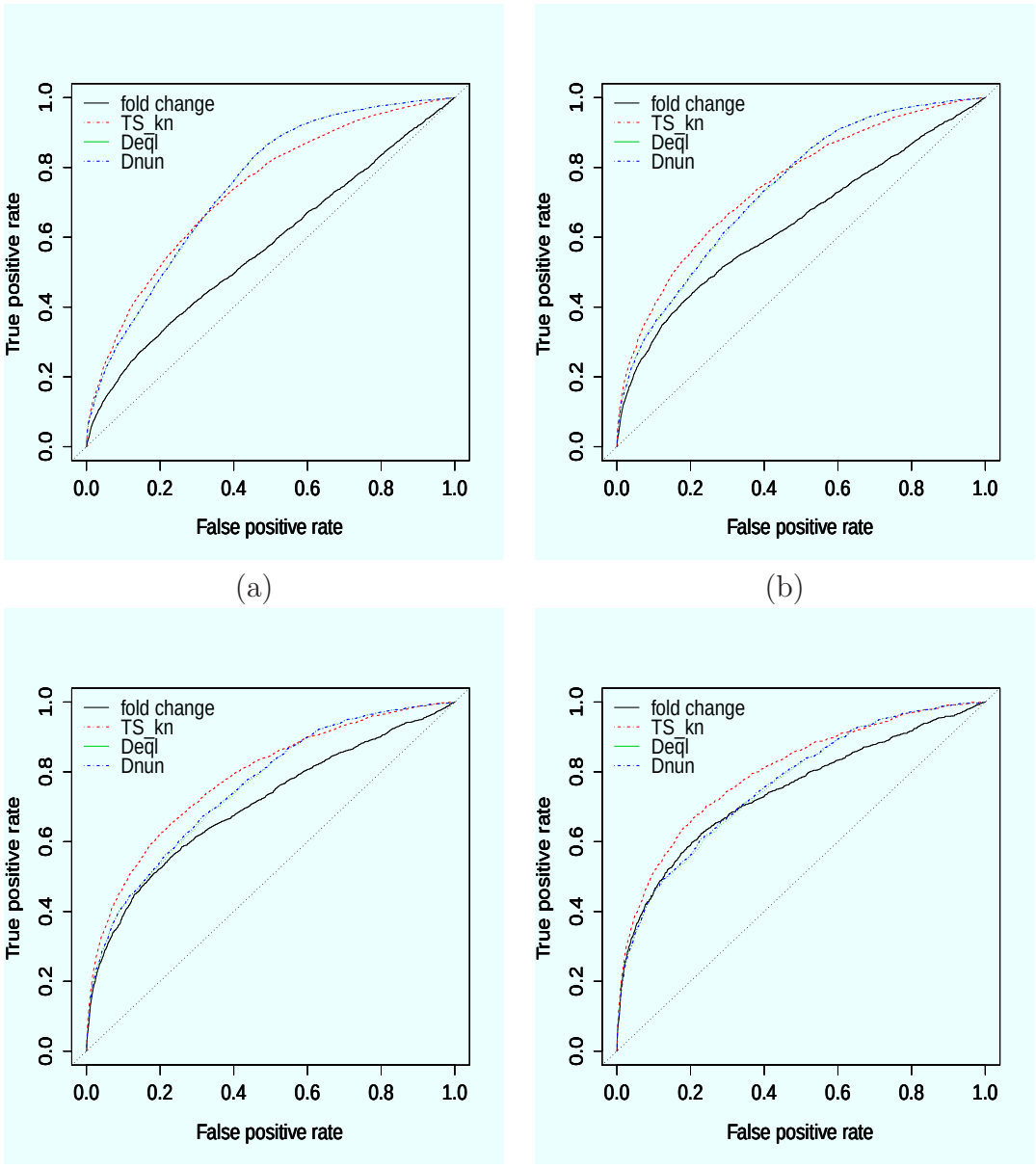
(c)

Figure 4.2: Comparison of different statistics for the mouse adipogenesis ChIP-seq data: (a) the proposed statistics with unequal variance estimation vs the statistics with equal variance estimation; (b) the proposed statistics with unequal variance estimation and the kernel-smoothing based statistics; (c) the proposed statistics with unequal variance estimation and the fold-change statistics.

where $\delta = 1$ represents at least two-fold change in gene expression. Figure 4.3 shows the ROC curves for gene expression changes using different statistics and different cutoff values. We observed that both $Z_{\lambda,WH}$ and Z_{diff} outperform the fold-change statistics in the ROC curves for different cutoff value δ from 0.5, 1, 1.5 to 2, especially when a smaller cutoff value is used.

Finally, we also calculate the proportions of true differentially expressed genes among the top 100 to 10000 genes selected by each method as the True Positive Rate (TPR). The definition of true differentially expressed gene is the same as we used in (4.12). In Figure 4.4, we observe that $Z_{\lambda,WH}$ shows a much higher proportions of true positives among the very top genes (100-2000) than the other methods. The performance between $Z_{diff,\ eqlvar}$ and $Z_{diff,\ unvar}$ are almost the same, which is reasonable since we use the rank list to make TPR plots and the two statistics only differ in the variance estimates. The TPR of fold-change statistics is much lower than other methods, especially for the top 500 genes, where the fold change statistics have very low TPR due to the small counts in both experimental conditions. This results are consistent with what we observed in Section 3.4 that a large proportion of top-ranked genes selected by fold-change is probably false. As a comparison, we found a similar pattern with different cutoff values between these test statistics as in Figure 4.3.

These results indicate that both the kernel-based statistics and nonparametric statistics with equal variance or unequal variances correspond to gene expression changes very well. As a comparison, kernel based methods with a fixed bandwidth have a slightly better performance. We also observe that the fold-change statistics perform the worst in detecting the genes with real differential histone enrichments.



(a) (b)

(c) (d)

Figure 4.3: Plots of ROC curves for gene expression fold changes (2^{δ}) using four test statistics for different fold-change cutoff values: (a) $\delta = 0.5$, (b) δ=1.0, (c) δ=1.5, (d) δ=2.

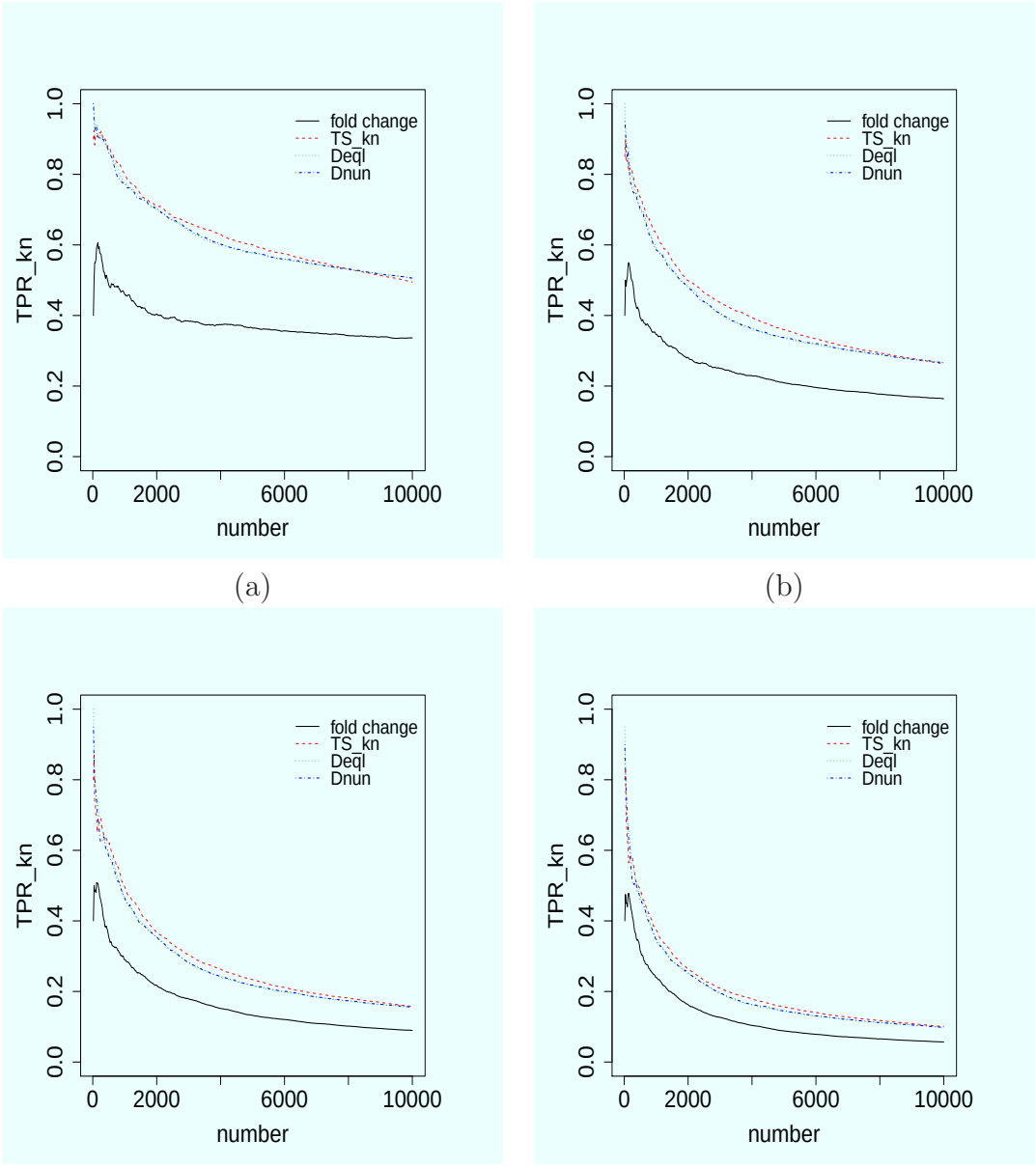


Figure 4.4: Plots of true positive rate curves for gene expression fold changes (2^{δ}) using four test statistics for different cutoff values: (a) $\delta = 0.5$, (b) δ=1.0, (c) δ=1.5, (d) δ=2.

4.3.4. Application to an ENCODE ChIP-seq Study with two replicates

To further demonstrate these nonparametric tests in term of false discovery, we analyze the ChIP-seq data reported in ENCODE Project Consortium et al. (2012) for two cell lines of human GM12878, B-lymphoblastoid cell and HeLa-S3, cervical carcinoma cell. Our analysis still focuses on H3K27ac mark at promoter regions of genes with $n = 280$ bins. There is a total of $m = 23807$ transcripts that can be mapped to 23807 genes. The data set includes two replicates for ChIP-seq data and two replicates for input data. Biologically, we should not expect many genes with differential enrichment between the two replicates.

We first compare the ChIP-seq profiles of the two input replicates where we calculate the test statistics $Z_{all,uneql}$ for each of the genes i, $i = 1, ..., m = 23807$. Based on (4.11), we obtain $\hat{T}_{diff}$ for 20300 genes with maximum value less than 5 and take the mean of $\hat{T}_{diff}$ as the value c. The histogram of $Z_{all,uneql}$ for all m genes in Figure 4.5 shows that the test statistics roughly follow the standard normal distribution. In addition, using a Bonferroni adjusted p-value of 0.05, our procedure only identifies 9 genes with test statistics greater than the threshold, which results in a less than 0.025 % false discovery rate. This example further demonstrates that our proposed nonparametric testing procedure is not only powerful enough to detect the true differential enrichment regions but also makes only a few false detections.

We also apply the nonparametric test to the two ChIP-seq replicates and calculate test statistics $Z_{all,uneql}$ for each gene i, $i = 1, ..., m = 23807$. We first normalize the reads counts data using a Poisson sampling (Li and Tibshirani, 2011), and then take $2 * \sqrt{\text{count} + 1/4}$ for the normalized data and also for the input ChIP-seq data. We then subtract the input counts from the ChIP-seq counts and calculate the test

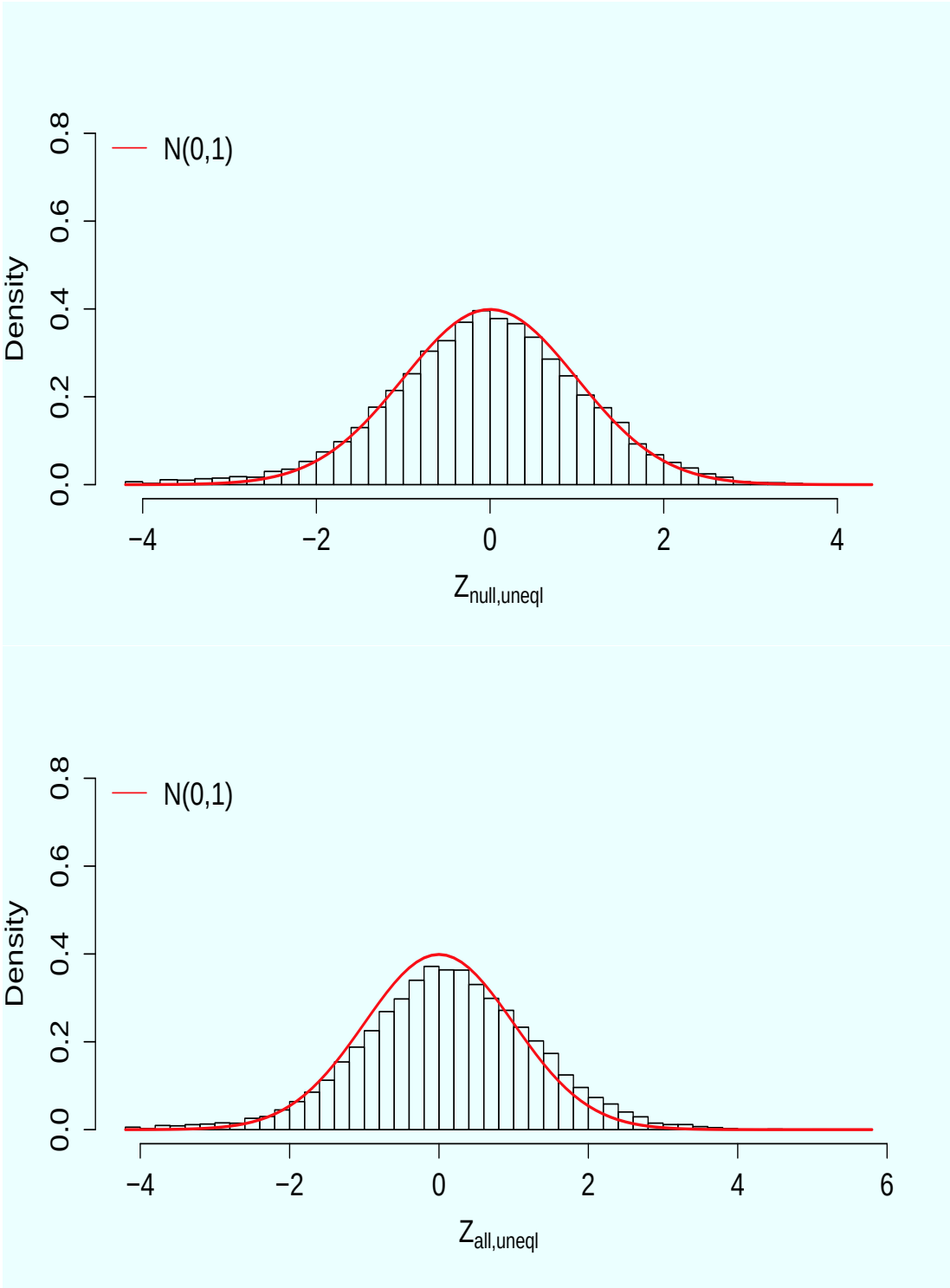


Figure 4.5: Comparison between two replicated ENCODE input data sets. Top: Histogram of test statistics $Z_{null,uneql}$ for 20300 NULL genes in the new data sets. Bottom: Histogram of test statistics $Z_{all,uneql}$ for all 23807 genes in the new data sets. The red curve represents the standard normal density.

statistic $Z_{all,uneql}$ for each gene. Figure 4.6 shows the histograms of the test statistics for genes with fewer than 5 read counts and for all the genes, which closely follow the standard normal distribution. We therefore should not expect many significant differentially enriched genes.

Table 4.2 shows the number of genes with DE regions identified by the different tests for several different comparisons. In general, we observe that both kernel smoothing and the nonparametric tests give a small number of false positives when we compare two the replicates of the GM12878 cell lines, or the two replicate of the input ChIP-seq data. The proposed procedures also give a very small number of false positives when we compare the genes that have only small number of counts. In contrast, results of the DBChIP test greatly depend on the threshold c used. For a small threshold, we observe many false positives. However, when the threshold is set too large, the test loses power to detect genes with differentially enriched regions.

Table 4.2: Numbers of genes with DE regions identified for the ENCODE data sets using different test statistics, $Z_{0\lambda,WH}$, $Z_{diff,unequal}$ and DBChIP test with allowable fold change value $c = 1.5$ (default), $c = 1$ and $c = 2$ (max). Four different comparisons are performed: (a) two GM12878 ChIP replicates; (b) two GM12878 Input replicates; (c) 9124 Null genes with maximum number of read count in both GM12878 and HeLa-S3 cell lines fewer than 5; (d) Two cell lines GM12878 and HeLa-S3.

			DBChIP		
	$Z_{0\lambda,WH}$	$Z_{diff,unequal}$	$c = 1.5$	$c = 1$	$c = 2$
GM12878 ChIP replicates	263	134	0	529	0
GM12878 Input replicates	11	9	0	222	0
9124 Null genes	23	14	2	333	0
GM12878 vs HeLa-S3	6647	7691	2202	7444	1074

4.3.5. A simulation comparison

We present Monte Carlo simulation to evaluate the performance of the different two-sample statistics, $Z_{\lambda,WH}$, $Z_{diff,\ eqlvar}$ and $Z_{diff,\ unvar}$. We demonstrate that these methods can control the Type I error at the desired levels, when we simulate data

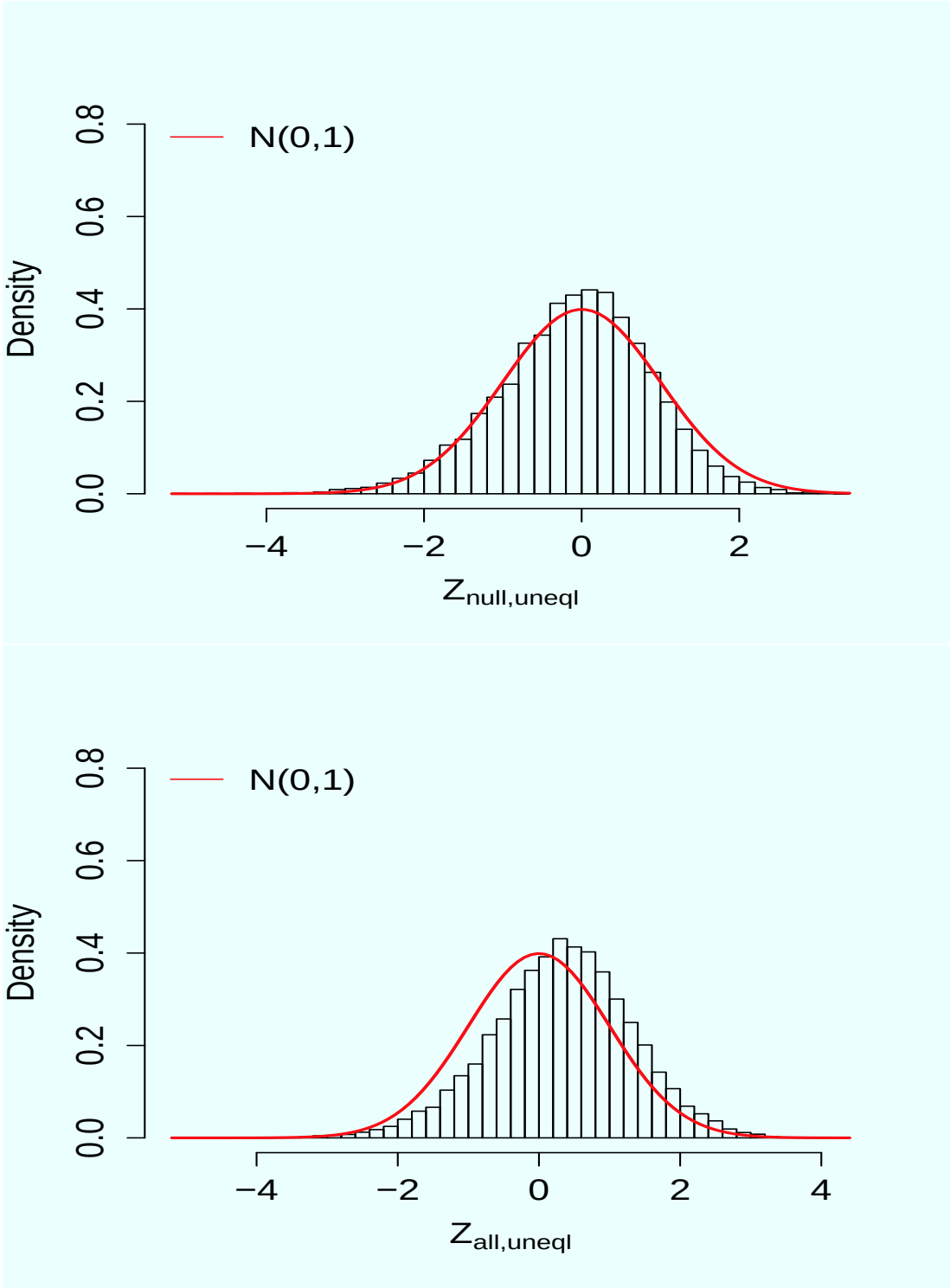


Figure 4.6: Histogram of test statistics $Z_{all,uneql}$ for all 23807 genes in the ENCODE data sets. The red curve represents the standard normal density.

under the null where signal functions are the same $f_1 = f_{N(0,1)}(t)$ and $f_2 = f_{N(0,1)}(t)$. We simulate the spatial histone enrichment profiles by a normal density function, $f_{N(\mu,s)}$ with mean μ and variance σ^2, where different parameters μ and s represent different differential enrichment profiles. The errors are still simulated from $N(0, \sigma^2(t))$. We consider three cases for the variance functions: (a) homoscedasticity and equal variance $\sigma_1^2 = \sigma_2^2 = 1$; (b) homoscedasticity and unequal variance $\sigma_1^2 = 0.5$, $\sigma_2^2 = 1$; (c) heteroscedastic variances $\sigma_1^2(t) = \sigma_2^2(t) = sin(t)$.

We simulate data for $m = 10,000$ genes and let the sample size for each gene be $n = 280$. The results are shown in Table 4.3. For simulations (a) and (b) under homoscedasticity assumption, all three statistics can control the type-I error reasonably well, where kernel-smoothing based methods show a slight inflation that may be due to the fixed bandwidth. The nonparametric tests are on the other hand slightly conservative since they do not make any assumptions on errors. For simulation (c) where we have heteroscedastic variances, only $Z_{diff,\ unvar}$ controls the type I error at the specified level. The other two methods have inflated errors, which implies that $Z_{diff,\ unvar}$ is robust and stable if the variance function is not always a constant across different bins. In real applications, we recommend using the test statistic $Z_{diff,\ unvar}$ for differential enrichment analysis.

Table 4.3: Simulation to evaluate the type 1 errors of three different two-sample test statistics under three different settings: (a) homoscedasticity and equal variance $\sigma_1^2 = \sigma_2^2 = 1$; (b) homoscedasticity and unequal variance $\sigma_1^2 = 0.5$, $\sigma_2^2 = 1$; (c) heteroscedastic variances, with $\sigma_1^2(t) = \sigma_2^2 = sin(t)$.

	$Z_{\lambda,WH}$		$Z_{diff,eql}$		$Z_{diff,uneql}$	
	0.05	0.01	0.05	0.01	0.05	0.01
(a)	0.0517	0.0122	0.0374	0.0034	0.0411	0.0049
(b)	0.0509	0.0146	0.0393	0.0046	0.0429	0.0062
(c)	0.0840	0.0357	0.0865	0.0212	0.0434	0.0080

4.4. Extension to Time-Course ChIP-seq Data

The two-sample nonparametric tests presented in Section 4.2 can also be extended to multiple-sample cases, in particularly to time-course ChIP-seq data. The ChIP-seq data sets (Mikkelsen et al., 2010) we analyzed include H3K27ac marks at four different time points, including proliferating (day -2) and confluent (day 0) preadipocytes, immature adipocytes (day 2) and mature adipocytes (day 7). Let X_{ikj} denote observed read counts X_{ikj} in bin k under condition j, for $i = 1, \cdots, m$, $k = 1, \cdots, n$ and $j = 1, 2, 3, L = 4$. The goal is to test the equality of functions among these 4 time points. For each gene, we assume the data follow a "signal + noise" model (omitting k),

$$X_j(t) = f_j(t) + W_j(t).$$

The null hypothesis of interest is

$$H_0 : f_1(t) = f_2(t) = f_3(t) = f_4(t) = f(t). \tag{4.13}$$

4.4.1. TSmax and TSmean test statistics

For the time-course ChIP-seq data with L time points, we are interested in changes of histone modification enrichment states between $L - 1$ adjacent time points. Let $TS_{L(L-1)}$, $\cdots$,TS_{21} be the pair-wise test statistics between two neighbouring time points. Based on (3.9) and (4.7), we know that under the $H_0 : f_i(t) = f_j(t)$, the two-sample statistic follows a $N(0, 1)$. Thus, under the global null hypothesis (4.13), the joint distribution of $TS = (TS_{L(L-1)},...,TS_{21})^T$ follows a multivariate normal distribution $N_{L-1}(\mu, \Sigma)$. Note that $TS_{(j+2)(j+1)}$ is independent with $TS_{j(j-1)}$, so $\rho_{ij} =$

0, for all $j > i + 1$.

$$\begin{pmatrix} TS_{12} \\ \vdots \\ TS_{(L-1)L} \end{pmatrix} \xrightarrow{H_0} N_{L-1} \left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{12} & 0 & \cdots & 0 \\ \rho_{12} & 1 & \rho_{23} & \ddots & \vdots \\ 0 & \rho_{23} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 1 \end{pmatrix} \right). \tag{4.14}$$

Based on this joint null distribution of the pair-wise test statistics, we use the mean of TS (TS_{mean}) or the maximum of TS (TS_{max}) as the test statistics for the hypothesis (4.13). Intuitively, TS_{max} is expected to perform better than TS_{mean} since a larger value of the positive test statistic represents differential enrichment. However, the signal of TS_{mean} can be diluted by taking the average of the pair-wise test statistics when some pair-wise statistics are negative.

4.4.2. Distribution of TS_{mean}

Based on the joint distribution of the pair-wise test statistics given in (4.14), the distribution of $TS_{mean} = \frac{1}{L-1} \sum_{j=1}^{L-1} TS_{(j+1)j}$ follows $N(0, \sigma^2_{mean})$, where

$$\sigma^2_{mean} = \frac{(L - 1 + 2 \sum_{i<j} \rho_{ij})}{(L-1)^2},$$

and therefore

$$Z_{mean} = \frac{TS_{mean}}{\sigma_{mean}} \xrightarrow{H_0} N(0, 1) \tag{4.15}$$

We discuss the estimation of ρ in Section 4.4.4 and the Appendix B.3.

4.4.3. Distribution of TS_{max}

The distribution of $TS_{max} = \max(TS_{jj+1}, j = 1, ..., L-1)$ is not simple since the pairwise statistics TS_{jj+1} are not independent. Arellano-Valle and Genton (2007, 2008) provided the exact distribution of the maximum of $X = (X_1, ..., X_n)^T \sim N_n(\mu, \Sigma)$. Let $X_{-i} = (X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n)^T$ for each i. We partition X by

$$X = \begin{pmatrix} X_{-i} \\ X_i \end{pmatrix}, \mu = \begin{pmatrix} \mu_{-i} \\ \mu_i \end{pmatrix},$$

and

$$\Sigma = \begin{pmatrix} \Sigma_{-i-i} & \Sigma_{-ii} \\ \Sigma_{i-i} & \Sigma_{ii} \end{pmatrix}.$$

Further we define $\mu_{-i,i} = \mu_{-i} + (x - \mu_i)\Sigma_{-ii}/\Sigma_{ii}$, and $\Sigma_{-i-i,i} = \Sigma_{-i-i} - \Sigma_{-ii}\Sigma_{-ii}^T/\Sigma_{ii}$. Using the general results of Arellano-Valle and Genton (2008) together with the covariance matrix Σ given in (4.14), the density function (PDF) of T_{max} can be written as

$$f_{max}(x) = \sum_{i=1}^{n} \phi(x)\Phi_{L-2}(X_{(L-2)\times 1}; \mu_{-i,i}, \Sigma_{-i-i,i}), \tag{4.16}$$

where $\phi(x)$ is the PDF of $N(0, 1)$, Φ_{L-2} is the cumulative distribution function (CDF) of multivariate normal distribution.

Jamalizadeh and Balakrishnan (2009) further provided the moment generating function of the maximum of a trivariate normal distribution and define this distribution as a sum of weighted generalized skew-normal (WGSN) distributions. Consider the

setting where we have L=4 time points, we have

$$TS = (TS_{12}, TS_{23}, TS_{34})^T \sim N_3 \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix} \right)$$

where $\rho_{13} = 0$. The PDF of TS_{max} is given by

$$f_{max}(t) = \sum_{i=1}^{3} \phi(t)\Phi_2(X_{a_i t, b_i t; \rho_i}), \tag{4.17}$$

where $\Phi_2(X_{a_i t, b_i t; \rho_i})$ is the CDF of bivariate normal distribution with correlation coefficient ρ_i, and

$$\begin{aligned}
a &= (a_1, a_2, a_3) = \left(\frac{1-\rho_{12}}{\sqrt{1-\rho_{12}^2}}, \frac{1-\rho_{12}}{\sqrt{1-\rho_{12}^2}}, \frac{1-\rho_{13}}{\sqrt{1-\rho_{13}^2}} = 1\right), \\
b &= (b_1, b_2, b_3) = \left(\frac{1-\rho_{13}}{\sqrt{1-\rho_{13}^2}} = 1, \frac{1-\rho_{23}}{\sqrt{1-\rho_{23}^2}}, \frac{1-\rho_{23}}{\sqrt{1-\rho_{23}^2}}\right), \\
\rho &= (\rho_1, \rho_2, \rho_3) = \left(\frac{\rho_{23}-\rho_{12}\rho_{13}}{\sqrt{1-\rho_{12}^2}\sqrt{1-\rho_{13}^2}}, \frac{\rho_{13}-\rho_{12}\rho_{23}}{\sqrt{1-\rho_{12}^2}\sqrt{1-\rho_{23}^2}}, \frac{\rho_{12}-\rho_{13}\rho_{23}}{\sqrt{1-\rho_{13}^2}\sqrt{1-\rho_{23}^2}}\right).
\end{aligned}$$

Further, the expectation and variance of TS_{max} is given by Jamalizadeh and Balakrishnan (2009)

$$E(TS_{max}) = \frac{1}{2\sqrt{\pi}}(\sqrt{1-\rho_{12}} + \sqrt{1-\rho_{13}} + \sqrt{1-\rho_{23}}), \tag{4.18}$$

$$\mathrm{Var}(TS_{max}) = 1 + \frac{\sqrt{M}}{2\pi} - E(TS_{max})^2 \tag{4.19}$$

where $M = 6 + 2(\rho_{12}\rho_{13} + \rho_{12}\rho_{23} + \rho_{23}\rho_{13}) - (1+\rho_{12})^2 - (1+\rho_{13})^2 - (1+\rho_{23})^2$.

To verify these results, we simulate 10,000 multivariate normal distributed samples

for $TS = (TS_{12}, TS_{23}, TS_{34})^T$ with $\mu = (0,0,0)^T$ and

$$\Sigma = \begin{pmatrix} 1 & 0.25 & 0 \\ 0.25 & 1 & 0.25 \\ 0 & 0.25 & 1 \end{pmatrix}.$$

We then calculate the test statistics $TS_{max} = max(TS_{12}, TS_{23}, TS_{34})$ for all 10,000 samples. Figure 4.7 shows the histogram of these statistics together with the fitted exact WGSN curve as the red line and normal density curve with mean (4.18) and variance (4.19) as the blue line. We observe that the WGSN curve is a little skewed to the left but both curves almost overlap in the tails. In real applications, we may use the normal distribution

$$Z_{max} = \frac{TS_{max} - E(TS_{max})}{\sqrt{Var(TS_{max})}} \xrightarrow{H_0} N(0,1) \tag{4.20}$$

to approximate the WGSN, especially at the tail.

4.4.4. Estimation of Covariance Matrix for Multiple-Sample Test Statistics

To calculate the statistics TS_{max} and TS_{mean}, we need the estimates of ρ_{12} and ρ_{23} ($\rho_{13} = 0$) in the covariance matrix Σ. Based on Munk and Dette (1998) (Appendix Lemma A3), $\hat{\rho}_{12}$ and $\hat{\rho}_{23}$ depend on the estimates of $\hat{f}_j\hat{f}_j$, and $\hat{f}_j\hat{f}_g$, $j = 1,..,L$, $g = 1,..,L$, and $g \neq j$. We present these estimates in the Appendix B.3 and B.4 under both equal variance and unequal variance assumptions. Here we only show the results of the estimates of ρ_{12} and ρ_{23} under the unequal variance assumption, which are given as

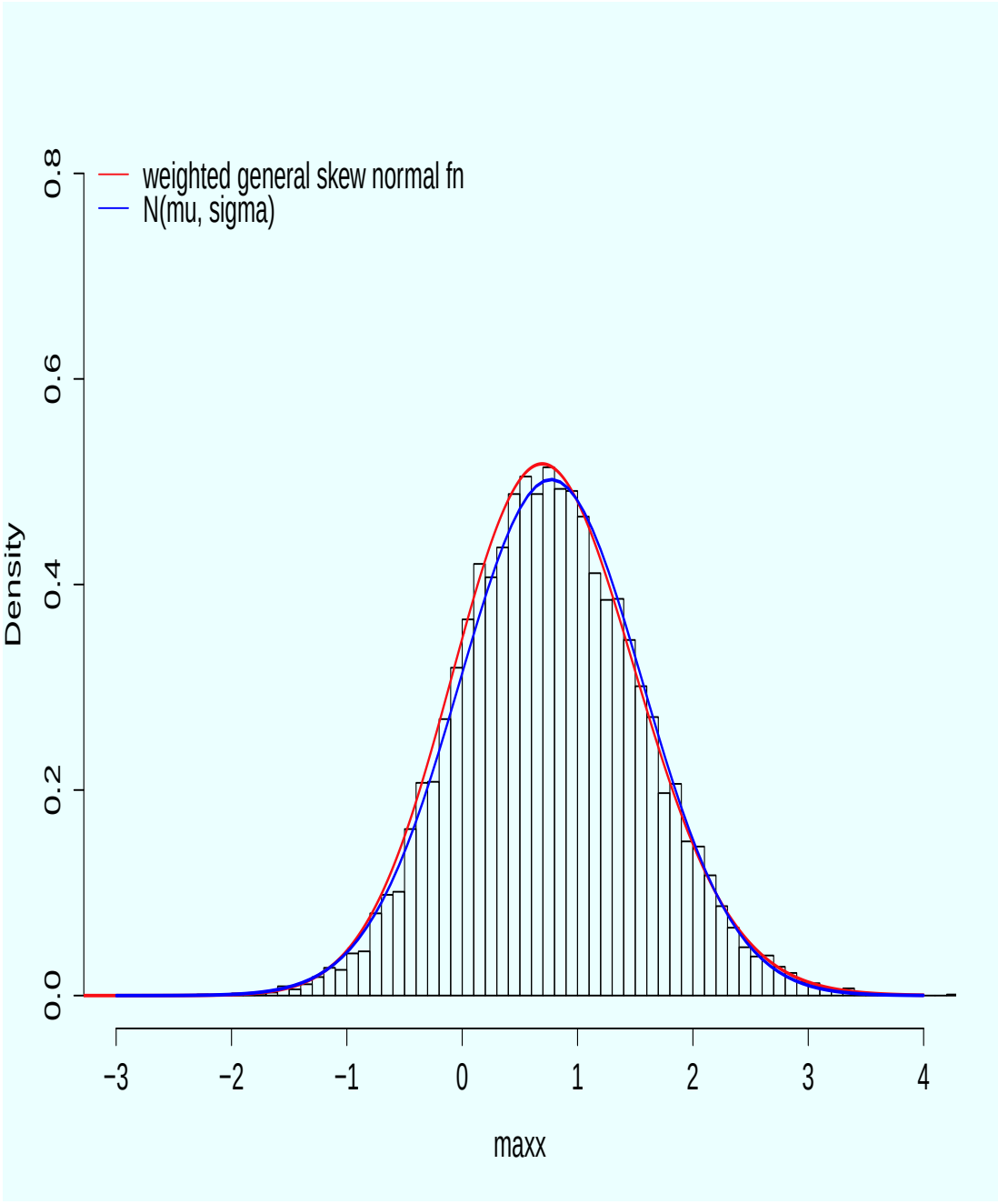


Figure 4.7: Histogram of the test statistics T_{max} for 10,000 samples simulated under the null multivariate normal distribution with WGSN and normal curves fitted.

$$\rho_{12}^{uneql} = \frac{||\sigma_2^2||^2 + 4(||\sigma_2 f_2||^2 - ||\sigma_2^2 f_1 f_2|| - ||\sigma_2^2 f_2 f_3|| + ||\sigma_2^2 f_1 f_3||)}{(n-1)\ \sigma_{1,2}^{unv}\ \sigma_{2,3}^{unv}}, \quad (4.21)$$

and

$$\rho_{23}^{uneql} = \frac{||\sigma_3^2||^2 + 4(||\sigma_3 f_3||^2 - ||\sigma_3^2 f_2 f_3|| - ||\sigma_3^2 f_3 f_4|| + ||\sigma_3^2 f_2 f_4||)}{(n-1)\ \sigma_{2,3}^{unv}\ \sigma_{3,4}^{unv}}. \quad (4.22)$$

Under the null hypothesis (4.13), we have

$$\rho_{12} \xrightarrow{H_0} \frac{||\sigma_2^2||^2}{\sqrt{||\sigma_1^2||^2 + ||\sigma_2^2||^2 + 2||\sigma_1\sigma_2||^2}\sqrt{||\sigma_2^2||^2 + ||\sigma_3^2||^2 + 2||\sigma_2\sigma_3||^2}},$$

$$\rho_{23} \xrightarrow{H_0} \frac{||\sigma_3^2||^2}{\sqrt{||\sigma_2^2||^2 + ||\sigma_3^2||^2 + 2||\sigma_2\sigma_3||^2}\sqrt{||\sigma_3^2||^2 + ||\sigma_4^2||^2 + 2||\sigma_3\sigma_4||^2}},$$

If we further assume constant variance $\sigma_j(t_k) = \sigma_j, j = 1, 2, 3, 4$ and $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$, we can obtain $\rho_{12} = \rho_{23} = \frac{1}{4}$. This value is used in the simulation presented in Figure 4.7.

4.5. Application to a Comparative Time Course ChIP-seq Study During Mouse Adiopogenesis

We apply the multi-sample test to the same ChIP-seq experiments data described in Section 4.3. Recall that there are $m = 29716$ genes and for each gene i, there are $n = 280$ observed read counts in bin k under condition j, for $i = 1, \cdots, m$, $k = 1, \cdots, n$ and $j = 1, 2, 3, L = 4$. For each gene, after the normal-transformation as in Section 3.2, we calculate the pair-wise statistics for each adjacent pair and their correlations and use mean or max of all the pair-wise statistics to identify genes with

differential enrichment during the time course experiment.

4.5.1. Comparison between TS_{max}, TS_{mean} for the ChIP-seq time-course experiments

We calculate the adjacent pair of test statistics TS_{12}, TS_{23} and TS_{34} for each gene. For each test, using the same procedure as in Section 4.2, we identify the genes with max counts in both conditions less than 5 and use these genes to estimate the c value, respectively. Here, the c value used in the test is $c_{12} = 0.805$, $c_{23} = 0.693$, and $c_{34} = 0.675$, respectively. The test statistics for these null genes should follow a $N(0, 1)$ distribution. Figure 4.8 shows the histograms of the statistics for these null genes, which are very close to the standard normal distribution.

To perform overall tests for differential enrichment over time, we calculate the statistics Z_{mean} and Z_{max} for each gene. Figure 4.8 also shows the histograms of these two statistics for the null genes, which closely follow the standard normal distribution.

Figure 4.9 and Figure 4.10 show the top 12 genes with largest test statistics by TS_{max} and TS_{mean}. Both plots show some genes with clear differences in ChIP-enriched profiles between all four time points and some genes are enriched in only one condition. For genes enriched at all four time points, the peak heights are very different. It seems that the top genes selected by TS_{max} show stronger differential enrichment among the four conditions than those identified by TS_{mean}.

4.5.2. Association with gene expression changes

We can similarly define test statistics based on pair-wise kernel-based statistics $Z_{\lambda,12}$, $Z_{\lambda,23}$ and $Z_{\lambda,34}$ between each pair of adjacent time points,

$$Z_{\lambda,max} = \max(Z_{\lambda,12},\ Z_{\lambda,23}, Z_{\lambda,34}), Z_{\lambda,mean} = \frac{Z_{\lambda,12} + Z_{\lambda,23} + Z_{\lambda,34}}{3}.$$

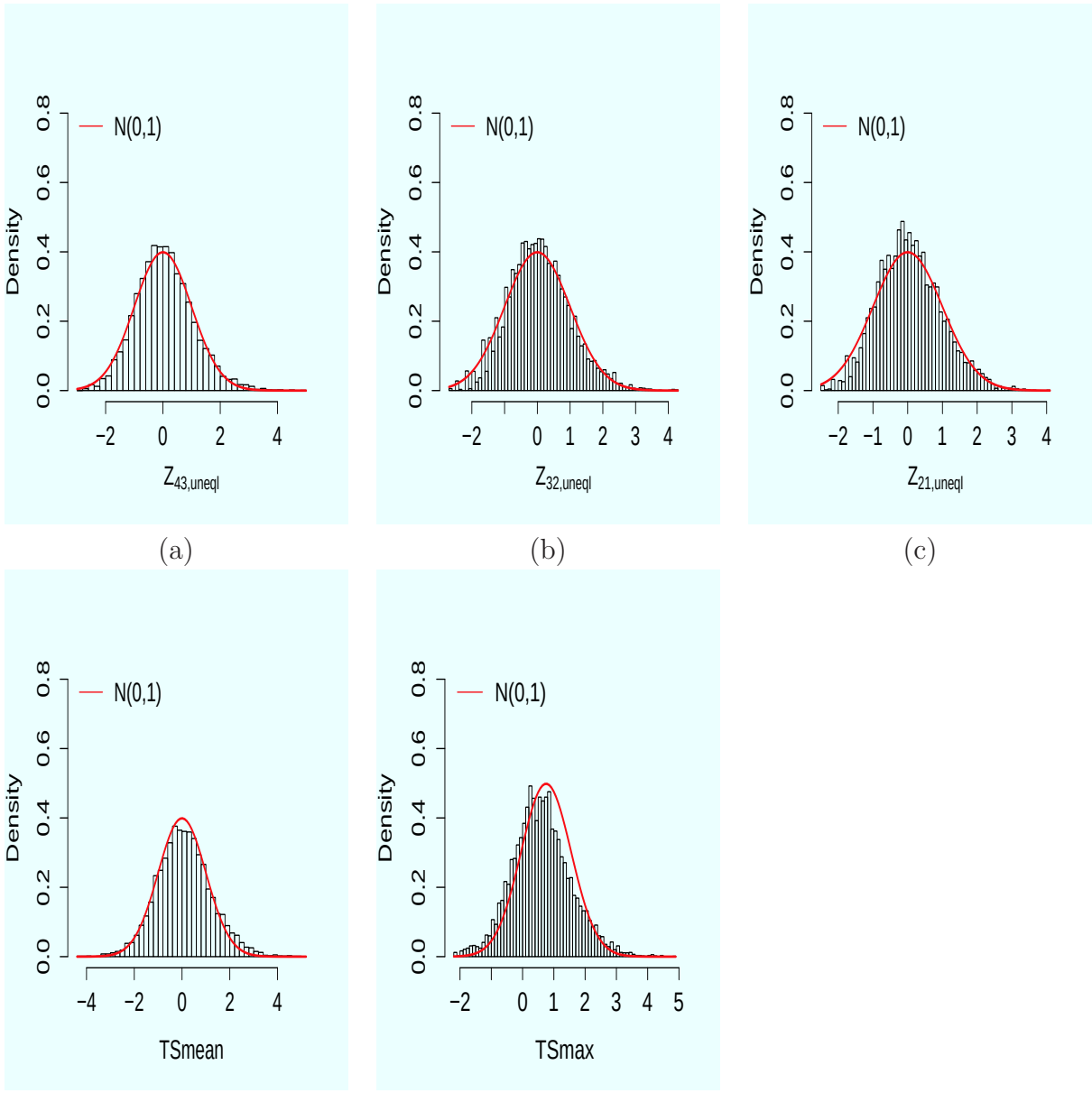


Figure 4.8: Histogram sof the test statistics for genes with reads counts fewer than 5 between each two adjacent time points and the overall test statistics. (a) TS_{43}; (b) TS_{32}; (c) TS_{21}; (d) TS_{mean}; (e) TS_{max}.

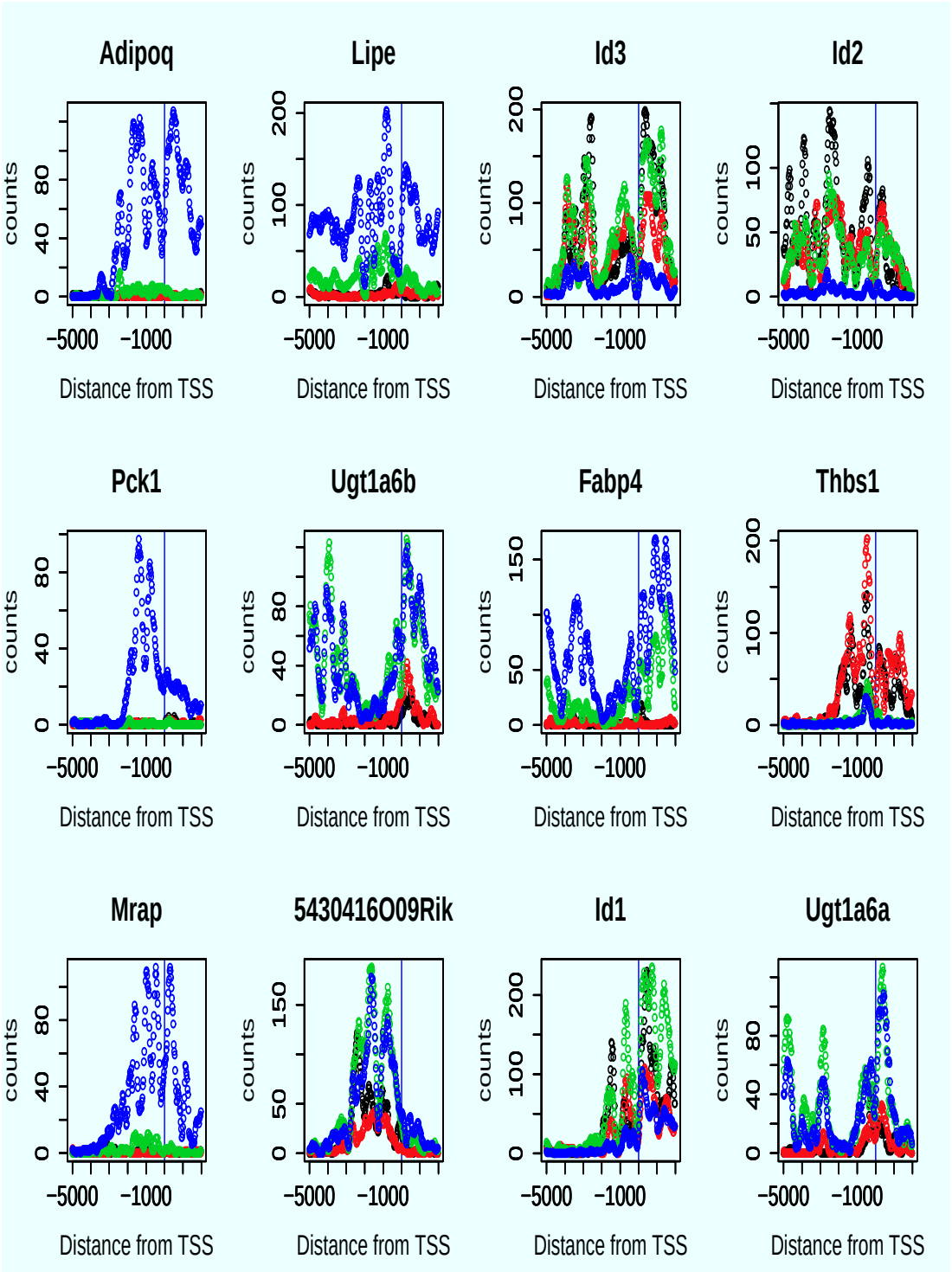


Figure 4.9: Observed ChIP-seq bin-counts for top twelve genes ranked by TS_{max} statistics over the promotor region for day -2 (black), day 0 (red), day 2 (green) and day 7 (blue). Vertical line represents the transcription starting site.

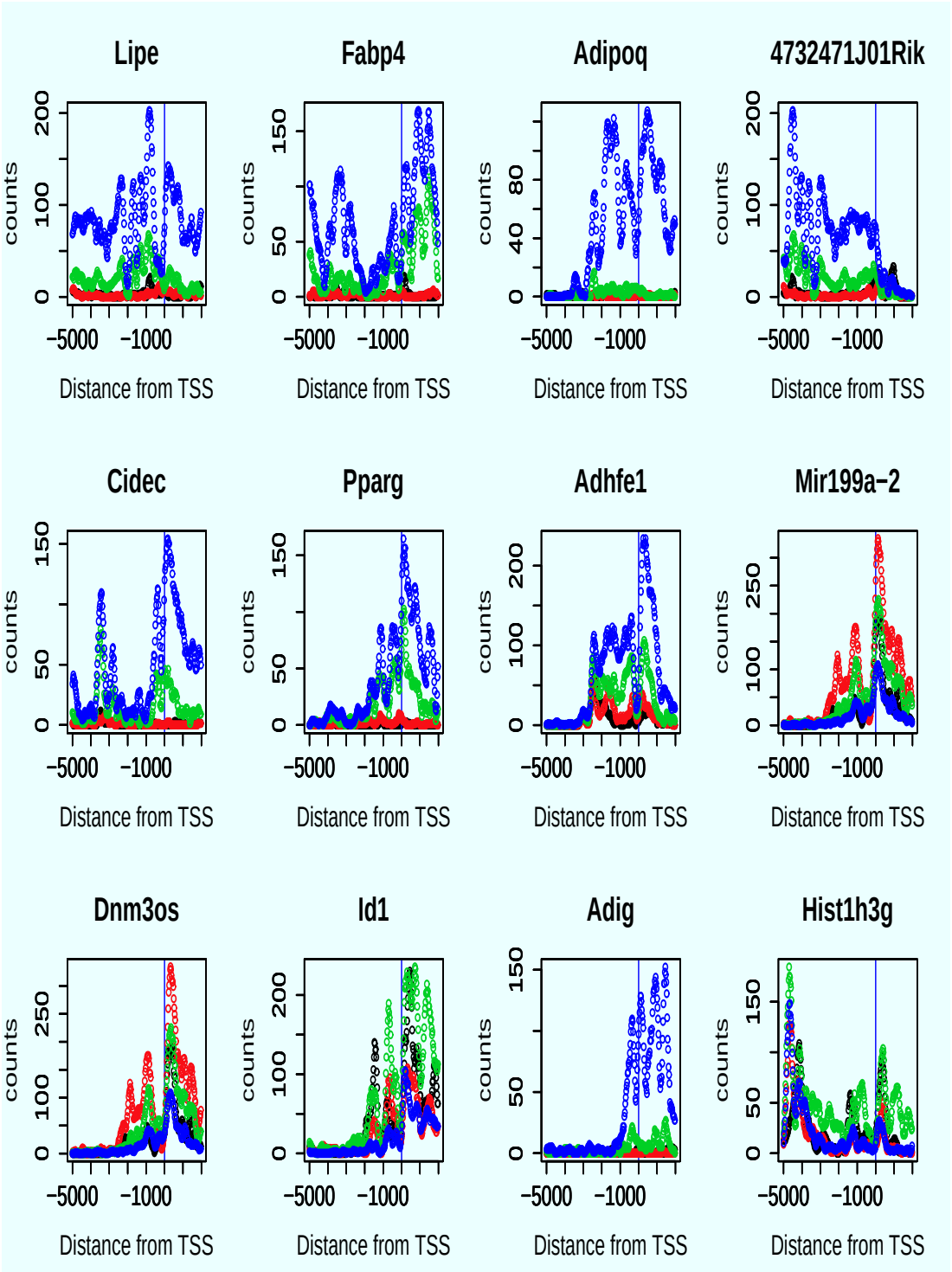


Figure 4.10: Observed ChIP-seq bin-counts for top twelve genes ranked by TS_{mean} statistics over the promotor region for day -2 (black), day 0 (red), day 2 (green) and day 7 (blue). Vertical line represents the transcription starting site.

To see how well these statistics are associated with gene expression changes over four time points, we define the classification indicator $\Delta_k = 1$ if the kth gene has at least one pair of gene expression change greater than two-fold,

$$\Delta_k = I\{|\log_2 \frac{W_{k1}}{W_{k2}}| > 1| \text{ or } |\log_2 \frac{W_{k2}}{W_{k3}}| > 1| \text{ or } |\log_2 \frac{W_{k3}}{W_{k4}}| > 1|\}, \tag{4.23}$$

where W_{kt} is the gene expression level for the kth gene at time point t, for $t = 1, 2, 3, 4$.

Figure 4.11 shows ROC curves using various multi-sample test statistics. We observe that in general using the maximum of the pair-wise statistics tends to result in better ROC curves than using the mean. In addition, $Z_{\lambda,max}$ seems to outperform other methods. Similar pattern are also observed in Figure 4.12, where $Z_{\lambda,max}$ shows a slightly higher proportion of true positive than the other methods, although overall they are close.

4.6. Conclusions and Discussion

We have applied two nonparametric test statistics for differential histone enrichment analysis between two or more conditions. The key of our approach is to apply the null genes or the input ChIP-seq data to define the biologically relevant null values. Compared to the kernel-based tests developed in Chapter 3, these two tests do not require smoothing and bandwidth selection. In addition, no parametric error assumption is needed for these two nonparametric tests. The test with heterogeneous variances also allows for heteroscedasticity where the variance function is no longer a constant across bins in a region.

The kernel smoothing-based test statistics minimize false positive identification of genes by choosing a relatively large bandwidth in order to smooth out the small local

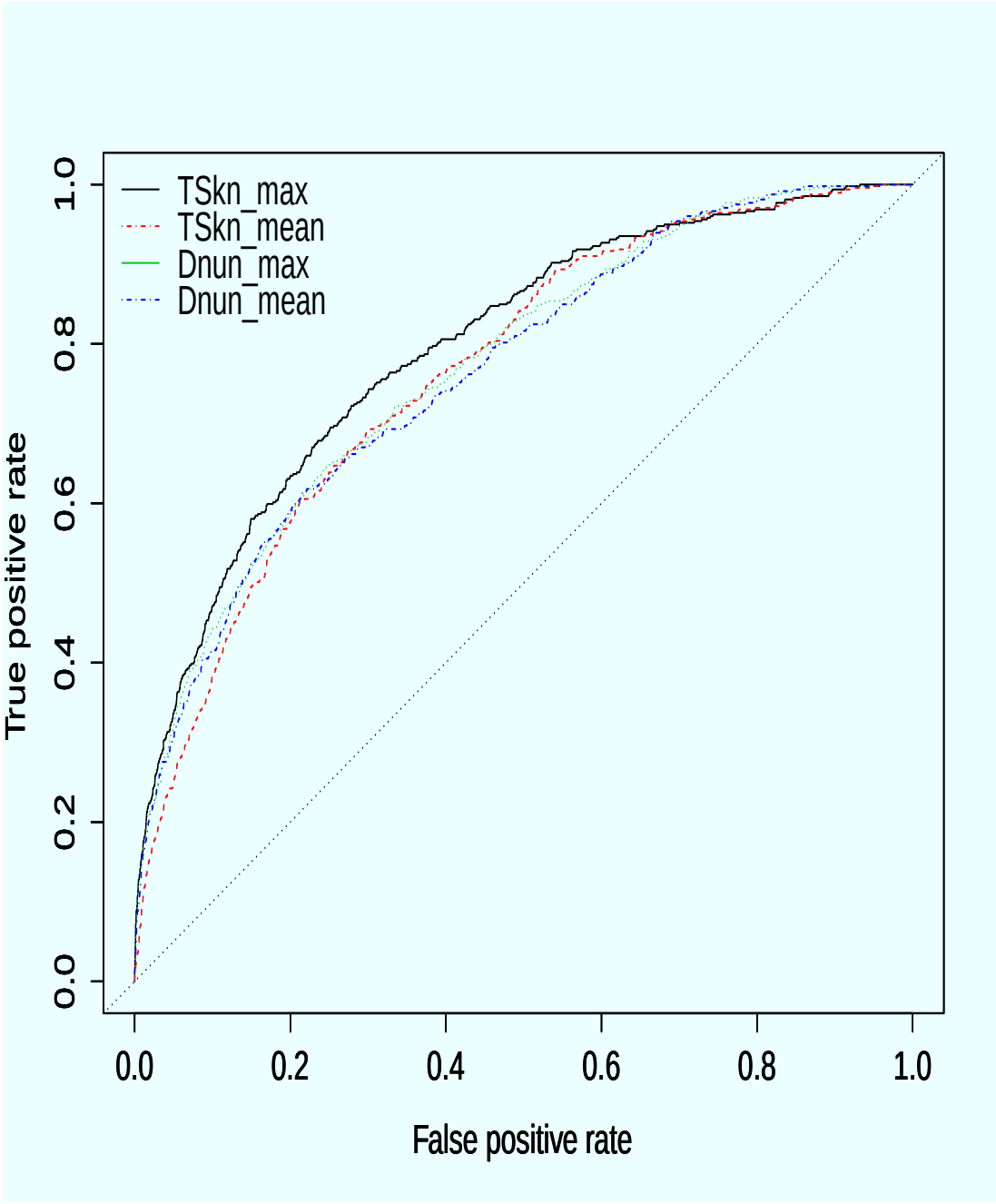


Figure 4.11: Plots of ROC curves for gene expression fold changes using four different test statistics.

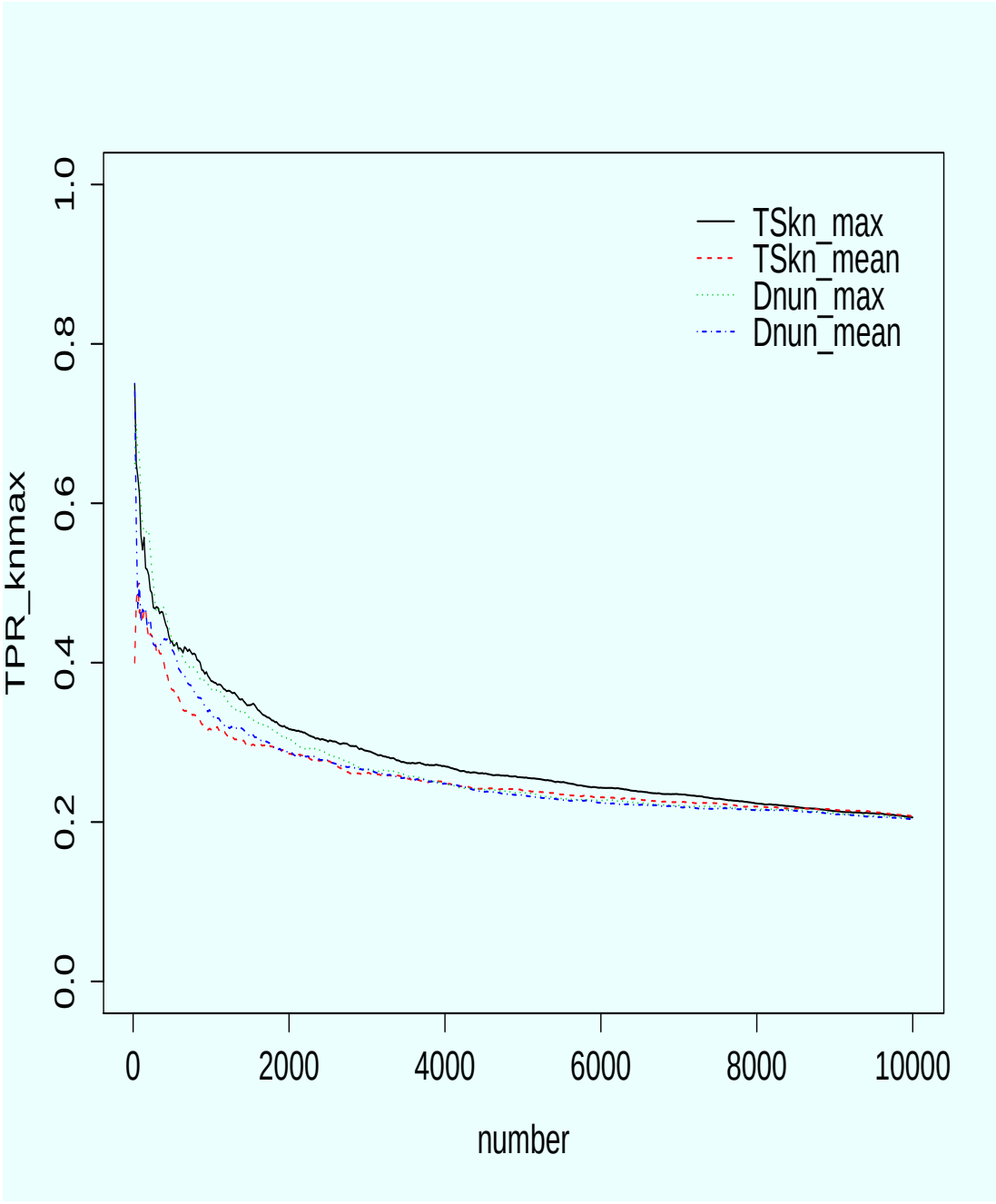


Figure 4.12: Plots of the TPR curves for gene expression fold changes using four different test statistics.

noises. The nonparametric tests considered in this chapter minimize possible false identification of genes with differential enrichment regions by using null genes or by using the input data. Overall, we observe that kernel-based tests slightly outperform the nonparametric tests in identifying genes with DE regions that show large gene expression changes.

Finally, we have also extended the two-sample statistics to multi-sample ChIP-seq analysis, such as time-course ChIP-seq experiments, by using the maximum or mean of the pair-wise test statistics. Such extensions allow us to identify genes with differential enrichment in multiple conditions or over time. We observe that TS_{max} tends to outperform the TS_{mean} since the signal of the mean statistic can be diluted by the negative values of the pair-wise test statistics but the maximum statistic always keeps the strongest signal.

CHAPTER 5

Conclusions and Future Work

This dissertation concerns the problem of signal detection in genomics in multiple-sample settings. We focus on the problem of identifying the genomic regions that show different characteristics than the background regions. The two problems we considered in this dissertation include copy number variants analysis and ChIP-seq data analyses, both have important biological applications.

In Chapter 2, we have developed a procedure to scan the genomes using score statistics in order to identify the CNVs that are associated with the phenotype. This procedure, CNVtest, identifies the CNVs and tests for the association between potential CNVs regions and the phenotype simultaneously. The power of identifying the trait-associated CNVs depends both on the jump size of the CNVs and also the strength of CNV association. CNVtest automatically allows for some shifts in the CNV boundaries among the carriers. The method is particularly effective when the CNV regions from the different carriers do not exactly cover the same intervals. The CNVtest is also flexible and can be applied to identify CNVs associated with different phenotypes through the use of generalized linear models (GLMs).

One interesting extension of CNVtest is to apply the test to CNV association analysis based on the next generation sequencing data. One can use the local median transformation procedure proposed in Cai et al. (2012) to transform the read-depth data to approximately normally distributed data and directly apply the CNVtest to the transformed data. We expect to have similar power and genome-wide error control as the intensity-based data. Similarly, since large-scale exome sequencing data have

been generated to study many diseases, it is also important to extend CNVtest to such data in order to identify the CNVs in the coding regions that are associated with phenotypes. In order to adjust for differential capture efficiencies of different exons or the GC contents, we can simple include the GC contents and the first few principal components of the exon counts data in the GLMs and develop similar score statistics for testing phenotype association.

In Chapters 3 and 4, we have developed methods for identifying genes with differential enrichment of histone modifications between two or more conditions. Instead of scanning the whole genomes, we take a hypothesis testing approach to test several specific regions of a given gene, including promoter, gene body and downstream region. We then apply several nonparametric tests for testing differential enrichment between two conditions for a given gene and a given region. In Chapter 3, we have developed kernel-based nonparametric tests to identify genes with differentially enriched regions between two or more conditions. The key of this approach is to smooth small local signals using a relatively large bandwidth. The top genes selected by our procedure clearly show patterns of differential enrichment and the test statistics correlate with gene expression changes well. In addition, it can be used to predict the fold changes of gene expressions.

In Chapter 4, we have investigated the use of two nonparametric tests without smoothing for differential enrichment analysis, which allow for possible heteroscedasticity in error variances. In addition, the tests do not make any parametric assumption on the error terms. The key of this approach is to use the input ChIP-seq data or null genes to choose the biologically meaningful null values in the hypothesis testing. The methods provide an effective way of applying the input data in differential enrichment analysis of ChIP-seq data.

In both Chapters 3 and 4, we have also extended the methods for differential enrichment analysis to multi-sample cases. These tests can be applied to time-course ChIP-seq experiments. We have investigated ANOVA-type statistics based on kernel-smoothing, and two nonparametric multivariate test statistics. In general, we observe that kernel-smoothing with large bandwidth performs better than nonparametric tests without smoothing in identification of genes with differential enrichment. However, kernel smoothing-based tests can be sensitive to the bandwidth used. One interesting idea that we explored in this dissertation is use the null genes to calibrate the bandwidth selection so that the test statistics of these null genes follow the expected null distribution.

It should be emphasized that the statistical validity of the proposed tests in Chapters 3 and 4 relies rather critically on the fact that the p-values are uniformly distributed in (0,1) under the null hypothesis of no differential enrichment. However, due to issues of ChIP-seq data normalization and local genome sequencing biases, the reference distributions used in calculating the p-values might be inaccurate and the statistical models on which the tests are based can be inadequate for rigorous statistical inferences. To deal with these potential issues, we took an approach of calibrating the Type I errors to the null genes. We used the genes with very small numbers of read counts as null genes and used their test statistics to calibrate the null distribution. This provides one feasible method for choosing the bandwidth in kernel-smoothing based tests discussed in Chapter 3. Alternatively, when input ChIP-seq data are available, we can use these data to determine the minimum biologically interesting null values when applying the nonparametric tests for differential enrichment analysis. Because of these complications, we should emphasize that the p-values from these nonparametric tests are indeed very useful for ranking the differentially enriched can-

didates, but the conventional use of significance tests based on these p-values should not be taken for granted. Similar conclusions have also been drawn for searching for genes with alternative splicing in term of using the p-values (Hu and He, 2012).

Although the real applications in this dissertation have no biological replications, it should be noted that our proposed tests in both Chapters 3 and 4 can be equally applied to differential enrichment analysis when biological replications are available. Since recent studies have indicated that global histone modification patterns predict risk of prostate cancer recurrence (Seligson et al., 2005) and breast cancer patient' outcome, we expect to see many large-scale ChIP-seq data being generated, especially for cancer studies. New statistical methods are needed to identify the genes whose histone modification differences are associated with cancer outcome.

APPENDIX A

Proof

Proof of Theorems in Chapter 2

A.1. Proof of Theorem 2.3.1

According to the construction of $S_{n,\tau}$ in the algorithm, we only need to show

$$P\{\max_{\tau\in\mathcal{I}_0\cap\mathcal{R}}|S_{n,\tau}|>\sqrt{2\log(\hat{r})}\}\to 0.$$

Based on the standard result for the score statistic, $S_{n,\tau}\to_L N(0,1)$ for $\tau\in\mathcal{I}_0$. Then it is enough to show

$$P\{\max_{\tau\in\mathcal{R}}|N(0,1)|>\sqrt{2\log(\hat{r})}\}\to 0. \tag{A.1}$$

The $\hat{r}$ is a random variable determined by the number of intervals included in $\mathcal{R}$. It can be shown that

$$\begin{aligned}
P(\hat{r}<q) &\leq P(\exists I_k\in\mathbb{I}:|\bar{X}_{iI_k}|\leq\nu \text{ for all } i\in\{1,\ldots,n\})\\
&\leq \sum_{I_k\in\mathbb{I}}P(|\bar{X}_{iI_k}|\leq\nu \text{ for } i \text{ being a carrier})\\
&= \sum_{I_k\in\mathbb{I}}P(|N(\mu_k\sqrt{|I_k|},\sigma_k^2)|\leq\nu)\\
&\leq \sum_{I_k\in\mathbb{I}}P\{N(0,\sigma_k^2)\leq\sqrt{2\log(mL)}-\sqrt{2(1+\epsilon)\log(m)}\}\\
&\leq qm^{-C}\to 0,
\end{aligned} \tag{A.2}$$

for some $C > 0$. The first inequality is by the definition of $\mathcal{R}$ and the condition $\bar{s} \leq L < \underline{d}$ in (2.5); the third inequality is by the choice of ν and condition (2.11); the forth inequality is by Mills' ratio and the condition $\log L = o(\log m)$ in (2.5); and the last step is by $\log q = o(\log m)$ in (2.10). Next, we have

$$\begin{aligned}
& P\{\max_{\tau\in\mathcal{R}} |N(0,1)| > \sqrt{2\log(\hat{r})}\} \\
= & \sum_{r=q}^{mL} P\{\max_{\tau\in\mathcal{R}} |N(0,1)| > \sqrt{2\log \hat{r}} \mid \hat{r} = r\} P(\hat{r} = r) + P(\hat{r} < q) \\
\leq & \sum_{r=q}^{mL} rP\{|N(0,1)| > \sqrt{2\log r}\} P(\hat{r} = r) + P(\hat{r} < q) \\
\leq & \sum_{r=q}^{mL} (C/\sqrt{\log r}) P(\hat{r} = r) + P(\hat{r} < q) \\
\leq & C/\sqrt{\log q} + P(\hat{r} < q).
\end{aligned}$$

Then (A.1) follows by (A.2) and the condition $q \to \infty$ as $n \to \infty$ in (2.10).

A.2. Proof of Theorem 2.3.2

By the asymptotic results for score statistic, S_{n,I_k} is asymptotically normally distributed with mean $\sqrt{n}\beta_{I_k} D(I_k)$ and variance 1. Then $P_{H_{1I_k}}\{|S_{n,I_k}| > \sqrt{2\log(\hat{r})}\}$ is approximately equal to

$$\begin{aligned}
& P\{|N(\sqrt{n}\beta_{I_k} D(I_k), 1)| > \sqrt{2\log(\hat{r})}\} \\
\geq & P\{|N(0,1)| > \sqrt{2\log(\hat{r})} - \sqrt{n}\beta_{I_k} D(I_k)\} \\
\geq & P\{|N(0,1)| > \sqrt{2\log(mL)} - \sqrt{2(1+\eta)\log m}\} \\
\geq & 1 - Cm^{-C}
\end{aligned}$$

for some $C > 0$, where the second inequality is by (2.13), and the last step is by $\eta = O(1)$ and Mill's Ratio. Therefore, $H_{I_k 0}$ is rejected with probability going to 1.

Now consider the rest of Theorem 2.3.2. By the asymptotic results of score statistics , $P(S_{n,I_k} > S_{n,\tau})$ is approximately equal to

$$\begin{aligned} & P\{N(\sqrt{n}\beta_{I_k} D(I_k), 1) - N(\sqrt{n}\beta_\tau D(\tau), 1) > 0\} \\ \geq\ & P[N(0,2) > \sqrt{n}\beta_{I_k}\{D(\tau) - D(I_k)\}] \qquad \text{(A.3)} \\ \geq\ & P[N(0,2) > \sqrt{n}\beta_{I_k} g'(\alpha)\sqrt{b''\{g(\alpha)\}/\gamma}\{\sqrt{Var(Z_\tau)} - \sqrt{Var(Z_{I_k})}\}], \qquad \text{(A.4)} \end{aligned}$$

where the first inequality is by $\beta_{I_k} > \beta_\tau$ for any τ s.t. $\tau \cap I_k \neq \emptyset$ and $\tau \neq I_k$. Then it is left to show that

$$Var(Z_\tau) < Var(Z_{I_k}). \qquad \text{(A.5)}$$

Since

$$Var(Z_\tau) = P(|\bar{X}_\tau| > \nu)\{1 - P(|\bar{X}_\tau| > \nu)\},$$

then by the monotonicity of function $f(x) = x(1-x)$ with $x < 1/2$, (A.5) is implied by

$$P(|\bar{X}_\tau| > \nu) < 1/2 \qquad \text{(A.6)}$$

and

$$P(|\bar{X}_\tau| > \nu) < P(|\bar{X}_{I_k}| > \nu) \qquad \text{(A.7)}$$

for any τ s.t. $\tau \cap I_k \neq \emptyset$ and $\tau \neq I_k$. Note that by (2.1) and (2.2), we have

$$\bar{X}_\tau \sim (1-\pi_k)N(0,1) + \pi_k N(\frac{|\tau \cap I_k|}{\sqrt{|\tau|}}\mu_k, \sigma_k^2).$$

Since $(1-\pi_k)P\{|N(0,1)| > \nu\} = o(1)$ given $\nu \gg 1$, and $\pi_k P\{|N(|\tau \cap I_k|\mu_k/\sqrt{|\tau|},\ \sigma_k^2)| > \nu\} < 1/2$ given $\pi_k < 1/2$, (A.6) follows. It is also easy to show that

$$P(|\bar{X}_\tau| > \nu) - P(|\bar{X}_{I_k}| > \nu) = \pi_k[P\{|N(\frac{|\tau \cap I_k|}{\sqrt{|\tau|}}\mu_k, \sigma_k^2)| > \nu\} - P\{|N(\sqrt{|I_k|}\mu_k, \sigma_k^2)| > \nu\}],$$

then (A.7) follows given the fact that $|\tau \cap I_k|/\sqrt{|\tau|} < \sqrt{|I_k|}$ for any τ s.t. $\tau \cap I_k \neq \emptyset$ and $\tau \neq I_k$.

APPENDIX B

Derivation

Derivations of $E(TS0_\lambda)$, $Var(TS0_\lambda)$, d and δ in Chapter 3

B.1. Derivation of $E(TS0_\lambda)$ and $Var(TS0_\lambda)$

From (3.3) and (3.5), $\tilde{Y}_\lambda(t)$ can be approximated by

$$\begin{aligned}\tilde{Y}_\lambda(t) &= \frac{1}{n\lambda}\sum_{i=1}^{n} K(\frac{t-s_i}{\lambda})Y(s_i) \\ &= \frac{1}{n\lambda}\left[K(\frac{t-s_1}{\lambda}),...,K(\frac{t-s_n}{\lambda})\right]^{\mathrm{T}}[Y(s_1),...,Y(s_n)] \\ &= S_\lambda(t)Y.\end{aligned}$$

Based on (3.5), under the null hypothesis,

$$Y=(Y_1,...,Y_n)\overset{H_0}{=}\sigma(W_{t_1},...,W_{t_n})=\sigma(N_1(0,1),...,N_n(0,1))=\sigma X,$$

where $N_i(0,1)$ represents for $i.i.d.$ standard normal random variable, $i=1,...,n$ and $X=(N_1(0,1),...,N_n(0,1))$. Based on (3.8), it can be shown that

$$T_{0\lambda}=\frac{1}{n\sigma^2}Y^{\mathrm{T}}S_\lambda^{\mathrm{T}}S_\lambda Y=\frac{1}{n}X^{\mathrm{T}}S_\lambda^{\mathrm{T}}S_\lambda X=X^{\mathrm{T}}\left(\frac{1}{n}S_\lambda^{\mathrm{T}}S_\lambda\right)X=X^{\mathrm{T}}AX,$$

where $A=1/nS_\lambda^{\mathrm{T}}S_\lambda$. Since $X\sim N_n(\mu,\Sigma)$, where $\mu=(0,...,0)_{n\times 1}$ and $\Sigma=I_n$, we have

$$\mathrm{E}(T_{0\lambda})=\mathrm{E}(X^{\mathrm{T}}AX)=\mathrm{trace}(\mathrm{A}\Sigma)=\mathrm{trace}(\mathrm{A})$$

where

$$a_{ii} = \frac{1}{n^3\lambda^2}\sum_{j=1}^{n} K^2(\frac{t_j - s_i}{\lambda}) \approx \frac{1}{n^2\lambda}\int K^2(\frac{t_j - s_i}{\lambda})d\frac{t_j - s_i}{\lambda} = \frac{1}{n^2\lambda}||K||^2,$$

and

$$a_{ik} = \frac{1}{n^3\lambda^2}\sum_{j=1}^{n} K(\frac{t_j - s_i}{\lambda})K(\frac{t_j - s_k}{\lambda}).$$

Then we can get

$$\mathrm{E}(T_{0\lambda}) = \mathrm{trace(A)} = \frac{1}{\mathrm{n}\lambda}||\mathrm{K}||^2,$$

which is the same as given by Eubank (1999). For Gaussian kernel, we have

$$\mathrm{E}(T_{0\lambda}) = \frac{1}{n\lambda 2\sqrt{\pi}}.$$

To derive the variance of the test statistics, it can be shown that

$$\mathrm{Var}(T_{0\lambda}) = \mathrm{Var}(X^{\mathrm{T}}AX) = \{E[(X^{\mathrm{T}}AX)^2] - E^2(X^{\mathrm{T}}AX)\}$$

and

$$(X^{\mathrm{T}}AX)^2 = \sum_i\sum_j\sum_k\sum_l a_{ij}a_{kl}x_ix_jx_kx_l,$$

where a_{ij}, a_{kl} are the elements of the A matrix. In addition, we have $\mu_4 = \mathrm{E}(x_i^4) = 3$, $\mu_2 = \mathrm{E}(x_i^2x_j^2) = 1, i \neq j$, $\mathrm{E}(x_i) = 0$,$\mathrm{E}(x_i^3x_j) = 0$,..., and all other combinations equal to zero. Based on Theorem 1.6 of Seber and Lee (2003),

$$E[(X^{\mathrm{T}}AX)^2] = (\mu_4 - 3\mu_2^2)a^{\mathrm{T}}a + \mu_2^2[\mathrm{trace(A)}^2 + 2\mathrm{trace(A}^2)]$$

where a is the $n \times 1$ vector of the diagonal elements of A, and

$$\mathrm{Var}(\frac{1}{n}X^{\mathrm{T}}AX) = \mathrm{trace(A)}^2 + 2\mathrm{trace(A}^2) - \mathrm{trace(A)}^2 = 2\mathrm{trace(A}^2)$$

Let $B = A^2$, and we only need the elements on the diagonal of B, where trace(B) = trace(A^2),

$$b_{ii} = \sum_{j=1}^{n} a_{ij}^2 = a_{ii}^2 + \sum_{j \neq i} a_{ij}^2 = \frac{1}{n^4\lambda^2}||K||^4 + \sum_{j \neq i} a_{ij}^2.$$

For Gaussian kernel,

$$\begin{aligned} a_{ik} &= \frac{1}{n^3\lambda^2} \sum_{n} \frac{1}{2\pi} \exp\left[-\frac{2(t_j - \frac{s_i+s_k}{2})^2 + s_i^2 + s_k^2 - \frac{(s_i+s_k)^2}{2}}{2\lambda^2} \right] \\ &= \frac{1}{2\sqrt{\pi}\lambda} \exp\left[-\frac{(s_i - s_k)^2}{4\lambda^2} \right]. \end{aligned}$$

So with $||K||^4 = \frac{1}{4\pi}$,

$$\begin{aligned} b_{ii} &= \frac{1}{n^4\lambda^2}||K||^4 + \sum_{j \neq i} a_{ij}^2 \\ &= \frac{1}{4\pi n^4\lambda^2} \sum_{j=1}^{n} \exp\left[-\frac{1}{2}(\frac{s_i - s_j}{\lambda}) \right] \\ &= \frac{\sqrt{2\pi}\lambda n}{4\pi n^4\lambda^2} \int \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(\frac{s_i - s_j}{\lambda}) d\frac{s_i - s_j}{\lambda} \right] \\ &= \frac{\sqrt{2\pi}}{4\pi n^3\lambda}, \end{aligned}$$

and therefore $\mathrm{Var}(TS0_\lambda) = 2\mathrm{trace(B)} = \frac{1}{\sqrt{2\pi}} \frac{1}{\mathrm{n}^2\lambda}$.

B.2. Derivation of d and δ

We have the approximate expectation and variance of the test statistics for Gaussian kernel under the null,

$$E(TS0_\lambda|\text{Gaussian Kernel}) = \frac{1}{2\sqrt{\pi}}\frac{1}{n\lambda},$$

$$Var(TS0_\lambda|\text{Gaussian Kernel}) = \frac{\sqrt{2\pi}}{2\pi}\frac{1}{n^2\lambda}.$$

It is easy to show that

$$E(TS0_\lambda) = \delta \times d,$$

$$Var(TS0_\lambda) = \delta^2 \times 2d.$$

This gives us the closed expressions for δ and d for Gaussian kernel:

$$\delta = \frac{1}{n}\frac{1}{\sqrt{2}},$$

$$d = \frac{1}{\sqrt{2\pi}}\frac{1}{\lambda}.$$

If d $>$ 50, based on central limited theorem,

$$Z_{\lambda,CLT} = \frac{TS_\lambda - E(TS_{0\lambda})}{\sqrt{Var(TS_{0\lambda})}} \overset{H_0}{\to} N(0,1)$$

But in our real data application with Gaussian kernel function and bandwidth $\lambda = 20/280$, $d = 5.59$ is far less than 10. Instead, we use Wilson-Hilferty transformation

to transform the χ^2 distribution to a standard normal distribution,

$$Z_{\lambda,WH} = \frac{\sqrt[3]{\frac{TS_\lambda}{\delta d}} - (1 - \frac{2}{9d})}{\sqrt{\frac{2}{9d}}} \xrightarrow{H_0} N(0,1)$$

Derivation of $\hat{\rho}_{j,j+1}$ in Chapter 4

B.3. Derivation of $\hat{\rho}_{j,j+1}$ under homoscedasticity assumption

Under the homoscedasticity assumption, $\hat{\rho}^{eql}$ is derived as follows,

$$\begin{aligned}\rho^{eql}_{j-1,j} &= \text{cov}(\hat{TS}_{j-1,j}\ ,\ \hat{TS}_{j,j+1}) \\ &= \frac{\sigma_j^4 + 4\sigma_j^2[\int f_j^2(t)dt - \int f_{j-1}(t)f_j(t)dt - \int f_{j+1}(t)f_j(t)dt + \int f_{j-1}(t)f_{j+1}(t)dt]}{n\ \sigma_{j-1,j}\ \sigma_{j,j+1}}.\end{aligned}$$

From (4.6) and (4.8), the estimates for $\sigma_{j,j+1}$ and σ_j are known.

We can estimate $\int f_j^2(t)dt$ and $\int f_j(t)f_l(t)dt$ by

$$\int f_j^2(t)dt = \frac{\sum_{k=1}^{n-1} X_j(t_k) \times X_j(t_{k+1})}{n-1}$$

and

$$\int f_j(t)f_l(t)dt = \frac{\sum_{k=1}^{n-1} X_j(t_k) \times X_l(t_{k+1}) + X_l(t_k) \times X_j(t_{k+1})}{2(n-1)}$$

where $j \neq l$. We therefore obtain

$$\rho^{eql}_{1,2} = \frac{\sigma_2^4 + 4\sigma_2^2[\int f_2^2 - \int f_1 f_2 - \int f_2 f_3 + \int f_1 f_3]}{(n-1)\ \sigma_{1,2}\ \sigma_{2,3}},$$

and

$$\rho_{2,3}^{eql} = \frac{\sigma_3^4 + 4\sigma_3^2[\int f_3^2 - \int f_2 f_3 - \int f_3 f_4 + \int f_2 f_4]}{(n-1)\ \sigma_{2,3}\ \sigma_{3,4}}.$$

B.4. Derivation of $\hat{\rho}_{j,j+1}$ under heteroscedasticity assumption

$$\rho_{j,j+1}^{uneql} = \frac{||\sigma_j^2||^2 + 4(||\sigma_j f_j||^2 - ||\sigma_j^2 f_{j-1} f_j|| - ||\sigma_j^2 f_{j+1} f_j|| + ||\sigma_j^2 f_{j-1} f_{j+1}||)}{(n-1)\ \sigma_{j-1,j}^{unv}\ \sigma_{j,j+1}^{unv}}$$

based on (4.6) and (4.10), under the heteroscedasticity assumption, we have estimates for $\sigma_{j,j+1}^{unv}$ and $||\sigma_j^2||^2$, and can estimate other terms by

$$||\sigma_j f_j||^2 = \frac{\sum_{k=1}^{n-3} X_j(t_{k+1}) X_j(t_k)(X_j(t_{k+3}) - X_j(t_{k+2}))^2}{2(n-3)},$$

and

$$||\sigma_i^2 f_j f_l|| = \frac{\sum_{k=1}^{n-3}(X_j(t_{k+1}) X_l(t_k) + X_l(t_{k+1}) X_j(t_k))(X_i(t_{k+3}) - X_i(t_{k+2}))^2}{4(n-3)}$$

where $j \neq l$. Therefore we have

$$\rho_{1,2}^{uneql} = \frac{||\sigma_2^2||^2 + 4(||\sigma_2 f_2||^2 - ||\sigma_2^2 f_1 f_2|| - ||\sigma_2^2 f_2 f_3|| + ||\sigma_2^2 f_1 f_3||)}{(n-1)\ \sigma_{1,2}^{unv}\ \sigma_{2,3}^{unv}},$$

and

$$\rho_{2,3}^{uneql} = \frac{||\sigma_3^2||^2 + 4(||\sigma_3 f_3||^2 - ||\sigma_3^2 f_2 f_3|| - ||\sigma_3^2 f_3 f_4|| + ||\sigma_3^2 f_2 f_4||)}{(n-1)\ \sigma_{2,3}^{unv}\ \sigma_{3,4}^{unv}}.$$

BIBLIOGRAPHY

C. Alkan, B.P. Coe, and E.E. Eichler. Genome structural variation discovery and genotyping. *Nat Rev Genet.*, 12:363–375, 2011.

A. Angel, J. Song, C. Dean, and M. Howard. A polycomb-based switch underlying quantitative epigenetic memory. *Nature*, 476:105–108, 2011.

R.B. Arellano-Valle and M.G. Genton. On the exact distribution of linear combinations of order statistics from dependent random variables. *Journal of Multivariate Analysis*, 98(10):1876–1894, 2007.

R.B. Arellano-Valle and M.G. Genton. On the exact distribution of the maximum of absolutely continuous dependent random variables. *Statistics & Probability Letters*, pages 27–35, 2008.

C. Barnes, V. Plagnol, T. Fitzgerald, R. Redon, J. Marchini, D. Clayton, and M.E. Hurles. A robust statistical method for case-control association testing with copy number variation. *Nature Genetics*, 40:1245–1252, 2008.

A. Barski, S. Cuddapah, K. Cui, T.Y. Roh, D.E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, 2007.

P.M. Bentler and J. Xie. Corrections to test statistics in principal hessian directions. *Statistics & probability letters*, 47(4):381–389, 2000.

L.D. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L.H. Zhao. Statistical analysis of a telephone call center: A queing science perspective. *Journal of the American Statistical Association*, 100:36–50, 2005.

L.D. Brown, T. Cai, R. Zhang, L. Zhao, and H. Zhou. The root-unroot algorithm for density estimation as implemented via wavelet block thresholding. *Probability and Related Fields*, 146:401–433, 2010.

T.T. Cai, X.J. Jeng, and H. Li. Robust detection and identification of sparse segments in ultra-high dimensional data analysis. *Journal of Royal Statistical Society, Series B.,*, 74:773–797, 2012.

Y. Chen, M. Jrgensen, R. Kolde, X. Zhao, B. Parker, E. Valen, J. Wen, and A. Sandelin. Prediction of rna polymerase ii recruitment, elongation and stalling from histone modification data. *BMC Genomics*, 12:544, 2011.

A. Cuevas, M. Febrero, and R. Fraiman. An anova test for functional data. *Computational statistics & data analysis*, 47(1):111–122, 2004.

H. Dette and N. Neumeyer. Nonparametric analysis of covariance. *the Annals of Statistics*, 29(5):1361–1400, 2001.

S.J. Diskin, C. Hou, J.T. Glessner, E.F. Attiyeh, M. Laudenslager, K. Bosse1, K. Cole1, Y.P. Moss, A. Wood, J.E. Lynch, K. Pecor, M. Diamond, C. Winter, K. Wang, C. Kim, E.A. Geiger, P.W. McGrady, A.I.F. Blakemore, W.B. London, T.H. Shaikh, J. Bradfield, S.F.A. Grant, H. Li, M. Devoto, E.R. Rappaport, H. Hakonarson, and J.M. Maris. Copy number variation at 1q21.1 associated with neuroblastoma. *Nature*, 459:987–991, 2009.

X. Dong, M.C. Greven, A. Kundaje, S. Djebali, J.B. Brown, C. Cheng, T.R. Gingeras, M. Gerstein, R. Guig, E. Birney, and Z. Weng. Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol.*, 13:R53, 2012.

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

ENCODE Project Consortium, Bernstein B.E., Birney E., Dunham I., Green E.D., Gunter C., and Snyder M. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.

R.L. Eubank. *Nonparametric Regression and Spline Smoothing (Second edition)*. CRC Press, 1999.

L. Feuk, A.R. Carson, and S.W. Scherer. Structural variation in the human genome. *Nature Review Genetics*, 7:85–97, 2006.

T. Gasser, A. Kneip, and W. Köhler. A flexible and fast method for automatic smoothing. *Journal of the american statistical association*, 86(415):643–652, 1991.

P. Hall and J.D. Hart. Bootstrap test for difference between means in nonparametric regression. *Journal of the American Statistical Association*, 85(412):1039–1049, 1990.

W. Hardle and J.S. Marron. Semiparametric comparison of regression curves. *anst*, pages 63–89, 1990.

T. Hastie and R. Tibshirani. *Generalized additive models*, volume 43. Chapman & Hall/CRC, 1990.

H.H. He, C. Meyer, H. Shin, S.T. Bailey, G. Wei, Q. Wang, Y. Zhang, K. Xu, M. Ni, M. Lupien, P. Mieczkowski, J.D. Lieb, K. Zhao, M. Brown, and X.S. Liu. Nucleosome dynamics defines transcriptional enhancers. *Nat Genet*, 42:343–347, 2010.

G. Hon, W. Wang, and B. Ren. Discovery and annotation of functional chromatin signatures in the human genome. *PLoS computational biology*, 5(11):e1000566, 2009.

J. Hu and X. He. Searching for alternative splicing with a joint model on probe measurability and expression intensities. *Journal of the American Statistical Association*, 107(499):935–945, 2012.

A. Jamalizadeh and N. Balakrishnan. Order statistics from trivariate normal and t distributions in terms of generalized skew normal and skew t distributions. *Journal of Statistical Planning and Inference*, pages 3799–3819, 2009.

X.J. Jeng, T.T. Cai, and H. Li. Optimal sparse segment identification with application in copy number variation analysis. *J. Am. Statist. Ass.*, 105:1156–1166, 2010.

X.J. Jeng, T.T. Cai, and H. Li. Simultaneous discovery of rare and common segment variants. *Biometrika*, 100:157–172, 2013.

H. Ji, H. Jiang, W. Ma, D.S. Johnson, R.M. Myers, and W.H. Wong. An integrated software system for analyzing chip-chip and chip-seq data. *Nature biotechnology*, 26(11):1293–1300, 2008.

D.S. Johnson, A. Mortazavi, R.M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science's STKE*, 316(5830):1497, 2007.

R. Karlic, H.R. Chung, J. Lasserre, K. Vlahovicek, and M. Vingron. Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci(USA)*, 107: 2926–2931, 2010.

E. King, J.D. Hart, and T.E. Wehrly. Testing the equality of two regression curves using linear smoothers. *Statistics & Probability Letters*, 12(3):239–247, 1991.

P.F. Kuan, D. Chung, G. Pan, J.A. Thomson, R. Stewart, and S. Kele. A statistical framework for the analysis of chip-seq data. *Journal of the American Statistical Association*, 106:891–903, 2011.

S.G. Landt et al. Chip-seq guidelines and practices of the encode and modencode consortia. *Genome Research*, 22:1813–1831, 2012.

B. Langmead, C. Trapnell, M. Pop, and S.L. Salzberg. Ultrafast and memory-efficient

alignment of short dna sequences to the human genome. *Genome Biology*, 10:R25, 2009.

O.V. Lepski and V.G. Spokoiny. Minimax nonparametric hypothesis testing: the case of an inhomogeneous alternative. *Bernoulli*, 5(2):333–358, 1999.

J. Li and R. Tibshirani. Finding consistent patterns: a nonparametric approach for identifying differential expression in rna-seq data. *Statistical methods in medical research*, 2011.

K. Liang and S. Keleş. Detecting differential binding of transcription factors with chip-seq. *Bioinformatics*, 28(1):121–122, 2012.

E.T. Liu, S. Pott, and M. Huss. Q&a: Chip-seq technologies and the study of gene regulation. *BMC biology*, 8(1):56, 2010.

P. Medvedev, M. Stanciu, and M. Brudno. Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods*, 6:S13–S20, 2009.

T.S. Mikkelsen, Z. Xu, X. Zhang, L. Wang, J.M. Gimble, E.S. Lander, and E.D. Rosen. Comparative epigenomic analysis of murine and human adipogenesis. *Cell*, 143:156–169, 2010.

A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628, 2008.

A. Munk and H. Dette. Nonparametric comparison of several regression functions: exact and asymptotic theory. *Annals of statistics*, pages 2339–2368, 1998.

N. Neumeyer and H. Dette. Nonparametric comparison of regression curves: an empirical process approach. *The Annals of Statistics*, 31(3):880–920, 2003.

H. O'Geen, L. Echipare, and P.J. Farnham. Using chip-seq technologe high-resolution profiles of histone modifications. *Methods Mol Biol.*, 791:265–286, 2011.

A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5 (4): 557–572, 2004.

P.J. Park. Chip-seq: advantages and challenges of a maturing technology. *Nature Review Genetics*, 10:669–680, 2009.

R. Redon, S. Ishikawa, K.R. Fitch, L. Feuk, G.H. Perry, T.D. Andrews, H. Fiegler, M.H. Shapero, A.R. Carson, W. Chen, E.K. Cho, S. Dallaire, J.L. Freeman, J.R. Gonzlez, M. Gratacs, J. Huang, D. Kalaitzopoulos, D. Komura, J.R. MacDonald, C.R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M.J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D.F. Conrad, X. Estivill, C. Tyler-Smith, N.P. Carter, H. Aburatani, C. Lee, K.W. Jones, S.W. Scherer, and M.E. Hurles. Global variation in copy number in the human genome. *Nature*, 444:444 – 454, 2006.

J. Rice. Bandwidth choice for nonparametric regression. *The Annals of Statistics*, 12 (4):1215–1230, 1984.

G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, et al. Genome-wide profiles of stat1 dna association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods*, 4(8):651–657, 2007.

A. Schwartzman, A. Jaffey, Y. Gavrilovz, and C.A. Meyer. Multiple testing of local maxima for detection of peaks in chip-seq data. *Annals of Applied Statistics*, 7: 471–494, 2013.

J. Sebat, B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Maner, H. Massa, M. Walker, M. Chi, and et al. Large-scale copy number polymorphism in the human genome. *Science*, 305:525–528–97, 2004.

G.A.F. Seber and A.J. Lee. *Linear Regression Analysis*. Wiley, 2003.

D.B. Seligson, S. Horvath, T. Shi, H Yu, S. Tze, M. Grunstein, and S.K. Kurdistani. Global histone modification patterns predict risk of prostate cancer recurrence. *Nature*, 435:1262–1266, 2005.

Z. Shao, Y. Zhang, G.C. Yuan, S. Orkin, and D. Waxman. Manorm: a robust model for quantitative comparison of chip-seq data sets. *Genome biology*, 13(3):R16, 2012.

D.O. Siegmund, B. Yakir, and N.R. Zhang. Detecting simultaneous variant intervals in aligned sequences. *Annals of Applied Statistics*, 5:645–668, 2010.

C. Spyrou, R. Stark, A.G. Lynch, and S. Tavaré. Bayespeak: Bayesian analysis of chip-seq data. *BMC bioinformatics*, 10(1):299, 2009.

R. Stark and G. Brown. Diffbind : differential binding analysis of chip-seq peak data. *Bioconductor*, 2011.

H. Stefansson, D. Rujescu, S. Cichon, O.P. Pietilainen, A. Ingason, A. Steinberg,

R. Fossdal, E. Sigurdsson, T. Sigmundsson, J.E. Buizer-Voskamp, and et al. Large recurrent microdeletions associated with schizophrenia. *Nature*, 455:178–179, 2008.

J.L. Stone, M.C. O'Donovan, H. Gurling, G.K. Kirov, D.H. Blackwood, A. Corvin, N.J. Craddock, M. Gill, C.M. Hultman, P. Lichtenstein, and et al. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*, 455:237–241, 2008.

C. Taslim, J. Wu, P. Yan, G. Singer, J. Parvin, T. Huang, S. Lin, and K. Huang. Comparative study on chip-seq data: normalization and binding pattern characterization. *Bioinformatics*, 25(18):2334–2340, 2009.

V.G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

A.E. Urban, J.O. Korbel, R. Selzer, T. Richmond, A. Hacker, G.V. Popescu, J.F. Cubells, R. Green, B.S. Emanuel, M.B. Gerstein, S.M. Weissman, and M. Snyder. High-resolution mapping of dna copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, 103: 45344539, 2006.

T. Walsh, J.M. McClellan, S.E. McCarthy, A.M. Addington, S.B. Pierce, G.M. Cooper, and et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science*, 320:539–543, 2008.

K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, S. Grant, H. Hakonarson, and M. Bucan. Penncnv: an integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome snp genotyping data. *Genome Research*, 17:1665–1674, 2007.

E.B. Wilson and M.M. Hilferty. The distribution of chi-squared. *Proc. Natl. Acad. Sci. USA*, 17:684–688, 1931.

R. Xie, L.J. Everett, H.W. Lim, N.A. Patel, J. Schug, E. Kroon, O.G. Kelly, A. Wang, K.A. DAmour, A.J. Robins, et al. Dynamic chromatin remodeling mediated by polycomb proteins orchestrates pancreatic differentiation of human embryonic stem cells. *Cell stem cell*, 12(2):224–237, 2013.

S.G. Young and A.W. Bowman. Non-parametric analysis of covariance. *Biometrics*, pages 920–931, 1995.

F. Zhang, W. Gu, M.E. Hurles, and J.R. Lupski. Copy number variation in human

health, disease and evolutions. *Annual Review of Genomics and Human Genetics*, 10:451–481, 2009.

N.R. Zhang, D.O. Siegmund, H. Ji, and J. Li. Detecting simultaneous change-points in multiple sequences. *Biometrika*, 00(0):1–18, 2008a.

Y. Zhang, T. Liu, C.A. Meyer, J. Eeckhoute, D.S. Johnson, B.E. Bernstein, C. Nusbaum, R.M. Myers, M. Brown, and W. Li. Model-based analysis of chip-seq (macs). *Genome Biol*, 9(9):R137, 2008b.