8-2017

A tail-based test for differential expression analysis and pathway analysis in RNA-sequencing data

Jiong Chen

A TAIL-BASED TEST FOR DIFFERENTIAL EXPRESSION ANALYSIS AND PATHWAY ANALYSIS IN RNA-SEQUENCING DATA

by

Jiong Chen. M.A.

APPROVED:

_____
Jianhua Hu, Ph.D.
Advisory Professor

_____
Kim-Anh Do, Ph.D.

_____
Jeffrey Morris, Ph.D.

_____
Jing Ning, Ph.D.

_____
Yun Wu, Ph.D.

_____
Ying Yuan, Ph.D.

APPROVED:

_____
Dean, The University of Texas
MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences

TAIL-BASED TEST FOR DIFFERENTIAL EXPRESSION ANALYSIS AND PATHWAY ANALYSIS IN RNA-SEQUENCING DATA

A

DISSERTATION

Presented to the Faculty of

The University of Texas

MD Anderson Cancer Center UTHealth

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

by

Jiong Chen M.A.

Houston, Texas, USA

August 2017

ACKNOWLEDGMENTS

Being a PhD student in the Department of Biostatistics at the University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences is truly a wonderful experience to me. I want to thank many people who make my dissertation possible.

I would like to express my deepest appreciation to my advisor, Dr. Jianhua Hu for all her guidance and patience. She has been with me every step of the way, encouraging and helping, as mentor and friend. I would like to thank my committee members, Dr. Kim-Anh Do, Dr. Jeffrey Morris, Dr. Jing Ning, Dr. Yun Wu, and Dr. Ying Yuan for their valuable comments and suggestions. I also would like to thank Dr. Xuming He for his valuable insights.

My special thanks go to my wife, Yang, who I met and married during my PhD study, for her support and companionship.

ABSTRACT

TAIL-BASED TEST FOR DIFFERENTIAL EXPRESSION ANALYSIS AND PATHWAY ANALYSIS IN RNA-SEQUENCING DATA

Jiong Chen, M.A.

Advisory Professor: Jianhua Hu, Ph.D

RNA sequencing data have been abundantly generated in biomedical research for biomarker discovery and pathway analysis. Such data at the exon-level are usually heavily tailed and correlated. Conventional statistical tests based on the mean or median difference for differential expression likely suffer from low power when the between-group difference occurs mostly in the upper or lower tail of the distribution of gene expression. We propose a tail-based test to make comparisons between groups in terms of a specific distribution area rather than a single location. The proposed test, which is derived from quantile regression, adjusts for covariates and accounts for within-sample dependence among the exons through a specified correlation structure. Through Monte Carlo simulation studies, we show that the proposed test is generally more powerful and robust in detecting differential expression than commonly used tests based on the mean or a single quantile. An application to TCGA lung adenocarcinoma data demonstrates the promise of the proposed method in terms of biomarker discovery. We also extend the proposed test to perform pathway analysis for a set of genes within the same pathway or share similar biological function. Genes in such sets are known to be dependent of each other and our test accounts for their pairwise correlation. Through simulation comparison with commonly used pathway

analysis methods, we show the proposed test yields better results. An application on non-small cell lung cancer pathways from KEGG pathway Database also demonstrates the proposed test is a powerful method in detecting differentially expressed pathways.

TABLE OF CONTENTS

Page

Page

LIST OF FIGURES

LIST OF TABLES

1. Introduction

1.1 Introduction

RNA sequencing (RNA-seq), also known as whole transcriptome shotgun sequencing, has become a popular technology for measuring gene expression levels. RNA-seq is designed to perform genome-wide transcriptome profiling. Specifically, this technology isolates and fragments RNA from cells and converts the RNA fragments into cDNA. Then the fragments are amplified through polymerase chain reaction, the cDNAs are sequenced, and the resulting reads are aligned to a reference genome for annotation. The number of sequencing reads mapped to an exon or a gene in the reference genome can be the output from the pipeline. RNA-seq is widely used in biomedical research because of its high efficiency and reproducibility (Auer and Doerge, 2010). Utilizing such data, researchers are able to extract rich genomic information from biological systems and advance our knowledge about various diseases, including cancer.

1.1.1 Differential Expression Analysis

An important objective in cancer research is to detect differential gene expression between cancer and normal tissue samples, with a goal of discovering cancer biomarkers. The Cancer Genome Atlas (TCGA) Research Network data, sponsored by the National Cancer Institute, has RNA-seq profiling data available for a large number of human tumor samples from various cancer types. This rich data resource provides an unprecedented opportunity for researchers to test and validate analytical methods and make scientific discoveries to advance cancer diagnosis and treatment. In our work,

we focus on TCGA lung adenocarcinoma data as lung adenocarcinoma has become the most common form of lung cancer for both smokers and non-smokers, accounting for nearly 40 % of lung cancer cases diagnosed in the United States (Subramanian and Govindan, 2007).

Several methods have been developed to detect differential gene expression in RNA-seq experiments. Jiang and Wong (2009) modeled the count data within a gene or transcript isoform as an independent random sampling process and used a Poisson distribution to approximate the observations. Bloom et al. (2009) and McCarthy et al. (2012) used Fisher's exact test and the likelihood ratio test for differential expression analysis. Because the conventional Poisson distribution cannot address the often-encountered large variation in the data, DESeq2 (Love et al., 2014) and edgeR (Robinson et al., 2010) adopted the negative binomial distribution to address the overdispersion problem. The two methods use different approaches to normalize the data and filter out outliers prior to estimating dispersion.$DESeq2$ uses a Wald test to make inference about differential gene expression while $edgeR$ uses an exact test adapted for overdispersed data. Limma+voom (Ritchie et al., 2015) is another method commonly used for differential expression (DE) analysis by normalizing the raw count data into log2 counts per million (logCPM) and then applying a linear mixed effect model to analyze differential gene expression. Laird and Ware (1982) detected the group difference while addressing the correlation structure within each gene. However, the normality assumption is usually not satisfied, even with data transformation (Bullard et al., 2010), for example, in data sets with excessive zeros or small counts. In fact, heavy tails are often the characteristic of distributions of gene intensities in the reads per kilobase per million mapped reads (RPKM) data, as we see in the lung adenocarcinoma data analyzed in this paper. These methods may have undesirable properties such as low power and inflated type I error rates according to Bullard et al. (2010) and Chu et al. (2015).

Alternative tests that are not sensitive to data distributions may be constructed based on quantile regression. Corresponding rank score tests based on single quantiles, typically the median, have been widely used (Gutenbrunner et al., 1993). Furthermore, Wang and He (2008) described a modified rank score test to account for correlations among smaller units within a gene in microarray studies. However, such tests based on single quantiles are known to yield low detection power, and it is difficult to know which specific quantiles should be chosen for testing in a given application.

Current DE analysis methods commonly use gene-level read counts by summarizing exon-level sequenced reads form gene-level data. These methods lose potentially useful information about the exon-level expression distribution (Laiho and Elo, 2014). In this paper, we propose a new tail-based test that uses exon-level expression data and accumulates the information on all the quantiles of a tail region. This is motivated by previous research on microarray expression data that shows that statistical testing on probe-level data can improve the detection of differential gene expression over that on gene-level data (Lader et al., 2006). The idea of using quantile aggregation was initially proposed by He et al. (2010), who focused on detecting treatment effects in clinical studies with independent observations of a response variable but ignoring the potential correction of outcomes. RNA degradation renders the read counts unevenly across the different exon regions and commonly cause biases towards the 3' end (Shanker et al., 2015). Hence, we focus on the upper tails in the test since high gene expression intensities are particularly meaningful in the applications. Nevertheless, the test can be easily tailored to the lower tails. In addition, exons belonging to a common gene tend to empirically correlate with each other, as Figure 1.1, which shows a compound symmetry correlation structure on gene $FHIT$. The proposed test is capable of adjusting for covariates and accounting for the inter-exon correlations within a gene.

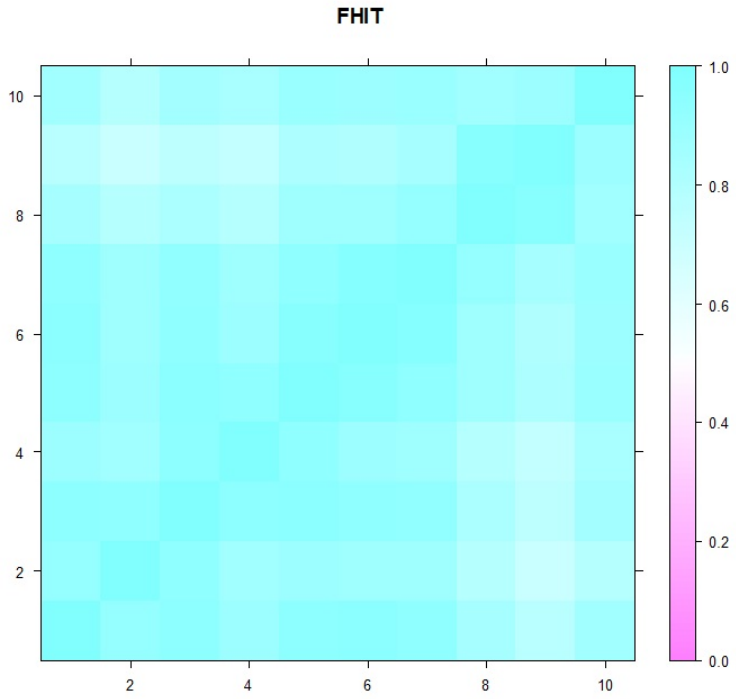

Figure 1.1. Heatmap of correlation on exon-level expression for gene $FHIT$ from TCGA lung adenocarcinoma data.

1.1.2 Pathway Analysis

Pathways or gene sets are a collection of genes that interact with each other and govern certain biological functions. As genes normally function as a group, analysis done on pathways or gene sets of interest would provide more biological insights than individual gene analysis. A current task of biomedical research is to understand the underlying mechanisms of pathways and their interaction with cancer. Researchers have assembled detailed information regarding cancer related pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database, which we use to obtain the pathway information for non-small cell lung cancer (NSCLS) analysis in Chapter 5 (Kanehisa and Goto, 2000). Recently, DE analysis methods have been extended from the detection of individual differential expressed gene to the detection of differential

expressed pathways and gene sets of interest between cancer and normal samples. This type of methods allows researchers to incorporate biological knowledge into the study and formalize systematic analysis on the pathological association and functional significance of pathways and gene sets of interest using hypothesis testing.

Several methods have been proposed to detect the DE of pathways and gene sets. The first category of these methods is called overrepresentation analysis. In this category, one of the most popular methods is the test of independence for 2 by 2 contingency table assessing the overrepresentation of the gene set, which has been discussed by Al-Shahrour et al. (2004); Khatri and Drghici (2005); Boyle et al. (2004). A threshold is selected to separate the genes into DE group and non-DE group in the contingency table and fisher exact test and hypergeometric distribution test are commonly used to conduct the statistical analysis. The popular methods and softwares in this category include GO-based tools by Rivals et al. (2007), BiNGO by Maere et al. (2005), and DAVID by Dennis et al. (2003). This approach is often criticized because the results of the analysis are highly depended on the choice of threshold for DE gene. As an alternative, Al-Shahrour et al. (2005) improve the 2 by 2 contingency table approach by simultaneously testing at various thresholds.

Another main category of pathway DE analysis methods is called functional class scoring. Functional class scoring assigns scores to each gene from the gene set of interest based on their expression level change and calculates the aggregated score of the gene set based on the individual gene scores. The main advantage of this approach is it utilizes the information of every gene from the gene set and the analysis no longer depend on the controversial threshold selection. Pavlidis et al. (2004) use geometric mean of the p-values as the aggregated score and find it generates more consistent results than the overrepresentation approach. Gene Set Enrichment Analysis (GSEA) by Subramanian et al. (2005) and Mootha et al. (2003) is one of the most popular methods developed in this category. Mootha et al. (2003) calculate the

p-values of the genes from the gene set and use a weighted Kolmogorov-Smirnov test to detect whether the ranking order of the p-values differ from a uniform distribution. Subramanian et al. (2005) improve the GSEA by including an ad-hoc modification and generate the null distribution using sample permutation approach. Tian et al. (2005) use an aggregation of t-test statistics and use permutation based method to assess the test significance. Irizarry et al. (2009) argue that the t-test statistics from Tian et al. (2005) are empirically independent and assess the significance of aggregated test statistics using a normal distribution. Unlike GSEA which uses sample permutation method to assess the significance of the test statistics, Generally Applicable Gene set Enrichment (GAGE) proposed by Luo et al. (2009) uses gene permutation method instead. GAGE significantly decreases the computation time and is able to handle data with different samples sizes and experiment designs. Vremo et al. (2013) present R Package *Piano* with a wide range of available functional class scoring methods which allow the choice of gene or sample permutation method.

Current pathway analysis methods for RNA-seq data commonly rely on standard gene DE analysis methods like Limma, edgeR, and DESeq2 to obtain the initial inputs such as Log Fold changes or test statistics for the pathway analysis. In this paper, we propose a tail-based pathway test for RNA-seq data that falls in the category of functional class scoring. The proposed pathway test utilizes the test statistics of individual genes from the tail-based test we proposed in the section 1.1.1 and computes the pathway test statistics. We hypothesize incorporating test statistics from our robust and powerful DE method will strengthen the downstream pathway analysis. Furthermore, many popular pathway analysis methods such as GAGE by Luo et al. (2009) and method by Irizarry et al. (2009) assume independence among the genes or test statistics. For NSCLC data, we observe a pairwise compound symmetry correlation structure for Calcium signaling pathway and *ErbB* signaling pathway as shown in Figure 5.2. Our proposed pathway test adjusts for this correlation structure

and the test statistics follows a standard normal distribution under null hypothesis, which is a desired property for hypothesis testing.

1.1.3 Covariate-adjusted Expected Shortfall Test

Our proposed tail-based test for individual genes is motivated by the COVariate-adjusted Expected Shortfall (COVES) test proposed by He et al. (2010) and Hsu (2010), who use quantile aggregation approach to accommodate observations with heavy tail distribution. He et al. (2010) focus on detecting treatment effects in clinical studies with independent observations of a response variable. This method adjusts for covariates effects which are influential to the outcomes but are independent of the treatment effect, then compare the distribution of the upper quantile region between groups. The limiting distribution of COVES's test statistics follows a standard normal distribution under null hypothesis. Simulation studies from He et al. (2010); Hsu (2010) have shown when the true difference lies in the upper quantile, COVES test performs significantly better than conventional tests such as t-test. However, this method ignores the potential correction of outcomes and cannot be directly applied to DE analysis for RNA-seq data, which is demonstrated in section 2.2.1. In this paper, we built on and tailored COVES test to address the characteristic of RNA-seq data. Our proposed test for DE gene detection is able to account for the correlation structure of inter-exon regions within a gene, and the test statistics can be used in the downstream pathway analysis.

This paper is organized as follows. In chapter 2, we introduce the model and notations and present the tail-based test for DE analysis and its limiting distribution under the null hypothesis. We perform Monte Carlo simulations on correlated data and make comparisons with several conventional tests and popular DE analysis methods. In chapter 3, we analyze TCGA lung adenocarcinoma data using the proposed test and compare with other methods. In Chapter 4, we propose the tail-based path-

way test and introduce its properties. We also conduct Monte Carlo simulations on correlated pathway data and make comparisons with several popular pathway analysis methods. In Chapter 5, we analyze NSCLS pathway data using the proposed pathway test and compare with other methods.

2. A tail-based test for differential expression analysis in RNA-sequencing data

2.1 Methodology

In biomedical applications of microarray studies involving, for example, exon-level RNA-seq data, it is often of interest to detect differential gene expression between disease groups. The proposed method is devised to meet this objective. We first introduce the notations. Let $\boldsymbol{Z}$ denote the gene expression intensity, which is treated as the response measure, wherein Z_{ij} indicates the intensity measurement of the jth exon location in a gene of interest for the ith sample. We use a dummy variable $D = 0, 1$ to denote the control and diseased patient groups, respectively, wherein D_i corresponds to the disease status of sample i. We use $\mathbf{C}$ to indicate K covariates and assume them to be independent of D, and a $K \times 1$ design vector $\boldsymbol{C}_i$ corresponding to the covariates with sample i. The integers n_0 and n_1 respectively indicate the number of patient samples for the groups of $D = 0$ and $D = 1$, and $n = n_0 + n_1$. We use m_i to denote the total number of exon locations belonging to the target gene for the ith sample and N_d to denote the total number of exon locations belonging to the target group of $D = 0$ and $D = 1$.

We express the τth quantile of $\boldsymbol{Z}$, given $\boldsymbol{D}$ and $\boldsymbol{C}$, as

$$Q_{\boldsymbol{Z}}(\tau \mid \mathbf{D}, \mathbf{C}) = \alpha(\tau) + \boldsymbol{D}\delta(\tau) + \mathbf{C}\boldsymbol{\gamma}(\tau) = \mathbf{X}\boldsymbol{\beta}(\tau), \tag{2.1}$$

where $X = (\mathbf{1}_{n\times 1}, \boldsymbol{D}_{n\times 1}, \boldsymbol{C}_{n\times K})$ and $\boldsymbol{\beta}(\tau) = (\alpha(\tau), \delta(\tau), \boldsymbol{\gamma}(\tau)^T_{K\times 1})^T$. Correspondingly, the model for the individual gene intensity measure Z_{ij} can be written as

$$Z_{ij} = \alpha(\tau) + D_i\delta(\tau) + \mathbf{C_i}^T\boldsymbol{\gamma}(\tau) + e_{ij}(\tau), \tag{2.2}$$

where the residuals $e_{ij}(\tau)$ have the value of 0 as the τth conditional quantile. We assume that the inter-exon correlation satisfies $cov(e_{ij}, e_{ij'}) \neq 0$ and $cov(e_{ij}, e_{i'j'}) = 0$. Given $(Z_{ij}, D_i, \mathbf{C_i})$, we obtain the consistent estimate $\hat{\alpha}(\tau), \hat{\delta}(\tau), \hat{\boldsymbol{\gamma}}(\tau)$ at the τth quantile via quantile regression (Koenke et al., 1978). We denote the corresponding empirical residuals as $\hat{e}_{ij}(\tau) = Z_{ij} - \hat{\alpha}(\tau) - D_i\hat{\delta}(\tau) - \mathbf{C_i}^T\hat{\boldsymbol{\gamma}}(\tau)$.

To detect the between-group difference in the gene expression intensity, we define a new tail-based test statistic (TTS) as follows:

$$T_\tau^{TTS}(n_1, n_0) = TTS_\tau(1) - TTS_\tau(0), \tag{2.3}$$

where $TTS_\tau(d) = \sum_{D_i=d}\sum_{j=1}^{m_i} w_{d,i,j}(Z_{ij} - \mathbf{C_i}^T\hat{\boldsymbol{\gamma}}(\tau)), d = 0, 1$. Let $e_{ij}^+ = I(e_{ij} > 0)$ and $e_{ij}^- = I(e_{ij} < 0)$. Herein, $w_{d,i,j} = S_d^{-1}e_{ij}^+(\tau)$, $S_d = \sum_{D_i=d}\sum_{j=1}^{m_i} e_{ij}^+(\tau)$, and $w_{d,i,j}$ serves as a weight for the ith sample at the jth exon location within group $d = 0$ or 1.

Note that $TTS_\tau(d)$ includes the information on residual directions and covariate adjusted residuals, and hence measures the average expression intensity above the τth quantile in group d after adjusting for the covariates. For example, if τ is chosen as the 50th quantile, $TTS_{0.5}(d)$ measures the information for the whole region above the 50th quantile for group d. Accordingly, the test statistic is powered to detect the distributional difference above the 50th quantile between two groups.

Let $\bar{D}_\tau(d)$, $\bar{\boldsymbol{C}}_\tau(d)$, and $\bar{e}_\tau(d)$ be the averages of all the D_i, $\boldsymbol{C}_i$, and e_{ij}, respectively, in group d that are above the τ-th conditional quantile. Specifically, $\bar{\boldsymbol{C}}_\tau(d) = S_d^{-1}\sum_{D_i=d}\sum_j^{m_i}\boldsymbol{C}_i\hat{e}_{ij}^+(\tau)$, and $\bar{e}_\tau(d) = S_d^{-1}\sum_{D_i=d}\sum_j^{m_i}(Z_{ij} - \alpha(\tau) - D_i\delta(\tau) - \mathbf{C_i}^T\boldsymbol{\gamma}(\tau))\hat{e}_{ij}^+(\tau)$. Replacing Z_{ij} with $e_{ij}(\tau) + \alpha(\tau) + D_i\delta(\tau) + \mathbf{C_i}^T\boldsymbol{\gamma}(\tau)$ in $T_\tau^{TTS}(n_1, n_0)$, we can express the test statistic as

$$T_\tau^{TTS}(n_1, n_0) = \delta(\tau) - (\bar{\boldsymbol{C}}_\tau^T(1) - \bar{\boldsymbol{C}}_\tau^T(0))(\hat{\boldsymbol{\gamma}}(\tau) - \boldsymbol{\gamma}(\tau)) + (\bar{e}_\tau(1) - \bar{e}_\tau(0)). \tag{2.4}$$

To perform the test, we establish the asymptotic distribution of $T_\tau^{TTS}(n_1, n_0)$ as $n_0, n_1 \to \infty$ under the null hypothesis of no difference between the two groups. We

first estimate the conditional density function f_{ij} of e_{ij} given (D_i, $\mathbf{C_i}$) evaluated at 0, denoted as $\hat{f}_{n(0)}$. Then, we let $(U_f)_{K\times K} = \sum_i \hat{f}_{n(0)} \boldsymbol{C}_i^* \boldsymbol{C}_i^{*T}$, in which U_f is a combination of the f_{ij} and can be estimated consistently even when the conditional densities vary with C_i (He et al., 2010). We also denote the transformed D and C via Gram-Schmidt orthogonalization as follows,

$$D_i^* = D_i - n_d^{-1} \sum_i D_i I(D_i = d) \tag{2.5}$$

$$\boldsymbol{C}_i^* = \boldsymbol{C}_i - n_d^{-1} \sum_i \boldsymbol{C}_i I(D_i = d), \tag{2.6}$$

In addition, let

$$V_d = \sum_{D_i=d} \sum_{j=1}^{m_i} var(e_{ij} e_{ij}^+) + \sum_{D_i=d} \sum_{j\neq j'} cov(e_{ij} e_{ij}^+, e_{ij'} e_{ij'}^+), \tag{2.7}$$

and $\zeta = P(e_{ij} < 0, e_{ij'} < 0)$.

Lemma 2.1.1

If $\lim_{n_1,n_0\to\infty}(n_1+n_0)^{-1}U_f$ *exists,* $E\|C_i\|_1^3 < \infty$*, the number of exon region* m_i *is some fixed number, and* f_{ij} *are uniformly bounded away from 0 and infinity for all* ij*, then we have the Bahadur representation of* $\hat{\boldsymbol{\gamma}}(\tau)$*,*

$$\hat{\boldsymbol{\gamma}}(\tau) - \boldsymbol{\gamma}(\tau) = U_f^{-1} \textstyle\sum_i m_i^{-1} \sum_{j=1}^{m_i} \boldsymbol{C}_i^* \psi_\tau(e_{ij}(\tau)) + o_p((n_0+n_1)^{-\frac{1}{2}}),$$

and the representation of $\bar{e}_\tau(d)$*,*

$$\bar{e}_\tau(d) = (\textstyle\sum_{D_i=d} \sum_j^{m_i} e_{ij}^+(\tau))^{-1} \sum_{D_i=d} \sum_j^{m_i} e_{ij}(\tau) e_{ij}^+(\tau) + o_p((n_0+n_1)^{-\frac{1}{2}}).$$

Proof of Lemma 2.1.1

This proof is based on the Lemma 2.1.1 from Hsu (2010) with few modifications.

The Bahadur representation of the $(K+2)\times 1$ parameter estimator $\hat{\boldsymbol{\beta}}(\tau)$, according to Koenker (2005) equation 4.4, can be written as

$$\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau) = D_\beta^{-1}(n_0+n_1)^{-1} \textstyle\sum_i m_i^{-1} \sum_{j=1}^{m_i} \mathbf{x}_i{}^* \psi_\tau(e_{ij}(\tau)) + (n_0+n_1)^{-1/2} R_n,$$

where diagonal matrix $D_\beta = \lim_{(n_0+n_1)\to\infty}(n_0+n_1)^{-1}\sum_i \hat{f}_{n(0)}\mathbf{x_i^*}\mathbf{x_i^{*T}}$, $\hat{f}_{n(0)}$ is the estimated conditional density function of e_{ij} given $(D_i, \boldsymbol{C}_i)$ evaluated at 0, $\mathbf{x}_i^* = (1, D_i^*, \mathbf{C_i}^*)$,$R_n = o_p(1)$, and $\psi_\tau(e_{ij}(\tau)) = \tau - e_{ij}^-$.

Then, as $n_0, n_1 \to \infty$, $\sum_i \hat{f}_{n(0)}\mathbf{x}_i^*\mathbf{x}_i^{*T} = \begin{pmatrix} \sum_i \hat{f}_{n(0)} & 0 & 0 \\ 0 & \sum_i \hat{f}_{n(0)} D_i^{*2} & 0 \\ 0 & 0 & \sum_i \hat{f}_{n(0)} \boldsymbol{C}_i^* \boldsymbol{C}_i^{*T} \end{pmatrix}$,

so the diagonal matrix $D_\beta = \begin{pmatrix} \frac{\sum_i \hat{f}_{n(0)}}{(n_0+n_1)} & 0 & 0 \\ 0 & \frac{\sum_i \hat{f}_{n(0)} D_i^{*2}}{(n_0+n_1)} & 0 \\ 0 & 0 & \frac{\sum_i \hat{f}_{n(0)} \boldsymbol{C}_i^* \boldsymbol{C}_i^{*T}}{(n_0+n_1)} \end{pmatrix} + o_p(1)$.

Using the right bottom corner of D_β^{-1}, we can obtain the following,

$\hat{\boldsymbol{\gamma}}(\tau) - \boldsymbol{\gamma}(\tau)$

$= \left[\left\{ \frac{\sum_i \hat{f}_{n(0)} \boldsymbol{C}_i^* \boldsymbol{C}_i^{*T}}{(n_0+n_1)} \right\}^{-1} + o_p(1) \right] (n_0+n_1)^{-1} \sum_i m_i^{-1} \sum_{j=1}^{m_i} \boldsymbol{C}_i^* \psi_\tau(e_{ij}(\tau)) + o_p((n_0 + n_1)^{-\frac{1}{2}})$.

$= \left(\sum_i \hat{f}_{n(0)} \boldsymbol{C}_i^* \boldsymbol{C}_i^{*T} \right)^{-1} \sum_i m_i^{-1} \sum_{j=1}^{m_i} \boldsymbol{C}_i^* \psi_\tau(e_{ij}(\tau)) + o_p((n_0+n_1)^{-\frac{1}{2}})$.

The last equality follows from the central Limit Theorem for $\sum_i m_i^{-1} \sum_{j=1}^{m_i} \boldsymbol{C}_i^* \psi_\tau(e_{ij}(\tau))$. The proof of the second part of Lemma 2.1.1 is equivalent to proving

$$\begin{aligned} &\left\{ \sum_{D_i=d} m_i^{-1} \sum_j^{m_i} \hat{e}_{ij}^+(\tau) \right\}^{-1} \sum_{D_i=d} m_i^{-1} \sum_j^{m_i} e_i(\tau) \hat{e}_{ij}^+(\tau) \\ &- \{n_d(1-\tau)\}^{-1} \sum_{D_i=d} m_i^{-1} \sum_j^{m_i} e_{ij}(\tau) e_{ij}^+(\tau) \\ &= o_p((n_0+n_1)^{-\frac{1}{2}}) \end{aligned}$$

Then, we need to verify the first and second equations below:

$$n_d^{-1} \sum_{D_i=d} m_i^{-1} \sum_j^{m_i} \hat{e}_{ij}^+(\tau) = 1 - \tau + o_p((n_0+n_1)^{-\frac{1}{2}}) \tag{2.8}$$

$$n_d^{-1} \sum_{D_i=d} m_i^{-1} \sum_j^{m_i} e_{ij}(\tau) \left\{ \hat{e}_{ij}^+(\tau) - e_{ij}^+(\tau) \right\} = o_p((n_0+n_1)^{-\frac{1}{2}}). \tag{2.9}$$

We can demonstrate the first equation (2.8) by the second inequality in corollary 2.1 of Koenker (2005),

$n_d^{-1}\sum_{D_i=d} m_i^{-1}\sum_j^{m_i}\hat{e}_{ij}^{+}(\tau) \le 1-\tau \le n_d^{-1}\sum_{D_i=d} m_i^{-1}\sum_j^{m_i}\hat{e}_{ij}^{+}(\tau) + n_d^{-1}p;$

hence, $n_d^{-1}\sum_{D_i=d} m_i^{-1}\sum_j^{m_i}\hat{e}_{ij}^{+}(\tau) = (1-\tau) + o_p(n_d^{-1}) = (1-\tau) + o_p((n_1+n_0)^{-1/2})$.

To prove the second equation (2.9), we can use Lemma 4.6 of He and Shao (1996) and Lemma 11.2 of Owen (2001).

Assume $\{x_i, i \ge 1\}$ are independent random variables drawn from probability distributions $F_{i,\theta} = 1, ..., n$, with a unknown parameter $\theta \in \Theta$, an open subset of $R^m, m \ge 1$. Let a score function $\psi(x_i, \theta)$ with $\lambda_i(\theta) = E\psi(x_i, \theta)$ and $\Lambda_n(\theta) = \sum_{i=1}^n E\psi(x_i, \theta)$, the M-estimator $\hat{\theta}_n$ of θ_0 that satisfies

$$\sum_{i=1}^{n} \psi(x_i, \hat{\theta}_n) = o(\delta_n),$$

where δ_n is a sequence of positive numbers. Let $\mu(x, \theta, d) = \sup_{\upsilon-\theta}|\psi(x, \upsilon) - \psi(x, \theta)|$, where $|.|$ is defined as $|\theta| = \max(|\theta_1|, ..., |\theta_m|)$, and $Z_n(\upsilon, \theta) = |\sum_{i=1}^n\{\psi(x_i, \upsilon) - \psi(x_i, \theta) - \lambda_i(\upsilon) + \lambda_i(\theta)\}|$.

Lemma 4.6 of He and Shao (1996) requires the following conditions:

(B1) $\psi(x, \theta)$ is Borel measurable for fixed $\theta \in \Theta$,.

(B2) There exists $\theta_0 \in \Theta$ such that $\Lambda_n(\theta_0) = 0$ and $|\hat{\theta}_0 - \theta_0| \to 0$ a.s. as $n \to \infty$.

(B3) There exist $r > 0$, $d_0 > 0$ and positive numbers $\{a_i, i \le 1\}$ such that $Eu^2(x_i, \theta, d) \le a_i^2 d^r$ for $|\theta - \theta_0| \le d_0$ and $d \le d_0$.

(B4) $A_{2n} = O(A_n)$, where $A_n = \sum_{i=1}^n a_i^2$.

(B5') For decreasing sequence of positive numbers d_n such that $d_n = O(d_{2n}) = o(1)$, $\max_{1\le i\le n} u(x_i, \theta_0, d_n) = O(A_n^{1/2} d_n^{r/2}(logn)^{-2})$ a.s.

Lemma 4.6 of He and Shao (1996)

Assume that (B1), (B3), and (B5') are satisfied. Then we have

$$\limsup_{n\to\infty}\sup_{|\upsilon-\theta_0|\le d_n} \frac{Z_n(\upsilon, \theta_0)}{(A_n d_n^r + 1)^{1/2}(\text{loglog}(n + A_n))^{1/2}} \le C \text{ a.s.},$$

for some constant $C < \infty$.

Let $\theta_0 = 0$, $m_i^{-1}\sum_j^{m_i} e_{ij}(\tau) = e_i^*(\tau)$ which is independent between samples, $\psi(e_i^*, \theta) = e_i^*(\tau)\{I(e_i^*(\tau) > x_i^T\theta) - I(e_i^*(\tau) > 0\}$, then

$$\begin{aligned}
\lambda_i(\theta) &= E\psi(e_i^*(\tau), \theta) = E(e_i^*(\tau)I(e_i^*(\tau) > x_i^T\theta) - I(e_i^*(\tau) > 0)), \\
Z_{nd}(\upsilon, \theta_0) &= |\sum_{D_i=d}\{\psi(e_i^*(\tau), \upsilon) - \psi(e_i^*(\tau), \theta_0) - \lambda_i(\upsilon) + \lambda_i(\theta_0)\}| \\
&= |\sum_{D_i=d} e_i\{I(e_i^*(\tau) > x_i^T\upsilon) - I(e_i^*(\tau) > 0)\} \\
&\quad + \sum_{D_i=d} E(e_i^*(\tau)\{I(e_i^*(\tau) > 0) - I(e_i^*(\tau) > x_i^T\upsilon)\})|.
\end{aligned}$$

First, we have

$$\begin{aligned}
&n_d^{-1}\sum_{D_i=d} E(e_i^*(\tau)\{I(e_i^*(\tau) > 0) - I(e_i^*(\tau) > x_i^T\upsilon)\}) = n_d^{-1}\sum_{D_i=d}\int_0^{x_i^T\upsilon} e_i^*(\tau)f(e_i^*(\tau))de_i^*(\tau) \\
&= n_d^{-1}\sum_{D_i=d} x_i^T\upsilon\xi_i f(\xi_i) \le n_d^{-1}\sum_{D_i=d}(x_i^T)^2 f(\xi_i) = \upsilon^T(\sum_{D_i=d} f(\xi_i)x_i x_i^T/n_d)\upsilon \\
&= O(\|\upsilon\|_2^2),
\end{aligned}$$

where ξ_i is between 0 and $x_i^T\upsilon$. Therefore,

$$n_d^{-1}Z_{nd}(\upsilon, \theta_0) = |n_d^{-1}\sum_{D_i=d} e_i^*(\tau)\{I(e_i^*(\tau) > x_i^T\upsilon) - I(e_i^*(\tau) > 0)\}| + O(\|\upsilon\|_2^2)$$

Conditions(B1), (B3), and (B5') are checked as:

(B1) $\psi(e_i^*(\tau), \theta) = e_i^*(\tau)\{I(e_i^*(\tau) > x_i^T\theta) - I(e_i^*(\tau) > 0)\}$ is Borel measurable for fixed θ.

(B3)$u(e_i^*(\tau), \theta, d) = \sup_{|\upsilon-\theta|\le d}|e_i^*(\tau)\{I(e_i^*(\tau) > x_i^T\upsilon) - I(e_i^*(\tau) > x_i^T\theta)\}|$

$= \sup_{|\upsilon-\theta|\le d}|e_i^*(\tau)I(x_i^T\upsilon < e_i^*(\tau) < x_i^T\theta)| = |e_i^*(\tau)I(x_i^T\upsilon* < e_i^*(\tau) < x_i^T\theta)|$, where $\upsilon^* = \theta - d(1, 1, \mathrm{sgn}(C_i^T))^T$.

$Eu^2(e_i^*(\tau),\theta,d) = \int_{x_i^T\upsilon^*}^{x_i^T\theta} e_i^*(\tau)^2 f(e_i^*(\tau))de_i^*(\tau) \le M(x_i^T\theta)^2\|x_i\|_1 d \le Md_0^2\|x_i\|_1^3 d,$

where $|\theta| \le d_0$, and $\|x_i\|_1 = 1 + D_i + \|C_i\|_1$. Condition (B3) holds if we take $r = 1$, and $a_i^2 = Md_0^2\|x_i\|_1^3$.

(B5') Let $d_{n_d} = n_d^{-1/2}\log n_d$, we have

$$\begin{aligned}
\frac{\max_{1\le i\le n_d} u(e_i^*(\tau),\theta_0,d_{n_d})}{A_{n_d}^{1/2} d_{n_d}^{1/2}(\log n_d)^{-2}} &= \frac{\max_{1\le i\le n_d}|e_i^*(\tau) I(x_i^T\upsilon^* < e_i^*(\tau) < 0)|}{\{Md_0^2\sum_{D_i=d}\|x_i\|_1^3\}^{1/2} d_{n_d}^{1/2}(\log n_d)^{-2}} \\
&\le \frac{d_{n_d}\max_{1\le i\le n_d}\|x_i\|_1}{M^{1/2}d_0(\sum_{D_i=d}\|x_i\|_1^3)^{1/2} d_{n_d}^{1/2}(\log n_d)^{-2}} \\
&= M^{-1/2}d_0^{-1}\frac{\max_{1\le i\le n_d}\|x_i\|_1 n_d^{-1/2}}{(\sum_{D_i=d}\|x_i\|_1^3/n_d)^{1/2}} d_{n_d}^{1/2}(\log n_d)^2 \to 0 \text{ a.s.} \\
&\quad \text{as } n_d \to \infty,
\end{aligned} \tag{2.10}$$

where $\max_{1\le i\le n_d}\|x_i\|_1 \le 2 + \max_{1\le i\le n_d}\|C_i\|_1 = O(n_d^{1/2})$ according to Lemma 11.2 in Owen (2001), $\sum_{D_i=d}\|x_i\|_1^3/n_d$ is bounded away from 0, and $d_{n_d}^{1/2}(\log n_d)^2 = o(1)$.

Lemma 11.2 of Owen (2001) *Let Y_i be independent random variables with a common distribution and $E(Y_i^2) < \infty$. Let $Z_n = max_{1\le i\le n}|Y_i|$. Then $Z_n = o(n^{1/2})$.*

When (B1), (B3) and (B5') are hold, according to Lemma 4.6 of He and Shao (1996), we have

$$\limsup_{n_d\to\infty}\sup_{|\upsilon-\theta_0|\le d_n}\frac{n_d^{-1}Z_n(\upsilon,\theta_0)}{n_d^{-1}(A_n d_n^r + 1)^{1/2}(\text{loglog}(n + A_n))^{1/2}} \le C \text{ a.s.},$$

The denominator is

$$
\begin{aligned}
&n_d^{-1} Z_n(\upsilon, \theta_0) n_d^{-1} (A_n d_n^r + 1)^{1/2} (\text{loglog}(n + A_n))^{1/2} \\
&= n_d^{-1} (M d_0^2 \sum_{D_i = d} \|x_i\|_1^3) n_d^{-1} (n_d d_{nd} + 1)^{1/2} (\text{loglog}(n_d + M d_0^2 \sum_{D_i = d} \|x_i\|_1^3))^{1/2} \\
&= O(n_d^{-1} (n_d n_d^{-1/2} \text{log} n_d)^{1/2} (\text{loglog} n_d)^{1/2}) = o((n_1 + n_0)^{-1/2}),
\end{aligned}
$$

where $\sum_{D_i=d} \|x_i\|_1^3 n_d^{-1} = O(1)$.

And the numerator is

$$
n_d^{-1} Z_n(\upsilon, \theta_0) = o_p((n_1 + N_0)^{-1/2}), \text{uniformly in } \{\upsilon : |\upsilon - \theta_0| \leq d_{n_d}\}.
$$

Take $\upsilon = \hat{\beta}(\tau) - \beta(\tau)$, we have

$$
n_d^{-1} [\sum_{D-i=d} e_i^*(\tau) \{I(\hat{e}_i^*(\tau) > 0) - I(e_i^*(\tau) > 0)\}] + O(\|\hat{\beta}(\tau) - \beta(\tau)\|_2^2) = o_p((n_1 + n_0)^{-1/2}),
$$

where $O(\|\hat{\beta}(\tau) - \beta(\tau)\|_2^2) = O((n_1 + n_0)^{-1}) = o_p((n_1 + n_0)^{-1/2})$.

Therefore, we have

$$
n_d^{-1} \sum_{D_i = d} m_i^{-1} \sum_{j}^{m_i} e_{ij}(\tau) \left\{ \hat{e}_{ij}^+(\tau) - e_{ij}^+(\tau) \right\} = o_p((n_0 + n_1)^{-\frac{1}{2}}).
$$

Theorem 2.1.1

If $\lim_{n_1,n_0\to\infty} \frac{n_0}{n_0+n_1} \to q \in (0,1)$ *and* $\lim_{n_1,n_0\to\infty}(n_1+n_0)^{-1}U_f$ *exists,* $E\|\boldsymbol{C}_i\|_1^3 < \infty$*, and* f_{ij} *are uniformly bounded away from 0 and infinity, then under the null hypothesis, in which the distribution of the two groups* $F_{Z|\mathbf{C},D=1}$=$F_{Z|\mathbf{C},D=0}$*, we have*

$$T_\tau^{TTS}(n_1,n_0)/s_{n_0,n_1} \to N(0,1) \text{ as } n_1,n_0 \to \infty. \tag{2.11}$$

Proof of Theorem 2.1.1

According to Lemma 2.1.1, and $\delta(\tau)=0$, under the null hypothesis, we can write

$$\begin{aligned} T_\tau^{TTS}(n_1,n_0) = \quad & \left\{ \sum_{D_i=1}\sum_{j=1}^{m_i} e_{ij}(\tau)e_{ij}^+(\tau)/N_1 - \sum_{D_i=0}\sum_{j=1}^{m_i} e_{ij}(\tau)e_{ij}^+(\tau)/N_0 \right\}(1-\tau)^{-1} \\ & -(\bar{\boldsymbol{C}}_\tau^T(1)-\bar{\boldsymbol{C}}_\tau^T(0))U_f^{-1}\sum_{i=1}^{n} m_i^{-1}\sum_{j=1}^{m_i}\boldsymbol{C}_i^*\psi_\tau(e_{ij}(\tau)) + o_p((n_0+n_1)^{-1/2}) \\ & = T_\tau^*(n_1,n_0) + o_p((n_0+n_1)^{-1/2}). \end{aligned}$$

where

$$\begin{aligned} T_\tau^*(n_1,n_0) = \quad & \left\{ \sum_{D_i=1}\sum_{j=1}^{m_i} e_{ij}(\tau)e_{ij}^+(\tau)/N_1 - \sum_{D_i=0}\sum_{j=1}^{m_i} e_{ij}(\tau)e_{ij}^+(\tau)/N_0 \right\}(1-\tau)^{-1} \\ & -(\bar{\boldsymbol{C}}_\tau^T(1)-\bar{\boldsymbol{C}}_\tau^T(0))U_f^{-1}\sum_{i=1}^{n} m_i^{-1}\sum_{j=1}^{m_i}\boldsymbol{C}_i^*\psi_\tau(e_{ij}(\tau)). \end{aligned}$$

Under the null hypothesis, the mean and variance of the test statistics are

$$\begin{aligned} E(T_\tau^*(n_1,n_0)) = \quad & \left\{ \sum_{D_i=1}\sum_{j=1}^{m_i} E(e_{ij}(\tau)e_{ij}^+)/N_1 - \sum_{D_i=0}\sum_{j=1}^{m_i} E(e_{ij}(\tau)e_{ij}^+)/N_0 \right\}(1-\tau)^{-1} \\ & = (1-\tau)^{-1}E(e_{ij}(\tau)e_{ij}^+)(1-1) = 0. \end{aligned}$$

$$
\begin{aligned}
&Var(T_\tau^*(n_1,n_0)) = \\
&(1-\tau)^{-2}(V_1/N_1^2 + V_0/N_0^2) \\
&+\left\{\bar{\boldsymbol{C}}_\tau^T(1)-\bar{\boldsymbol{C}}_\tau^T(0)\right\}U_f^{-1}\left\{\sum_i m_i^{-2}\sum_{j=1}^{m_i}\boldsymbol{C}_i^*\boldsymbol{C}_i^{*T}\tau(1-\tau)\right\}U_f^{-1}\left\{\bar{\boldsymbol{C}}_\tau(1)-\bar{\boldsymbol{C}}_\tau(0)\right\} \\
&+\left\{\bar{\boldsymbol{C}}_\tau^T(1)-\bar{\boldsymbol{C}}_\tau^T(0)\right\}U_f^{-1}\left\{\sum_i m_i^{-2}\sum_{j\neq j'}\boldsymbol{C}_i^*\boldsymbol{C}_i^{*T}(\zeta-\tau^2)\right\}U_f^{-1}\left\{\bar{\boldsymbol{C}}_\tau(1)-\bar{\boldsymbol{C}}_\tau(0)\right\} \\
&+(1-\tau)^{-1}\left\{\sum_{D_i=1}\sum_{j=1}^{m_i}e_{ij}(\tau)e_{ij}^+(\tau)/N_1\right\}\left\{\bar{\boldsymbol{C}}_\tau^T(1)-\bar{\boldsymbol{C}}_\tau^T(0)\right\}U_f^{-1}\sum_{D_i=1}^{n}m_i^{-1}\sum_{j=1}^{m_i}\boldsymbol{C}_i^*\psi_\tau(e_{ij}(\tau)) \\
&-(1-\tau)^{-1}\left\{\sum_{D_i=0}\sum_{j=1}^{m_i}e_{ij}(\tau)e_{ij}^+(\tau)/N_0\right\}\left\{\bar{\boldsymbol{C}}_\tau^T(1)-\bar{\boldsymbol{C}}_\tau^T(0)\right\}U_f^{-1}\sum_{D_i=0}^{n}m_i^{-1}\sum_{j=1}^{m_i}\boldsymbol{C}_i^*\psi_\tau(e_{ij}(\tau)) \\
&-(1-\tau)^{-1}\left\{\bar{\boldsymbol{C}}_\tau^T(1)-\bar{\boldsymbol{C}}_\tau^T(0)\right\}U_f^{-1}/N_1\sum_{D_i=1}^{n}\sum_{j=1}^{m_i}\boldsymbol{C}_i^*m_i^{-1}e_{ij}e_{ij}^+\psi_\tau(e_{ij}(\tau)) \\
&+(1-\tau)^{-1}\left\{\bar{\boldsymbol{C}}_\tau^T(1)-\bar{\boldsymbol{C}}_\tau^T(0)\right\}U_f^{-1}/N_0\sum_{D_i=0}^{n}\sum_{j=1}^{m_i}\boldsymbol{C}_i^*m_i^{-1}e_{ij}e_{ij}^+\psi_\tau(e_{ij}(\tau)) \\
&=(1-\tau)^{-2}(V_1/N_1^2+V_0/N_0^2) \\
&+\left\{\bar{\boldsymbol{C}}_\tau^T(1)-\bar{\boldsymbol{C}}_\tau^T(0)\right\}U_f^{-1}\left[\sum_i m_j^{-2}\left\{\sum_{k=1}^{m_j}\boldsymbol{C}_i^*\boldsymbol{C}_i^{*T}\tau(1-\tau)+\sum_{j\neq j'}\boldsymbol{C}_i^*\boldsymbol{C}_i^{*T}(\zeta-\tau^2)\right\}\right] \\
&\times U_f^{-1}\left\{\bar{\boldsymbol{C}}_\tau(1)-\bar{\boldsymbol{C}}_\tau(0)\right\} \\
&-(1-\tau)^{-1}\left\{\bar{\boldsymbol{C}}_\tau^T(1)-\bar{\boldsymbol{C}}_\tau^T(0)\right\}U_f^{-1}/N_1 \\
&\times\left\{\sum_{D_i=1}^{n}\sum_{j_1=1}^{m_i}\sum_{j_2=1}^{m_i}\boldsymbol{C}_i^*m_i^{-1}e_{ij_1}e_{ij_1}^+\psi_\tau(e_{ij_2}(\tau))-\sum_{D_i=1}^{n}\sum_{j_1=1}^{m_i}e_{ij_1}e_{ij_1}^+\sum_{j_2=1}^{m_i}m_i^{-1}\boldsymbol{C}_i^*\psi_\tau(e_{ij_2}(\tau))\right\} \\
&+(1-\tau)^{-1}\left\{\bar{\boldsymbol{C}}_\tau^T(1)-\bar{\boldsymbol{C}}_\tau^T(0)\right\}U_f^{-1}/N_0 \\
&\times\left\{\sum_{D_i=0}^{n}\sum_{j_1=1}^{m_i}\sum_{j_2=1}^{m_i}\boldsymbol{C}_i^*m_i^{-1}e_{ij_1}e_{ij_1}^+\psi_\tau(e_{ij_2}(\tau))-\sum_{D_i=0}^{n}\sum_{j_1=1}^{m_i}e_{ij_1}e_{ij_1}^+\sum_{j_2=1}^{m_i}m_i^{-1}\boldsymbol{C}_i^*\psi_\tau(e_{ij_2}(\tau))\right\}
\end{aligned}
$$

where

$$V_d = \sum_{D_i=d} \sum_{j=1}^{m_i} var(e_{ij}e_{ij}^+) + \sum_{D_i=d} \sum_{j\neq j'} cov(e_{ij}e_{ij}^+, e_{ij'}e_{ij'}^+),$$

and $\zeta = P(e_{ij} < 0, e_{ij'} < 0)$.

which can be estimated by $s^2_{n_0,n_1}$.

$$\begin{aligned}
s^2_{n_0,n_1} &= (1-\tau)^{-2}\left\{V_1/\left(\sum_{D_i=1} m_i\right)^2 + V_0/\left(\sum_{D_i=0} m_i\right)^2\right\} \\
&+\left\{\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0)\right\} U_f^{-1}\left[\sum_i m_j^{-2}\left\{\sum_{k=1}^{m_j} \boldsymbol{C}_i^* \boldsymbol{C}_i^{*T}\tau(1-\tau) + \sum_{j\neq j'} \boldsymbol{C}_i^* \boldsymbol{C}_i^{*T}(\zeta-\tau^2)\right\}\right] \\
&\times U_f^{-1}\left\{\bar{\boldsymbol{C}}_\tau(1) - \bar{\boldsymbol{C}}_\tau(0)\right\} \\
&-(1-\tau)^{-1}\left\{\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0)\right\} U_f^{-1}/N_1 \\
&\times\left\{\sum_{D_i=1}^{n}\sum_{j_1=1}^{m_i}\sum_{j_2=1}^{m_i} \boldsymbol{C}_i^* m_i^{-1}\hat{e}_{ij_1}\hat{e}^+_{ij_1}\tau - \sum_{D_i=1}^{n}\sum_{j_1=1}^{m_i}\hat{e}_{ij_1}\hat{e}^+_{ij_1}\sum_{j_2=1}^{m_i} m_i^{-1}\boldsymbol{C}_i^*\psi_\tau(\hat{e}_{ij_2}(\tau))\right\} \\
&+(1-\tau)^{-1}\left\{\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0)\right\} U_f^{-1}/N_0 \\
&\times\left\{\sum_{D_i=0}^{n}\sum_{j_1=1}^{m_i}\sum_{j_2=1}^{m_i} \boldsymbol{C}_i^* m_i^{-1}\hat{e}_{ij_1}\hat{e}^+_{ij_1}\tau - \sum_{D_i=0}^{n}\sum_{j_1=1}^{m_i}\hat{e}_{ij_1}\hat{e}^+_{ij_1}\sum_{j_2=1}^{m_i} m_i^{-1}\boldsymbol{C}_i^*\psi_\tau(\hat{e}_{ij_2}(\tau))\right\}
\end{aligned}$$

By the central limit theorem, $T^*_\tau(n_1, n_0)$ is asymptotically normal with mean 0 and variance. Thus, by Lemma 2.1.1 and $T_\tau(n_1, n_0) - T^*_\tau(n_1, n_0) = o_p((n_0+n_1)^{-1/2})$, we prove the asymptotic normality of the test statistic $T_\tau(n_1, n_0)$.

Remark (a): A consistent estimate of U_f can be obtained using the kernel density estimate of f_{ij} based on empirical residuals $\hat{e}_{ij}(\tau)$ (Hardcastle and Kelly, 2010;

Koenker, 2005). We use a Gaussian kernel function to carry out the kernel density estimation in our analysis and select a rule of thumb bandwidth as $h = 0.9A(n_1+n_0)^{(-1/5)}$, as provided by Silverman (1986), where A is the minimum of the standard deviation and interquartile range/1.34 of the empirical residuals.

Remark (b): The term ζ is intended to account for the dependence of exons within a common gene. If the residuals are independent, ζ becomes τ^2 and the rightmost term in the expression of $s^2_{n_0,n_1}$ becomes 0. Empirically, we can estimate ζ and V_d based on $\hat{e}_{ij}$, as follows,

$$\hat{\zeta} = \left\{ \sum_i m_i(m_i-1)/2 - K \right\}^{-1} \sum_i \sum_{j \neq j'} \hat{e}^-_{ij} \hat{e}^-_{ij'}, \tag{2.12}$$

$$\hat{V}_d = \sum_{D_i=d} \sum_{j=1}^{m_i} \left(\hat{e}^2_{ij} \hat{e}^+_{ij} \right) - N_d^{-1} \left(\sum_{D_i=d} \sum_{j=1}^{m_i} \hat{e}_{ij} \hat{e}^+_{ij} \right)^2 \tag{2.13}$$

$$+ \sum_{D_i=d} \sum_{j \neq j'} \left[\left\{ \sum_{D_i=d} m_i(m_i-1) \right\}^{-1} \sum_{D_i=d} \sum_{j \neq j'} \hat{e}_{ij} \hat{e}^+_{ij} \hat{e}_{ij'} \hat{e}^+_{ij'} - {n_d}^{-1} \left(\sum_{D_i=d} \sum_{j} \hat{e}_{ij} \hat{e}^+_{ij} \right)^2 \right]$$

where K is the dimension of $\boldsymbol{C}_i$. We can plug in the estimate of f_{ij} to obtain the variance estimate of $T^{TTS}_{\tau}(n_1, n_0)$.

2.2 Simulation

2.2.1 Simulation studies versus quantile rank score test, linear mixed effect model, and COVariate-adjusted Expected Shortfall test

We conducted simulation studies to investigate the statistical validity and power of the proposed test, TTS. In the first set of simulation studies, we compared TTS to conventional statistical tests, including the quantile rank score test, assuming independent errors (called QRS), the quantile rank score test, assuming correlated errors

(called QRS_c), the Wald test for coefficient estimates of the linear mixed effect model (called LME), and the COVariate-adjusted Expected Shortfall test (called $COVES$) by He et al. (2002). We generated exon-level gene expression data from the following model,

$$Z_{ij} = 5 + \gamma C_i + \delta_1 I(D_i = 1) + \delta_2 I(e_{ij} > 0) I(D_i = 1) e_{ij} + e_{ij}, \text{where} \qquad (2.14)$$

Z_{ij} is the intensity value of exon j of a gene for subject sample i, C_i indicates the covariate value, and D_i indicates the disease status, normal tissue or cancer, of the patient sample i. The corresponding error terms are denoted by e_{ij}s. We investigated the following four scenarios.

Scenario 1: $C_i \sim N(2.5, 0.5^2)$, $\delta_2 = 0$, $\delta_1 = 0$ under H_0 or $\delta_1 = 0.5$ under H_1.

Scenario 2: $C_i \sim N(2.5, 0.5^2)$, $\delta_1 = 0$, $\delta_2 = 0$ under H_0 or $\delta_2 = 1.35$ under H_1.

Scenario 3: $C_i \sim N(2.5, 0.5^2)$ for $D_i = 0$; and $C_i \sim N(2.5, 1)$ for $D_i = 1$, $\delta_1 = 0$, $\delta_2 = 0$ under H_0 or $\delta_2 = 1.35$ under H_1.

Scenario 4: $C_i \sim N(2.5, 0.5^2)$ for $D_i = 0$; and $C_i \sim N(3, 0.5^2)$ for $D_i = 1$, $\delta_1 = 0$, $\delta_2 = 0$ under H_0 or $\delta_2 = 1.35$ under H_1.

In all the scenarios, $\gamma = 1$ and the error terms are normally distributed with unit variance and an exchangeable correlation structure $cor(e_{ij}, e_{ij'}) = 0.8$ and $cor(e_{ij}, e_{i'j'}) = 0$. To study the impact of sample size and gene length on the test, we considered the sample sizes of 50, 75, and 100 subjects per group and gene lengths of 5, 10 and 30 exon locations within a gene, respectively. In each scenario, we ran $5,000$ Monte Carlo samples. For the quantile related test, we used $\tau = 0.5$ for testing H_0 at nominal levels of 1% and 5%, and $\tau = 0.5$ and 0.75 for testing H_1 at the nominal level of 5%.

Scenario 1. In this scenario, the difference between the cancer and normal tissue samples is constant across all the quantiles. The type I error rates are shown in the upper panel of Table 2.1. We observe that QRS and $COVES$ fail to maintain appropriate type I error rates due to high correlation among the exons as their assumptions

Table 2.1

Type I error rates at the nominal levels of 1% and 5% for scenarios 1, 2, and 3. Scenarios 1 and 2 have identical type-I error rates. The values in the table are percentages.

Scenario 1, 2	Nominal Level			1%					5%		
Gene	Sample										
Length	Size	TTS	$COVES$	QRS_c	QRS	LME	TTS	$COVES$	QRS_c	QRS	LME
5	50	1.26	12.72	0.96	16.46	0.78	5.60	25.04	5.46	30.78	4.94
	75	0.98	11.40	0.70	16.12	0.96	4.86	22.74	4.74	28.26	4.74
	100	0.86	12.04	0.96	16.22	0.74	5.16	23.84	5.18	29.00	4.90
10	50	1.44	26.96	1.08	30.44	1.02	5.98	39.86	4.94	43.34	4.44
	75	1.22	27.28	0.98	30.56	0.78	5.72	40.12	5.04	44.40	5.12
	100	1.10	27.16	1.12	29.98	1.04	5.18	40.08	5.20	43.16	4.76
30	50	1.52	51.86	1.20	53.96	1.16	6.14	61.90	5.34	64.04	4.68
	75	1.36	51.16	1.24	54.96	1.16	5.62	61.52	5.20	64.86	5.02
	100	1.32	51.48	1.26	53.72	1.10	5.64	61.86	5.26	63.92	5.38
Scenario 3	Nominal Level			1%					5%		
Gene	Sample										
Length	Size	TTS	$COVES$	QRS_c	QRS	LME	TTS	$COVES$	QRS_c	QRS	LME
5	50	1.48	12.80	1.02	16.18	0.78	5.82	25.30	5.24	30.66	4.80
	75	1.06	11.02	0.86	15.76	0.90	5.06	22.14	5.12	28.78	4.84
	100	0.94	11.94	1.00	17.32	0.84	5.24	23.52	5.24	29.00	4.60
10	50	1.42	26.68	1.10	30.34	0.98	5.98	40.24	5.34	43.62	4.76
	75	1.24	27.68	0.92	31.36	0.70	5.68	40.70	5.30	44.46	5.14
	100	1.02	27.24	1.06	30.18	1.02	5.22	39.78	4.86	44.20	4.78
30	50	1.56	51.08	1.26	54.46	1.20	6.14	61.90	5.48	65.06	4.68
	75	1.32	50.78	1.32	55.04	1.10	5.58	61.82	5.56	64.62	4.98
	100	1.36	51.46	1.46	53.80	1.32	5.88	62.20	5.26	63.54	5.00

are violated. In contrast, TTS, QRS_c, and LME are able to preserve the type I error rates in various cases.

The power results are shown for TTS, QRS_c, and LME in the top panel of Table 2.4. We did not investigate QRS and $COVES$ further due to its statistical

Table 2.2

Type I error rates at the nominal levels of 1% and 5% for scenario 4. Scenarios 1 and 2 have identical type-I error rates. The values in the table are percentages.

Scenario 4	Nominal Level	1%					5%				
Gene Length	Sample Size	TTS	$COVES$	QRS_c	QRS	LME	TTS	$COVES$	QRS_c	QRS	LME
5	50	1.36	13.90	0.84	16.40	0.74	5.94	26.42	5.44	29.16	4.94
	75	1.10	11.88	0.88	16.42	0.88	4.84	23.42	5.20	29.22	4.56
	100	0.84	13.02	0.76	16.46	0.96	5.16	24.24	5.10	29.30	4.64
10	50	1.58	27.78	1.12	30.50	1.06	6.18	41.04	5.40	42.84	5.02
	75	1.08	28.30	1.04	31.00	1.04	5.64	41.72	5.40	43.10	5.18
	100	1.06	28.12	0.98	30.18	1.04	5.28	40.88	5.28	44.56	5.34
30	50	1.52	52.66	1.38	55.30	1.14	5.88	62.78	5.32	64.66	5.00
	75	1.30	51.64	0.86	54.44	0.98	5.12	62.08	5.56	64.66	4.82
	100	1.28	51.44	1.18	53.42	1.00	5.36	62.38	5.08	63.48	4.80

Table 2.3

Difference of mean and quantiles, and the ratio of the variances between cancer and normal groups in scenario 2.

Quantile τ	0.5	0.6	0.7	0.75	0.8	0.9	0.99	Mean	Var ratio
	0.02	0.3	0.68	0.89	1.25	1.61	3.43	0.55	2.58

invalidity. With a constant group difference across the quantiles, it appears that the tests conducted at a single quantile had satisfactory performance. In fact, TTS displayed slightly lower power than LME and QRS_c, which could be caused by the inclusion of additional noise in the upper tails.

Scenario 2. In this scenario, the cancer group ($D_i = 1$) has a heavier right tail and larger variance than the normal group ($D_i = 0$). The difference between the two

groups is relatively small at the median and becomes larger in the upper quantiles as shown in 2.3.

For example, the difference is 0.02 at the median versus 0.89 at the 75th quantile. The ratio of the two groups' variances under H_1 is 2.58. The type I error rates are the same as those in *Scenario 1*. The power results are shown in the middle panel of Table 2.4. In this case, QRS_c shows extremely poor performance at $\tau = 0.5$ since the median group difference is small. TTS, with its capability of utilizing the information in the upper quantile region, shows superior performance at different values of τ compared to both LME and QRS, which only utilize the information of a single, prespecified quantity. The advantage of TTS is more prominent when analyzing smaller sample sizes (e.g., 50), which are often encountered in practice. For example, TTS achieves improvements in power of 40% and 77%, respectively, compared to that achieved by QRS_c and LME in the case of 50 subjects and 5 exons in a gene at $\tau = 0.75$.

Scenarios 3 and 4. These two cases are similar to *Scenario 2*, except that the covariate C_i is generated with either different variances between the two groups in *Scenario 3* or different means in *Scenario 4*. The type I error rates are shown in the lower panel of Table 2.1 and 2.2. The type I error rates of the proposed test, TTS, are well maintained at the corresponding nominal level in the various setups. The power results displayed in Table 2.5 support the superior performance of TTS over that of the other two tests in both scenarios.

Remark: Without prior knowledge of which quantiles show the true difference between groups, TTS shows satisfactory detection power overall as it utilizes information across multiple quantiles in a tail region.

Table 2.4
Power for scenarios 1 and 2 at quantiles $\tau = 0.5$ and 0.75 at the significance level of 0.05. The values in the table are percentages.

Scenario 1		$\tau = 0.5$			$\tau = 0.75$		
Gene	Sample						
Length	Size	TTS	QRS_c	LME	TTS	QRS_c	LME
5	50	65.62	65.90	75.10	53.84	59.00	75.10
	75	83.48	83.78	90.82	70.30	78.32	90.82
	100	91.80	92.14	96.66	81.58	88.28	96.66
10	50	68.70	71.14	78.36	56.52	64.04	78.36
	75	83.88	86.24	91.40	72.00	80.48	91.40
	100	92.10	93.56	96.68	82.34	89.52	96.68
30	50	69.32	71.92	77.62	57.58	65.16	77.62
	75	85.42	87.40	92.00	73.92	83.22	92.00
	100	92.82	94.04	96.46	83.76	91.30	96.46
Scenario 2		$\tau = 0.5$			$\tau = 0.75$		
Gene	Sample						
Length	Size	TTS	QRS_c	LME	TTS	QRS_c	LME
5	50	85.60	6.92	55.00	97.92	69.62	55.00
	75	96.06	6.42	72.44	99.72	84.22	72.44
	100	99.04	6.68	83.80	100.00	92.56	83.80
10	50	86.12	6.74	54.88	98.32	71.40	54.88
	75	96.14	6.16	73.64	99.88	86.12	73.64
	100	99.28	6.14	85.00	100.00	93.44	85.00
30	50	87.96	6.74	57.20	98.78	75.24	57.20
	75	96.80	6.12	74.22	99.92	87.52	74.22
	100	99.34	6.60	86.24	100.00	94.96	86.24

Table 2.5
Power for scenarios 3 and 4 at quantiles $\tau = 0.5$ and 0.75 at the significance level of 0.05. The values in the table are percentages.

Scenario 3		$\tau = 0.5$			$\tau = 0.75$		
Gene	Sample						
Length	Size	TTS	QRS_c	LME	TTS	QRS_c	LME
5	50	86.08	6.58	55.10	98.06	70.32	55.10
	75	96.26	6.18	72.86	99.80	85.08	72.86
	100	99.00	6.26	83.22	100.00	92.76	83.22
10	50	86.54	6.42	54.76	98.44	71.86	54.76
	75	96.18	6.12	73.68	99.92	86.48	73.68
	100	99.28	5.88	84.86	100.00	93.68	84.86
30	50	88.14	6.62	57.20	98.82	75.72	57.20
	75	96.88	6.12	74.58	99.94	87.92	74.58
	100	99.26	6.04	86.14	100.00	95.18	86.14
Scenario 4		$\tau = 0.5$			$\tau = 0.75$		
Gene	Sample						
Length	Size	TTS	QRS_c	LME	TTS	QRS_c	LME
5	50	78.66	6.30	42.00	94.66	57.58	42.00
	75	93.20	6.44	58.80	99.32	74.82	58.80
	100	97.06	6.10	65.56	99.72	81.22	65.56
10	50	79.52	6.00	41.26	95.92	59.06	41.26
	75	93.58	6.56	59.96	99.64	76.66	59.96
	100	97.68	5.84	67.94	99.92	83.44	67.94
30	50	82.06	6.34	43.08	96.84	62.12	43.08
	75	94.86	6.20	61.12	99.58	79.62	61.12
	100	98.12	5.74	67.30	99.94	85.42	67.30

2.2.2 Simulation studies versus *edgeR*, *DESeq2*, and *Limma*, part 1

In the second set of simulation studies, we compared TTS to state-of-the-art DE analysis methods including edgeR (called *edgeR*), DESeq2 (called *DESeq2*), and Limma+voom (called *Limma*). We generated exon level gene expression data in Log2-RPKM format from the following model to fit our model, and converted the measurement to gene-level raw counts to fit other DE analysis methods.

$$Z_{ij} = \alpha + \gamma C_i + \delta I(e_{ij} > 0) I(D_i = 1) e_{ij} + e_{ij}, \text{where} \tag{2.15}$$

Z_{ij} is the intensity value of exon j of a gene for subject sample i, C_i indicates the covariate value, and D_i indicates the disease status, normal tissue or cancer, of the patient sample i. The corresponding error terms are denoted by e_{ij}s. We investigated the following two scenarios.

Scenario *DE-1* (null hypothesis): $\delta = 0$.

Scenario *DE-2* (alternative hypothesis): $\delta = 0$ for 90% of the expression data to simulate non-DE genes and $\delta \sim uniform(1, 2)$ for 10% of the expression data to simulate DE genes.

In both scenarios, we used $\alpha \sim uniform(2, 10)$ to denote the baseline gene expression. We used $C_i \sim N(2.5, 0.5^2)$ to denote the covariates and let $\gamma = 1$. The error terms are normally distributed with unit variance and an exchangeable correlation structure $cor(e_{ij}, e_{ij'}) = 0.8$ and $cor(e_{ij}, e_{i'j'}) = 0$. To study the impact of sample size and gene length on the test, we considered the sample sizes of 40, 60, and 80 subjects per group and gene lengths of 5, 10 and 30 exon locations within a gene, respectively. In each scenario, we ran $5,000$ Monte Carlo samples. For quantile related tests, we used $\tau = 0.5$ for testing scenario *DE-1* at the nominal levels of 1% and 5%, and $\tau = 0.5$ for testing scenario *DE-2* at the nominal level of 5%.

Table 2.6
FPRs at the nominal levels of 1% and 5% for scenario *DE-1*. The values in the table are percentages.

Scenario	Nominal	1%				5%			
DE-1	Level								
Gene	Sample								
Length	Size	*TTS*	*edgeR*	*DESeq2*	*Limma*	*TTS*	*edgeR*	*DESeq*	*Limma*
5	40	1.20	1.68	1.84	0.92	5.56	7.32	7.48	4.92
	60	1.38	1.82	2.02	0.96	5.56	7.66	8.02	5.24
	80	1.46	2.16	2.30	1.20	5.34	7.76	8.16	5.30
10	40	1.38	2.26	2.38	0.88	5.72	7.62	8.50	5.12
	60	1.34	2.06	2.30	1.02	6.04	7.68	7.88	4.68
	80	1.34	2.26	2.40	0.96	5.64	8.30	8.82	4.98
30	40	1.56	2.02	2.36	0.98	6.04	7.72	8.68	5.30
	60	1.20	1.74	2.16	1.08	5.60	7.58	8.26	4.68
	80	1.28	1.86	2.04	0.86	5.74	7.72	8.62	4.84

We calculated the average false positive rates (FPRs) and true positive rates (TPRs) to measure and compare the performance of the aforementioned 4 methods.

Scenario DE-1. The FPRs are shown in Table 2.6. We observe that *edgeR* and *DESeq*2 are sensitive to noise and show inflated FPRs. In contrast, *TTS* and *Limma* can maintain the FPRs around the nominal value.

Scenario DE-2. In this scenario, the cancer group ($D_i = 1$) has a heavier right tail and larger variance than the normal group ($D_i = 0$) for DE genes. The difference between the two groups is relatively small at the median and becomes larger in the upper quantiles as shown in Figure 2.1. As shown in Table 2.7, *edgeR* and *DESeq*2 are sensitive to noise and result in inflated FPRs, while *TTS* is able to preserve the

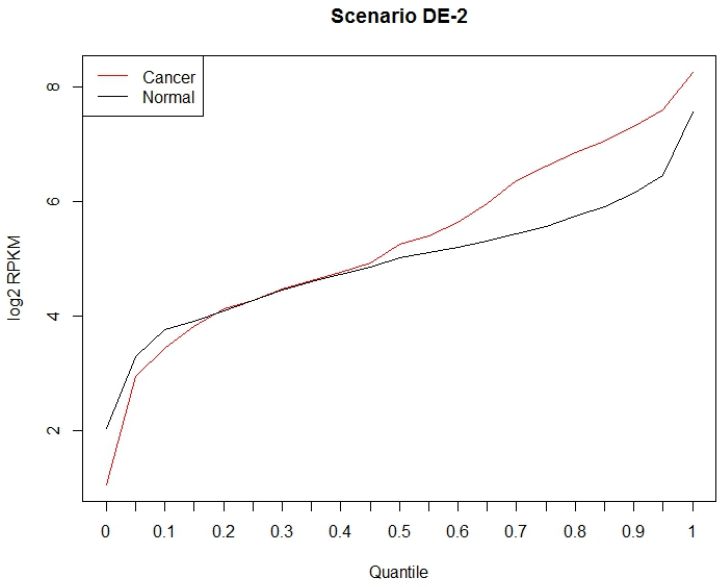

Figure 2.1. Quantile intensity plots of normal tissue and cancer samples for scenario DE-2

FPRs at appropriate levels. For TPRs, TTS has the same performance compared to both *edgeR* and $DESeq2$, while *Limma* has inferior performance. Overall, TTS has better performance than *edgeR*, $DESeq2$, and *Limma* as it outperforms *edgeR* and $DESeq2$ in FPRs and *Limma* in TPRs.

Table 2.7
FPRs and TPRs at the nominal level of 5% for scenarios *DE-2*. The values in the table are percentages.

Scenario		FPR				TPR			
DE-2									
Gene Length	Sample Size	*TTS*	*edgeR*	*DESeq*2	*Limma*	*TTS*	*edgeR*	*DESeq*	*Limma*
5	40	6.18	8.11	9.68	5.53	96.40	98.40	98.00	82.60
	60	5.42	8.96	10.84	6.82	99.60	99.60	99.60	95.60
	80	5.89	10.27	11.72	7.89	100.00	100.00	100.00	96.00
10	40	5.96	7.98	10.32	5.62	97.20	98.00	98.00	84.60
	60	5.53	9.47	10.96	6.71	99.40	99.40	99.40	93.40
	80	5.91	10.18	11.40	7.71	99.80	99.60	99.60	96.60
30	40	6.33	8.49	11.36	6.60	95.20	96.80	97.40	84.40
	60	5.62	9.44	12.04	6.93	99.80	99.60	99.80	96.00
	80	5.58	9.24	12.12	6.78	99.80	99.80	99.80	98.40

2.2.3 Simulation studies versus *edgeR*, *DESeq2*, and *Limma*, part 2

In the third set of the simulation, we generated raw counts of gene-level expression data. We fitted edgeR, DESeq2, and Limma+voom using raw count data and converted the gene level measurements to exon-level Log2-RPKM measurements to fit our methods.

For a gene g, its mean expression level γ_g was generated from an exponential distribution with mean 100. We generated covariate C_i from a normal distribution $N(2.5, 0.5^2)$. Then we let the regulating factor $\delta_g = 1$ for the normal group. We generated the count data for N_{gj} of gene g for subject i from a negative binomial distribution.

We investigated the following two scenarios.

Scenario *DE-3* (null scenario): $\delta_g = 1$ for all genes in the cancer group.

Scenario *DE-4* (alternative scenario): For the cancer group, $\delta_g = 1 + X_g$ for 5% of the expression data to simulate up-regulated DE genes and $\delta_g = (1 + X_g)^{-1}$ for 5% of the expression data to simulate down-regulated DE genes, where X_g follows an exponential distribution with rate=2. Let $\delta_g = 1$ for the remaining 90% of the expression data to simulate non-DE genes.

In each scenario, we ran $5,000$ Monte Carlo samples. For the quantile related test, we used $\tau = 0.5$ for testing both scenarios at a nominal level of 5%.

To convert the gene-level count data to exon-level count data, we allocated the count of gene g from subject i to m_i exon regions with probabilities p_1^g, ..., p_j^g and $\sum_{j=1}^{m_i} p_j^g = 1$. Following the allocation method of Lin and Sun (2012), we generated p_j^g by $p_j^g = P_j^g / \sum_{j=1}^{m_i} P_j^g$, where P_j^g follows the standard exponential distribution. The majority of the reads were mapped to 1 or 2 exon regions when $k \leq 5$.

The results for scenario DE-3 are shown in Table 2.8. The FPRs of the four tests considered here are all around the nominal level.

Table 2.8
FPRs at the nominal level of 5% for scenario *DE-3*. The values in the table are percentages.

Scenario *DE-3*	Nominal Level	5%			
Gene Length	Sample Size	TTS	$edgeR$	$DESeq2$	$Limma$
5	40	5.60	4.34	5.04	4.94
	60	5.38	4.30	5.00	5.06
	80	5.64	4.62	5.08	5.10
10	40	5.48	4.50	5.08	4.96
	60	5.38	4.76	5.46	4.96
	80	5.16	4.52	5.02	5.00
30	40	5.26	3.98	4.74	4.54
	60	5.60	4.22	5.04	4.94
	80	5.02	4.20	4.82	4.80

The results for scenario DE-4 are shown in Table 2.9. All methods have correct FPRs at the appropriate level and achieve similar TPRs for various exon lengths and sample sizes. Such results demonstrate that the proposed test is robust and comparable with $edgeR$, $DESeq2$, and $Limma$ even when the data do not follow our assumed model.

Remark: In scenario DE-1 and DE-2, TTS is able to control FPRs appropriately while $edgeR$ and $DESeq2$ have inflated FPRs. TTS also achieves better TPRs than $Limma$. In scenarios DE-3 and DE-4, TTS controls FPRs and achieve similars TPRs similar to those of state-of-the-art DE methods.

Table 2.9
FPRs and TPRs at the nominal level of 5% for scenarios *DE-4*. The values in the table are percentages.

Scenario		FPR				TPR			
DE-4									
Gene Length	Sample Size	*TTS*	*edgeR*	*DESeq*2	*Limma*	*TTS*	*edgeR*	*DESeq*	*Limma*
5	40	5.60	4.29	5.53	5.60	60.80	60.20	61.80	60.40
	60	5.47	4.69	5.36	5.84	72.60	71.80	73.60	71.00
	80	5.49	4.73	5.33	5.00	75.80	76.60	76.80	75.40
10	40	5.69	4.36	5.22	5.18	62.00	61.40	63.60	59.60
	60	4.93	4.09	4.71	5.02	71.60	72.00	72.00	70.60
	80	5.40	4.40	5.07	5.69	78.80	77.80	79.00	76.80
30	40	5.47	4.33	5.18	5.07	65.40	63.00	65.00	61.60
	60	4.98	4.47	4.87	5.56	71.80	70.80	72.00	69.60
	80	5.00	4.51	4.87	5.18	77.00	75.40	76.80	76.40

3. An application on TCGA lung adenocarcinoma data to detect differential expressed Genes

3.1 Introduction

We analyzed the lung adenocarcinoma data accessible at the TCGA public data portal, with the RNA-seq data profiled from 50 cancer and 50 normal tissue samples at the exon-level and gene-level. The gene expression data were normalized into Log2-RPKM following standard protocols, then the non-expressed genes in both groups were eliminated (Mortazavi et al., 2008) prior to our downstream analysis. As ancillary clinical information, we also considered gender and smoking status in our study. The objective was to detect genes differentially expressed between cancer and normal tissue samples. In particular, our focus was chromosome 3, which has been shown to harbor genes that have potentially important associations with lung adenocarcinoma (Marileila, 2010). We applied the proposed test, TTS, the quantile rank score test, QRS_c of (Wang and He, 2008) at single quantile levels, and the Wald test from the linear mixed model, LME, to each gene, and used a 5% false discovery rate (FDR) adjustment to control for multiple testing (Benjamini and Hochberg, 1995). We also applied standard gene-level differential expression analysis methods including likelihood ratio test from edgeR (Robinson et al., 2010), Wald test from DESeq2 (Love et al., 2014), and ordinary linear model t-test from Limma (Ritchie et al., 2015).

3.2 Results

We included gender and smoking status, defined as current smoker, reformed smoker, and nonsmoker, as covariates in the analysis. TTS detected 537 and 465 genes at $\tau = 0.5$ and 0.75, respectively; and QRS_c detected 484 and 519 genes at $\tau = 0.5$ and 0.75, respectively, while LME detected 501 genes. The top Venn diagrams in Figure 3.1 show the number of overlapping gene among the three tests. We observed that 75% and 84% of the genes detected by TTS were also detected by QRS_c at $\tau = 0.5$ and 0.75, respectively. Moreover, 83% and 77% of the genes selected by TTS at $\tau = 0.5$ and 0.75, respectively, also appear in the list of genes selected by LME. *Limma* detected 684 genes, *edgeR* detected 700 genes, and $DESeq2$ detected 70 genes. The bottom Venn diagrams in Figure 3.1 show the number of overlapping gene among the four tests. We observed that 91% and 86% of the genes detected by TTS were also detected by *limma* at $\tau = 0.5$ and 0.75, respectively; 91% and 90% of the genes selected by TTS were also detected by *edgeR* at $\tau = 0.5$ and 0.75 respectively; and 6% and 5% of the genes selected by TTS were also detected by $DESeq2$ at $\tau = 0.5$ and 0.75, respectively.

Some of the genes detected by TTS were not detected by the other tests. To evaluate the performance of the proposed test, we used prior knowledge from the literature regarding the important genes associated with lung adenocarcinoma. Specifically, six tumor suppressor genes on chromosome 3 have been reported to have strong associations with lung adenocarcinoma, namely, $FHIT$, $RASSF1$, $TUSC2$, $SEMA3B$, $SEMA3F$, and $MLH1$ (Marileila, 2010). For example, $FHIT$ is an identified tumor-suppressor gene that has abnormal expression in lung cancer. In Table 3.1, we report the p-value of these six genes obtained by TTS and QRS_c at $\tau = 0.5$ and by LME with and without the covariates of gender and smoking status.

TTS, LME, *Limma*, and *edgeR* were able to detect $SEMA3B$, $RASSF1$. TTS and *edgeR* also detected $FHIT$, while LME detected $SEMA3F$ with a modest FDR

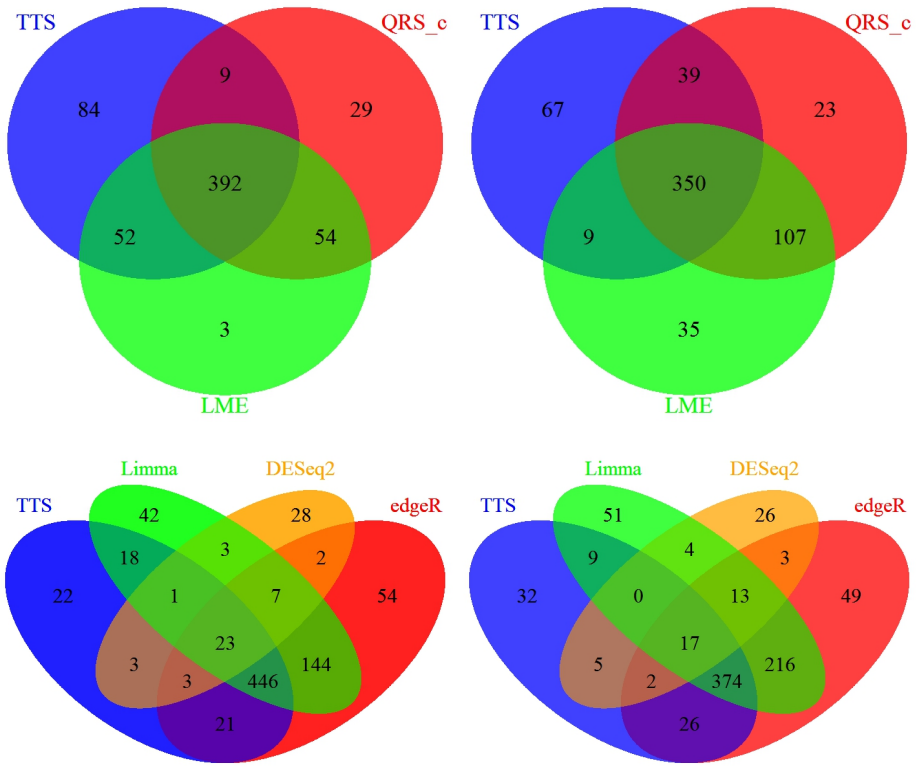

Figure 3.1. Venn diagram of number of overlapping genes among TTS, QRS_c, LME at top and TTS, $edgeR$, $DESeq2$, $Limma$ at bottom, for $\tau = 0.5$ at left and 0.75 at right.

of 0.03. In contrast, QRS_c detected only $SEMA3B$ and $SEMA3F$, where $SEMA3F$ was discovered with a modest FDR of 0.02. $DESeq2$ detected only $TUSC2$ with a modest FDR of 0.04.

To understand the discrepancy in the results between the methods, we first compared the results from our methods with those from conventional test methods, including QRS_c and LME. We plot the exon-level group differences at various covariate-adjusted quantiles for the genes $RASSF1$ and $SEMA3B$ in Figure 3.2.

It is not surprising that $SEMA3B$ could be detected by TTS, QRS_c, and LME due to its large group differences at most quantiles, including the median. QRS_c failed to detect $RASSF1$, which is understandable because of the trivial differences

Table 3.1
P-values of the six genes based on TTS, QRS_{cor}, and LME are reported. The detected genes with false discovery rates ≤ 0.05 are highlighted in blue.

Gene	TTS	QRS_c	LME	$Limma$	$edgeR$	$DESeq2$
$FHIT$	2.35e-03	3.81e-01	5.11e-02	1.16e-01	3.17e-03	3.18e-01
$RASSF1$	2.28e-19	7.61e-01	3.39e-06	9.10e-15	8.38e-15	8.40e-01
$TUSC2$	4.23e-01	3.38e-01	9.19e-01	9.93e-01	5.63e-01	4.33e-02
$SEMA3B$	9.30e-13	2.60e-14	3.10e-18	1.05e-17	4.50e-09	9.97e-01
$SEMA3F$	6.67e-02	2.00e-02	3.23e-02	5.82e-02	1.57e-01	7.85e-01
$MLH1$	9.91e-01	3.81e-01	2.92e-01	9.31e-01	4.75e-01	7.55e-01

between the normal tissue and cancer samples at the single point of the median. In contrast, TTS's ability to leverage the information across quantiles in the tail region substantially increased the detection power, since the upper quantiles show much larger group differences than the median. For example, the group differences at the median versus the 75% quantile were respectively 0.50 versus 0.72 for $RASSF1$.

Moreover, 32 other genes detected by TTS at $\tau = 0.5$ but not by QRS_c are likely associated with lung cancer according to the medical literature. The complete list of genes and their associated citations are presented in the upper part of Table 7.1 in Appendix.

Here are some examples. Expression of $FOXP1$ improves the survival rate of non-small cell lung cancer patients. $SIAH2$ suppresses lung carcinoma cells by antagonizing $TYK2 - STAT3$ signaling. $CTNNB1$ is involved in tumorigenesis of a subset of lung cancer. $GSK3B$ has been validated as a prognostic factor for lung carcinomas. Knockdown of VHL has been shown to promote epithelial-mesenchymal

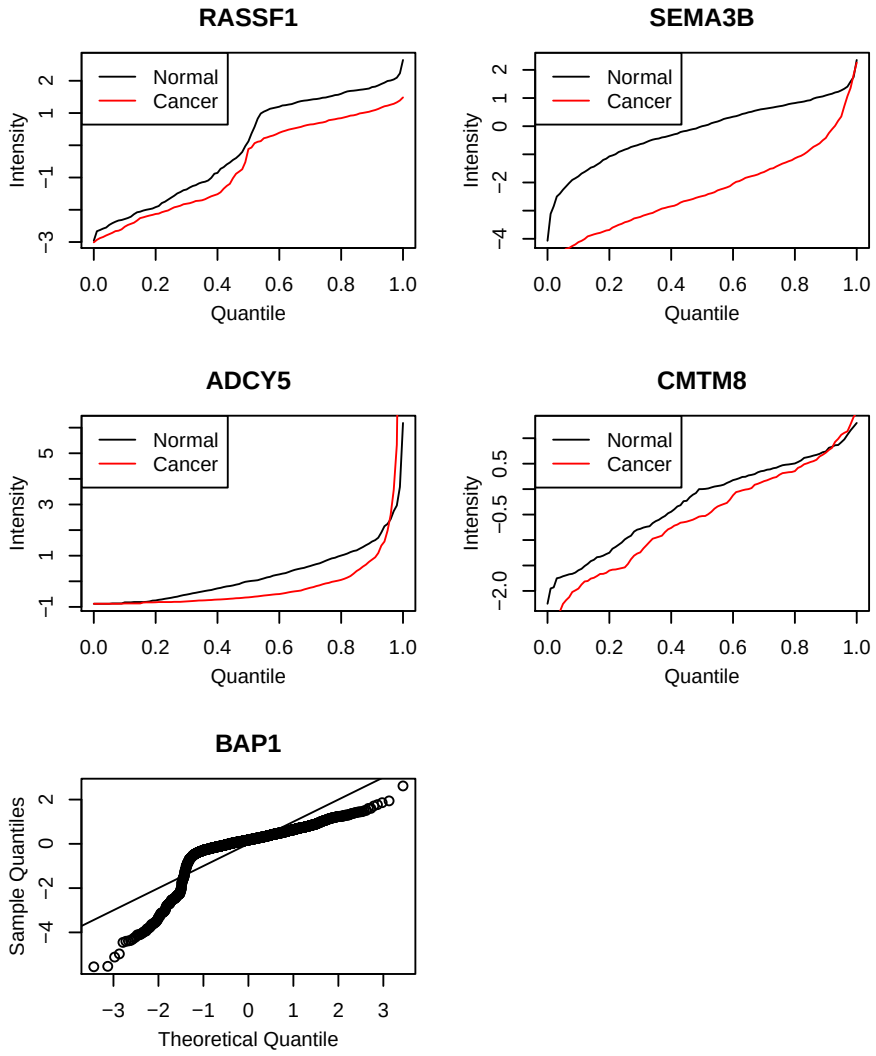

Figure 3.2. Top two rows are exon-level covariate-adjusted quantile intensity plots of normal tissue and cancer samples for genes $RASSF1$, $SEMA3B$, $ADCY5$, and $CMTM8$; bottom row is QQ-plot of the standardized residuals obtained from linear mixed model for gene $BAP1$.

transition in lung cancer cells, and $EAF2$ knockout has been found to cause lung adenocarcinoma.

We also looked into the genes that were detected by QRS_c but not by TTS, which account for 17% and 25% of genes detected by QRS_c at $\tau = 0.5$ and 0.75,

respectively. For example, with the FDR of 4.25×10^{-6}, QRS_c identified $ADCY5$ as being associated with lung adenocarcinoma. In Figure 3.2, we plot the group difference at various quantiles for $ADCY5$. We observe that the quantiles from cancer and normal tissue samples cross each other and the group differences are overturned in the upper tail region. As a result, QRS_c claims the group difference at the median. In contrast, TTS measures all the information across the quantiles in the upper tail region and concludes that the two groups are insignificantly different due to the offset of the opposite effects in the upper tail region.

In addition, 20 genes that were detected by TTS but not by LME have been shown to be associated with lung cancer in the literature. They are listed in the lower panel of Table 7.1 in Appendix.

Among these genes, $IQCB1$ displays patterns of alternative splicing in primary non-small cell lung tumors that are different from those of normal tissues. $RPL14$ has a lower heterozygous rate in non-small cell lung cancer cell lines compared to normal cells and has been shown to be a useful marker for lung cancer. Examination of human non-small cell lung cancer tissue shows positive correlation with $VPRBP$ expression.

We noticed that LME missed these genes mainly because of the violation of the required normal distribution assumption. As an example, we show the QQ-plot of the standardized residuals, obtained from linear mixed models, for $BAP1$ in the bottom row of Figure 3.2. It is clear that normality does not hold for this gene.

We also looked into the genes that were detected by LME but not by TTS at $\tau = 0.5$ and 0.75, which respectively account for 28% and 11% of genes detected by LME. For example, with the respective FDR of 0.0076, LME identified $CMTM8$ as being associated with lung adenocarcinoma. In the second row of Figure 3.2, we plot the group differences at various quantiles for $CMTM8$. We observe that the group difference is overall relatively small, especially the difference is gradually

diminishing in the upper tail region. Therefore, TTS concludes that the two groups are insignificantly different due to the modest difference in the upper tail region.

Then we compared the results of our method with those of standard DE analysis methods including *Limma*, *edgeR*, and $DESeq2$. Likely associated with lung cancer according to the medical literature are 15 genes detected by TTS at $\tau = 0.5$ but not by *Limma*, 12 genes detected by TTS at $\tau = 0.5$ but not by *edgeR*, and 143 genes detected by TTS at $\tau = 0.5$ but not by $DESeq2$. The complete list of genes and their literature citations are presented in Tables 7.2, 7.3, and 7.4 in Appendix. For example, $GSK3B$ is involved in the histogenesis of lung carcinomas, and its overexpression indicates worse prognosis in lung carcinoma. $SETD2$ is a potential tumor suppressor in lung adenocarcinoma and its inactivation has led to accelerated tumor progression. $TRIM59$ upregulates cell-cycle-related proteins to promote the proliferation and migration of non-small cell lung cancer cells.

We plot the group differences at various exon-level covariate-adjusted quantiles and gene-level read counts for the genes $TP63$ and $GSK3B$ in Figure 3.3. $TP63$ was detected by TTS but missed by *Limma* and $DESeq2$. $GSK3B$ was detected by TTS but missed by *edgeR* and $DESeq2$. Both $TP63$ and $GSK3B$ show trivial differences between the normal tissue and cancer samples before the median, which also causes shrinkage of the mean difference. Hence, the standard mean-based DE analysis methods are unable to detect these genes. In contrast, TTS's focus on the tail region substantially increased the detection power, since the upper quantile regions show much larger group differences than the mean.

We also looked into the genes that were detected by standard DE analysis methods but not by TTS. Genes that were detected by *Limma* but not by TTS account for 29% and 42% of genes detected by *edgeR* at $\tau = 0.5$ and 0.75, respectively. Genes that were detected by *edgeR* but not by TTS account for 30% and 40% of genes detected by *edgeR* at $\tau = 0.5$ and 0.75, respectively. Genes that were detected by $DESeq2$ but

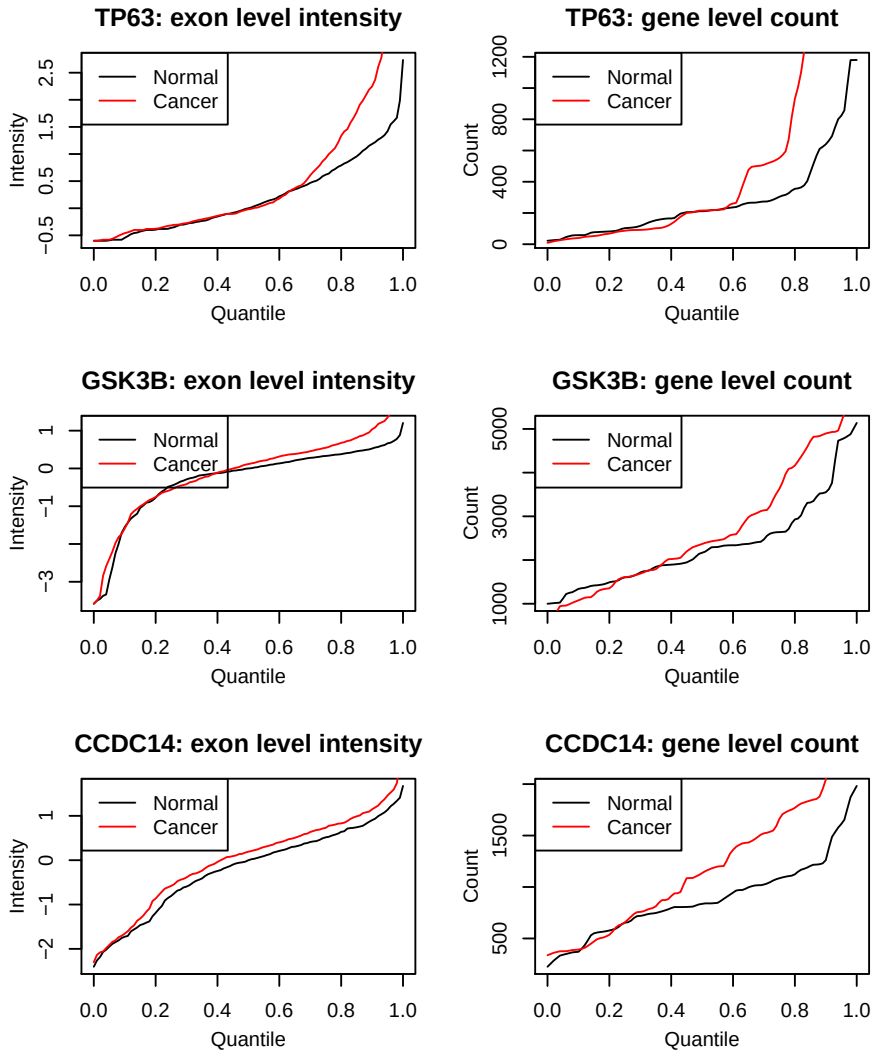

Figure 3.3. Left column: exon-level covariate-adjusted quantile intensity plots of normal tissue and cancer samples for genes $TP63$, $GSK3B$, and $CCDC14$; right column: gene-level read count quantile plot for the corresponding genes.

not by TTS account for 57% and 66% of genes detected by $DESeq2$ at $\tau = 0.5$ and 0.75, respectively. For example, $Limma$, $edgeR$, and $DESeq2$ identified $CCDC14$ with the respective FDRs of 0.018, 0.033, and 0.008. In Figure 3.3, we plot the group difference at various quantiles for $CCDC14$ regarding the exon-level covariate-

adjusted intensity and gene level read counts. We observe that the quantiles from cancer and normal tissue samples cross each other, and exon-level group differences are only modest across all quantiles, and the difference is larger at gene-level. As a result, *Limma*, *edgeR*, and *DESeq*2 claim a group difference. However, TTS concludes that the two groups are insignificantly different due to the modest difference in the upper tail region.

In summary, TTS shows better performance than QRS_c and LME due to its ability to utilize all the information in the upper quantile region and its robustness to model distributions and individual outliers. TTS is also a good supplement method to use along with standard DE methods, as it is able to include potential biomarkers that are missed by *Limma*, *edgeR*, and *DESeq*2. Our proposed method can detect many exclusive genes when there are consistent and considerable differences between two groups across the upper quantile region. TTS loses it power advantage when the group difference is overturned or is very modest in the upper tail region, but those are cases in which caution must be exercised when inferring statistical significance from other tests. Overall, our proposed method offers a powerful and robust supplement for biomarker discovery by utilizing the information in the whole region of interest.

4. A tail-based test for pathway analysis in RNA-sequencing data

4.1 Methodology

In biomedical applications of genome-wide expression studies, pathway analysis, rather than individual gene analysis, is gaining popularity. In pathway analysis, we incorporate known biological information of pathways to generate gene set, and then test whether the generated gene set of interest has differential expression between disease groups. The proposed method is devised to meet this objective. The test first conducts TTS test on individual gene in the pathway to obtain the TTS test statistics, then combine the individual TTS test statistics and compute the pathway test statistics P_{TTS}.

Since we are introducing the gene set concept, we redefine the notations to add a new layer for the TTS test. Let $\boldsymbol{Z}$ denote the gene expression intensity, which is treated as the response measure, wherein Z_{ijk} indicates the intensity measurement of the kth exon location in jth gene of interest for the ith sample. We use a dummy variable $D = 0, 1$ to denote the control and diseased patient groups, respectively, wherein D_i corresponds to the disease status of sample i. We use $\mathbf{C}$ to indicate P covariates and assume them to be independent of D, and a $P \times 1$ design vector $\boldsymbol{C}_i$ corresponding to the covariates with sample i. The integers n_0 and n_1 respectively indicate the number of patient samples for the groups of $D = 0$ and $D = 1$, and $n = n_0 + n_1$. We use m_j to denote the total number of exon locations belonging to the jth gene in one sample. We also use $N_{j,d}$ to denote the total number of exon

locations belonging to the jth gene in the target group of $D = 0$ and $D = 1$, and N_j to denote the total number of exon locations of all samples in the jth gene.

We express the τth quantile of $\boldsymbol{Z}$, given $\boldsymbol{D}$, $\boldsymbol{C}$, as

$$Q_{\boldsymbol{Z}}(\tau \mid \boldsymbol{D}, \mathbf{C}) = \alpha(\tau) + \boldsymbol{D}\delta(\tau) + \mathbf{C}\boldsymbol{\gamma}(\tau) = X\boldsymbol{\beta}(\tau), \tag{4.1}$$

where $X = (\mathbf{1}_{N\times 1}, \boldsymbol{D}_{N\times 1}, \boldsymbol{C}_{N\times P})$ and $\boldsymbol{\beta}(\tau) = (\alpha(\tau), \delta(\tau), \boldsymbol{\gamma}(\tau)^T_{P\times 1})^T$.
We perform the Gram-Schmidt orthogonalization for $\boldsymbol{D}$ and $\boldsymbol{C}$ to $\boldsymbol{D}^* = \boldsymbol{D} - \bar{D}$ and $\boldsymbol{C}^* = \boldsymbol{C} - n_d^{-1}\sum_i \boldsymbol{C}_i I(D_i = d)$. $\bar{D}$ is the overall mean for D.

Correspondingly, the model for the jth gene intensity measure Z_{ijk} can be written as

$$Z_{ijk} = \alpha_j(\tau) + D_i\delta(\tau) + \mathbf{C_i}^T\boldsymbol{\gamma}(\tau) + e_{ijk}(\tau), \tag{4.2}$$

where the residuals $e_{ijk}(\tau)$ have the value of 0 as the τth conditional. We assume the following correlation: (1) the inter-exon correlation satisfies $cov(e_{ijk}, e_{ijk'}) \neq 0$ where $k \neq k'$. (2) the gene-wise correlation satisfies $cov(e_{ijk_1}, e_{ij'k_2}) \neq 0$ where $j \neq j'$ and for all k. (3) no sample-wise correlation such that $cov(e_{ijk}, e_{i'j'k'}) = 0$ where $i \neq i'$. (4) the inter-exon correlation and the gene-wise correlation are compound symmetry.

Given $(Z_{ijk}, D_i, \mathbf{C_i})$, we obtain the estimate $\hat{\alpha}(\tau), \hat{\delta}(\tau), \hat{\boldsymbol{\gamma}}(\tau)$ at the τth quantile via quantile regression (Koenke et al., 1978). We denote the corresponding empirical residuals as $\hat{e}_{ijk}(\tau) = Z_{ijk} - \hat{\alpha}(\tau) - D_i\hat{\delta}(\tau) - \mathbf{C_i}^T\hat{\boldsymbol{\gamma}}(\tau)$.
To detect the between-group difference in the gene expression intensity, we define a new tail-based test statistic (TTS) as follows:

$$T_\tau^{TTS}(n_1, n_0) = TTS_\tau(1) - TTS_\tau(0), \tag{4.3}$$

where $TTS_\tau(d) = \sum_{D_i=d}\sum_{k=1}^{m_i} w_{d,i,j}[Z_{ij} - \mathbf{C_i}^T\hat{\boldsymbol{\gamma}}(\tau)], d = 0, 1$. Let $e_{ij}^+ = I(e_{ijk} > 0)$ and $e_{ij}^- = I(e_{ijk} < 0)$. Herein, $w_{d,i,j,k} = S_d^{-1}\hat{e}_{ijk}^+(\tau)$, $S_d = \sum_{D_i=d}\sum_{k=1}^{m_i}\hat{e}_{ijk}^+(\tau)$, and $w_{d,i,j,k}$ serves as a weight for the ith sample at the jth gene and kth exon location within group $d = 0$ or 1.

Note that $TTS_\tau(d)$ includes the information on residual directions and covariate adjusted residuals, and hence measures the average expression intensity above the τth quantile in group d after adjusting for the covariates.

Let $\bar{D}_\tau(d)$, $\bar{\boldsymbol{C}}_\tau(d)$, and $\bar{e}_\tau(d)$ be the averages of all the D_i, $\boldsymbol{C}_i$, and e_{ijk}, respectively, in group d that are above the τ-th conditional quantile.
Specifically, $\bar{\boldsymbol{C}}_\tau(d) = S_d^{-1} \sum_{D_i=d} \sum_k^{m_i} \boldsymbol{C}_i \hat{e}_{ijk}^+(\tau)$,
and $\bar{e}_\tau(d) = S_d^{-1} \sum_{D_i=d} \sum_k^{m_j} [Z_{ijk} - \alpha(\tau) - D_i\delta(\tau) - \mathbf{C_i}^T\boldsymbol{\gamma}(\tau)]\hat{e}_{ijk}^+(\tau)$.

Replacing Z_{ijk} with $e_{ijk}(\tau) + \alpha(\tau) + D_i\delta(\tau) + \mathbf{C_i}^T\boldsymbol{\gamma}(\tau)$ in T_τ^{TTS}, we can express the individual gene test statistic as

$$T_\tau^{TTS}(n_1, n_0) = \delta(\tau) - (\bar{\boldsymbol{C}}_\tau^T(1) - \bar{\boldsymbol{C}}_\tau^T(0))(\hat{\boldsymbol{\gamma}}(\tau) - \boldsymbol{\gamma}(\tau)) + (\bar{e}_\tau(1) - \bar{e}_\tau(0)). \tag{4.4}$$

To perform the test, we establish the asymptotic distribution of $T_\tau^{TTS}(n_1, n_0)$ as $n_0, n_1 \to \infty$ under the null hypothesis of no difference between the two groups. We first estimate the conditional density function f_{ijk} of e_{ijk} given (D_i, $\mathbf{C_i}$) evaluated at 0, denoted as $\hat{f}_{n(0)}$. Then, we let $(U_f)_{P\times P} = \sum_i \hat{f}_{n(0)} \mathbf{C^*}_i \mathbf{C^*}_i^T$ in which U_f is a combination of the f_{ijk} and can be estimated consistently even when the conditional densities vary with C_i (He et al., 2010). We also denote the transformed D and C via Gram-Schmidt orthogonalization as follows,

$$D_i^* = D_i - n_d^{-1} \sum_i D_i I(D_i = d) \tag{4.5}$$

$$\boldsymbol{C}_i^* = \boldsymbol{C}_i - n_d^{-1} \sum_i \boldsymbol{C}_i I(D_i = d), \tag{4.6}$$

Note that, when conducted on each individual gene, this model is essentially the same as the model in Section 2.1 in this paper. We can use the Lemma 2.1.1 and Theorem 2.1.1 and convert them into Lemma 4.1.1 and Theorem 4.1.1 with some notation adjustments.

After we calculate the TTS test statistics for each gene in the gene set, we let $X = (T_{\tau,1}^{TTS}, ..., T_{\tau,j}^{TTS})$ be the vector of TTS test statistics on j genes. Let Σ denote the covariance matrix of the test statistics. In Lemma 4.1.2, we obtain $X^* = \Sigma^{-1/2}X \sim N_n(0, \mathbf{I_{j\times j}})$, and $T_{\tau,j}^{TTS^*}$ in $X^* = (T_{\tau,1}^{TTS^*}, ..., T_{\tau,j}^{TTS^*})$ follows a standard normal distribution.

The proposed tail-based pathway test statistics P_{TTS} for the hypothesis of no difference between disease groups is defined as

$$P_{TTS} = \sqrt{j}\bar{T}_{\tau}^{TTS^*}, \text{ with } \bar{T}_{\tau}^{TTS^*} = \frac{1}{j}\sum_{p=1}^{j} T_{\tau,p}^{TTS^*}, \tag{4.7}$$

with j being the number of genes in pathway of interest. We use Theorem 4.1.2 to establish the standard normal distribution of P_{TTS} under null hypothesis.

Lemma 4.1.1

If $\lim_{n_1,n_0\to\infty}(n_1+n_0)^{-1}U_f$ *exists,* $E\|C_i\|_1^3 < \infty$*, the number of exon region* m_i *is some fixed number, and* f_{ijk} *are uniformly bounded away from 0 and infinity, then we have the Bahadur representation on* $\hat{\boldsymbol{\gamma}}(\tau)$*,*

$$\hat{\boldsymbol{\gamma}}(\tau) - \boldsymbol{\gamma}(\tau) = U_f^{-1}\textstyle\sum_i m_j^{-1}\sum_{j=1}^{m_j}\boldsymbol{C}_i^*\psi_\tau(e_{ijk}(\tau)) + o_p((n_0+n_1)^{-\frac{1}{2}}),$$

and the representation of $\bar{e}_\tau(d)$*,*

$$\bar{e}_\tau(d) = (\textstyle\sum_{D_i=d}\sum_j^{m_j} e_{ijk}^+(\tau))^{-1}\sum_{D_i=d}\sum_j^{m_j} e_{ijk}(\tau)e_{ijk}^+(\tau) + o_p((n_0+n_1)^{-\frac{1}{2}}).$$

Proof of Lemma 4.1.1

Refer to proof of Lemma 2.1.1.

Theorem 4.1.1

If $\lim_{n_1,n_0\to\infty}\frac{n_0}{n_0+n_1} \to q \in (0,1)$ *and* $\lim_{n_1,n_0\to\infty}(n_1+n_0)^{-1}U_f$ *exists,* $E\|\boldsymbol{C}_i\|_1^3 < \infty$*, and* f_{ijk} *are uniformly bounded away from 0 and infinity, then under the null hypothesis, in which the distribution of the two groups* $F_{Z|\mathbf{C},D=1}$=$F_{Z|\mathbf{C},D=0}$*, we have*

$$T_{\tau}^{TTS}(n_1,n_0)/s_{n_0,n_1} \to N(0,1) \text{ as } n_1, n_0 \to \infty. \tag{4.8}$$

Proof of Theorem 4.1.1

According to Lemma 4.1.1, and $\delta(\tau) = 0$, under the null hypothesis, we can write

$$\begin{aligned} T_\tau^{TTS} = & \left\{ \sum_{D_i=1} \sum_{k=1}^{m_j} e_{ijk}(\tau) e_{ijk}^+(\tau)/N_1 - \sum_{D_i=0} \sum_{k=1}^{m_j} e_{ijk}(\tau) e_{ijk}^+(\tau)/N_0 \right\} (1-\tau)^{-1} \\ & -\{\bar{\boldsymbol{C}}_\tau^T(1) - \bar{\boldsymbol{C}}_\tau^T(0)\} U_f^{-1} \sum_{i=1}^{n} m_j^{-1} \sum_{k=1}^{m_j} \boldsymbol{C}_i^* \psi_\tau(e_{ijk}(\tau)) + o_p((n_0+n_1)^{-1/2}) \\ = & \quad T_\tau^*(n_1, n_0) + o_p((n_0+n_1)^{-1/2}). \end{aligned}$$

where

$$\begin{aligned} T_\tau^*(n_1, n_0) = & \left\{ \sum_{D_i=1} \sum_{k=1}^{m_j} e_{ijk}(\tau) e_{ijk}^+(\tau)/N_1 - \sum_{D_i=0} \sum_{k=1}^{m_j} e_{ijk}(\tau) e_{ijk}^+(\tau)/N_0 \right\} (1-\tau)^{-1} \\ & -\{\bar{\boldsymbol{C}}_\tau^T(1) - \bar{\boldsymbol{C}}_\tau^T(0)) U_f^{-1} \sum_{i=1}^{n} m_j^{-1} \sum_{k=1}^{m_j} \boldsymbol{C}_i^* \psi_\tau(e_{ijk}(\tau)\} \end{aligned}$$

Under the null hypothesis, the mean of the test statistics are

$$\begin{aligned} E(T_\tau^*(n_1, n_0)) = & \left\{ \sum_{D_i=1} \sum_{k=1}^{m_j} E(e_{ijk}(\tau) e_{ijk}^+)/N_1 - \sum_{D_i=0} \sum_{k=1}^{m_j} E(e_{ijk}(\tau) e_{ijk}^+)/N_0 \right\} (1-\tau)^{-1} \\ & = (1-\tau)^{-1} E(e_{ijk}(\tau) e_{ijk}^+)(1-1) \\ & = 0. \end{aligned}$$

And the variance are

$$
\begin{aligned}
&Var(T^*_\tau(n_1,n_0)) = \\
&(1-\tau)^{-2}(V_1/N_1^2 + V_0/N_0^2) \\
&+\{\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0)\}U_f^{-1}\left\{\sum_i m_k^{-2}\sum_{k=1}^{m_j}\boldsymbol{C}^*_i\boldsymbol{C}^{*T}_i\tau(1-\tau)\right\} \\
&\times U_f^{-1}\{\bar{\boldsymbol{C}}_\tau(1) - \bar{\boldsymbol{C}}_\tau(0)\} \\
&+\{\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0)\}U_f^{-1}\left\{\sum_i m_j^{-2}\sum_{k\neq k'}\boldsymbol{C}^*_i\boldsymbol{C}^{*T}_i\psi_\tau(e_{ijk}(\tau))\psi_\tau(e_{ijk'}(\tau))\right\} \\
&\times U_f^{-1}\{\bar{\boldsymbol{C}}_\tau(1) - \bar{\boldsymbol{C}}_\tau(0)\} \\
&+(1-\tau)^{-1}\left\{\sum_{D_i=1}\sum_{k=1}^{m_j} e_{ijk}(\tau)e^+_{ijk}(\tau)/N_1\right\}\{\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0)\}U_f^{-1} \\
&\times\sum_{D_i=1}^{n} m_j^{-1}\sum_{k=1}^{m_j}\boldsymbol{C}^*_i\psi_\tau(e_{ijk}(\tau)) \\
&-(1-\tau)^{-1}\left\{\sum_{D_i=0}\sum_{k=1}^{m_j} e_{ijk}(\tau)e^+_{ijk}(\tau)/N_0\right\}\{\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0)\}U_f^{-1} \\
&\times\sum_{D_i=0}^{n} m_j^{-1}\sum_{k=1}^{m_j}\boldsymbol{C}^*_i\psi_\tau(e_{ijk}(\tau)) \\
&-(1-\tau)^{-1}\{\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0)\}U_f^{-1}/N_1\sum_{D_i=1}^{n}\sum_{k=1}^{m_j}\boldsymbol{C}^*_i m_j^{-1}e_{ijk}e^+_{ijk}\psi_\tau(e_{ijk}(\tau)) \\
&+(1-\tau)^{-1}\{\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0)\}U_f^{-1}/N_0\sum_{D_i=0}^{n}\sum_{k=1}^{m_j}\boldsymbol{C}^*_i m_j^{-1}e_{ijk}e^+_{ijk}\psi_\tau(e_{ijk}(\tau)) \\
&= (1-\tau)^{-2}(V_1/N_1^2 + V_0/N_0^2) \\
&+\{\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0)\}U_f^{-1}\left\{\sum_i m_k^{-2}\sum_{k=1}^{m_j}\boldsymbol{C}^*_i\boldsymbol{C}^{*T}_i\tau(1-\tau)\right\}U_f^{-1}\{\bar{\boldsymbol{C}}_\tau(1) - \bar{\boldsymbol{C}}_\tau(0)\} \\
&+\{\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0)\}U_f^{-1}\left\{\sum_i m_j^{-2}\sum_{k\neq k'}\boldsymbol{C}^*_i\boldsymbol{C}^{*T}_i\psi_\tau(e_{ijk}(\tau))\psi_\tau(e_{ijk'}(\tau))\right\} \\
&\times U_f^{-1}\{\bar{\boldsymbol{C}}_\tau(1) - \bar{\boldsymbol{C}}_\tau(0)\} \\
&-(1-\tau)^{-1}\{\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0)\}U_f^{-1}/N_1
\end{aligned}
$$

$$
\begin{aligned}
&\times\Bigg\{\sum_{D_i=1}^{n}\sum_{k_1=1}^{m_j}\sum_{k_2=1}^{m_j}\boldsymbol{C}_i^{*}m_j^{-1}e_{ijk_1}e_{ijk_1}^{+}\psi_\tau(e_{ijk_2}(\tau))\\
&-\sum_{D_i=1}^{n}\sum_{k_1=1}^{m_j}e_{ijk_1}e_{ijk_1}^{+}\sum_{k_2=1}^{m_j}m_j^{-1}\boldsymbol{C}_i^{*}\psi_\tau(e_{ijk_2}(\tau))\Bigg\}\\
&+(1-\tau)^{-1}(\bar{\boldsymbol{C}}_\tau^{T}(1)-\bar{\boldsymbol{C}}_\tau^{T}(0))U_f^{-1}/N_0\\
&\times\Bigg\{\sum_{D_i=0}^{n}\sum_{k_1=1}^{m_j}\sum_{k_2=1}^{m_j}\boldsymbol{C}_i^{*}m_j^{-1}e_{ijk_1}e_{ijk_1}^{+}\psi_\tau(e_{ijk_2}(\tau))\\
&-\sum_{D_i=0}^{n}\sum_{k_1=1}^{m_j}e_{ijk_1}e_{ijk_1}^{+}\sum_{k_2=1}^{m_j}m_j^{-1}\boldsymbol{C}_i^{*}\psi_\tau(e_{ijk_2}(\tau))\Bigg\}
\end{aligned}
$$

where

$$
\begin{aligned}
V_d = \quad & \sum_{D_i=d}\sum_{k=1}^{m_j}var(e_{ijk}e_{ijk}^{+})+\sum_{D_i=d}\sum_{k\neq k'}cov(e_{ij}e_{ijk}^{+},e_{ijk'}e_{ijk'}^{+}),\\
= \quad & \sum_{D_i=d}\sum_{k=1}^{m_j}\left(e_{ijk}^{2}\hat{e}_{ijk}^{+}\right)-N_d^{-1}\left(\sum_{D_i=d}\sum_{k=1}^{m_j}e_{ijk}e_{ijk}^{+}\right)^{2}\\
& +\sum_{D_i=d}\sum_{k\neq k'}\Bigg[\left\{\sum_{D_i=d}m_j(m_j-1)\right\}^{-1}\sum_{D_i=d}\sum_{k\neq k'}e_{ijk}e_{ijk}^{+}e_{ijk'}e_{ijk'}^{+}\\
& -{n_d}^{-1}\left\{\sum_{D_i=d}\sum_{k}e_{ijk}e_{ijk}^{+}\right\}^{2}\Bigg],
\end{aligned}
$$

which can be estimated by $s^2_{n_0,n_1}$.

$$
\begin{aligned}
s^2_{n_0,n_1} &= \{1-\tau)^{-2}\left(V_1/\left(\sum_{D_i=1} m_j\right\}^2 + V_0/\left(\sum_{D_i=0} m_j\right)^2\right) \\
&+\{\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0)\}U_f^{-1}\left\{\sum_i m_k^{-2}\sum_{k=1}^{m_j}\boldsymbol{C}^*_i\boldsymbol{C}^{*T}_i\tau(1-\tau)\right\}U_f^{-1}\{\bar{\boldsymbol{C}}_\tau(1) - \bar{\boldsymbol{C}}_\tau(0)\} \\
&+\{\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0)\}U_f^{-1}\left\{\sum_i m_j^{-2}\sum_{k\neq k'}\boldsymbol{C}^*_i\boldsymbol{C}^{*T}_i\psi_\tau(\hat{e}_{ijk}(\tau))\psi_\tau(\hat{e}_{ijk'}(\tau))\right\} \\
&\times U_f^{-1}\{\bar{\boldsymbol{C}}_\tau(1) - \bar{\boldsymbol{C}}_\tau(0)\} \\
&-(1-\tau)^{-1}\{\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0)\}U_f^{-1}/N_1\left\{\sum_{D_i=1}^{n}\sum_{k_1=1}^{m_j}\sum_{k_2=1}^{m_j}\boldsymbol{C}^*_i m_j^{-1}\hat{e}_{ijk_1}\hat{e}^+_{ijk_1}\psi_\tau(\hat{e}_{ijk_2}(\tau))\right. \\
&\left.-\sum_{D_i=1}^{n}\sum_{k_1=1}^{m_j}\hat{e}_{ijk_1}\hat{e}^+_{ijk_1}\sum_{k_2=1}^{m_j} m_j^{-1}\boldsymbol{C}^*_i\psi_\tau(\hat{e}_{ijk_2}(\tau))\right\} \\
&+(1-\tau)^{-1}(\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0))U_f^{-1}/N_0\left\{\sum_{D_i=0}^{n}\sum_{k_1=1}^{m_j}\sum_{k_2=1}^{m_j}\boldsymbol{C}^*_i m_j^{-1}\hat{e}_{ijk_1}\hat{e}^+_{ijk_1}\psi_\tau(\hat{e}_{ijk_2}(\tau))\right. \\
&\left.-\sum_{D_i=0}^{n}\sum_{k_1=1}^{m_j}\hat{e}_{ijk_1}\hat{e}^+_{ijk_1}\sum_{k_2=1}^{m_j} m_j^{-1}\boldsymbol{C}^*_i\psi_\tau(\hat{e}_{ijk_2}(\tau))\right\}
\end{aligned}
$$

By the central limit theorem, $T^*_\tau(n_1,n_0)$ is asymptotically normal with mean 0 and variance. Thus, by lemma 4.1.1 and $T_\tau(n_1,n_0) - T^*_\tau(n_1,n_0) = o_p((n_0+n_1)^{-1/2})$, we prove the asymptotic normality of the test statistic $T_\tau(n_1,n_0)$.

To construct the pathway test, we let $T_{\tau,j}^{TTS}$ denote the test statistics for jth gene, $Var(T_{\tau,j}^{TTS})$ denotes the variance for jth gene, $Cov(T_{\tau,j_1}^{TTS}, T_{\tau,j_2}^{TTS})$ denotes the covariance for the test statistics between j_1th and j_2th gene.

Lemma 4.1.2

Let $X = (T_{\tau,1}^{TTS}, ..., T_{\tau,j}^{TTS})$ be the vector of the TTS test statistics on j genes from a gene set. Let Σ denote the covariance matrix of the test statistics.

For $X \sim N_n(0, \Sigma)$ under the null hypothesis and Σ is positive definite, we have $X^* = \Sigma^{-1/2} X \sim N_n(0, \mathbf{I_{j \times j}})$, and $T_{\tau,j}^{TTS^*}$ in $X^* = (T_{\tau,1}^{TTS^*}, ..., T_{\tau,j}^{TTS^*})$ follows a standard normal distribution.

Proof of Lemma 4.1.2

According to Theorem 4.1.1 and the gene-wise correlation structure specified in equation (4.2), $X \sim N_n(0, \Sigma)$ under null hypothesis. The covariance matrix of the test statistics vector X is as following

$$\Sigma = \begin{pmatrix} Var(T_{\tau,1}^{TTS}) & Cov(T_{\tau,j_1}^{TTS}, T_{\tau,j_2}^{TTS}) & ... & ... \\ Cov(T_{\tau,j_2}^{TTS}, T_{\tau,j_1}^{TTS}) & Var(T_{\tau,2}^{TTS}) & ... & ... \\ ... & ... & ... & ... \\ ... & ... & ... & Var(T_{\tau,j}^{TTS}) \end{pmatrix}$$

where the variances $Var(T_{\tau,j}^{TTS})$ have been derived in Theorem 4.1.1. and

pairwise covariance is as following

$$
\begin{aligned}
&Cov(T^{TTS}_{\tau,j}, T^{TTS}_{\tau,j'}) \\
&= E\left(T^{TTS}_{\tau,j} T^{TTS}_{\tau,j'}\right) - E\left(T^{TTS}_{\tau,j}\right) E\left(T^{TTS}_{\tau,j'}\right) \\
&= E\Bigg(\Bigg[\Bigg\{ \sum_{D_i=1} \sum_{k_1=1}^{m_j} e_{ijk_1}(\tau) e^{+}_{ijk_1}(\tau)/N_1 - \sum_{D_i=0} \sum_{k_1=1}^{m_j} e_{ijk_1}(\tau) e^{+}_{ijk_1}(\tau)/N_0 \Bigg\} (1-\tau)^{-1} \\
&- \{\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0)\} U_f^{-1} \sum_{i=1}^{n} m_j^{-1} \sum_{k_1=1}^{m_j} \boldsymbol{C}^*_i \psi_\tau(e_{ijk_1}(\tau)) \Bigg] \\
&\times \Bigg[\Bigg\{ \sum_{D_i=1} \sum_{k_2=1}^{m'_j} e_{ij'k_2}(\tau) e^{+}_{ij'k_2}(\tau)/N_1 - \sum_{D_i=0} \sum_{k_2=1}^{m'_j} e_{ij'k_2}(\tau) e^{+}_{ij'k_2}(\tau)/N_0 \Bigg\} (1-\tau)^{-1} \\
&- \{\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0)\} U_f^{-1} \sum_{i=1}^{n} m_j'^{-1} \sum_{k_2=1}^{m'_j} \boldsymbol{C}^*_i \psi_\tau(e_{ij'k_2}(\tau)) \Bigg] \Bigg) \\
&- E\Bigg[\Bigg\{ \Big(\sum_{D_i=1} \sum_{k_1=1}^{m_j} e_{ijk_1}(\tau) e^{+}_{ijk_1}(\tau)/N_1 - \sum_{D_i=0} \sum_{k_1=1}^{m_j} e_{ijk_1}(\tau) e^{+}_{ijk_1}(\tau)/N_0 \Bigg\} (1-\tau)^{-1} \\
&- \{\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0)\} U_f^{-1} \sum_{i=1}^{n} m_j^{-1} \sum_{k_1=1}^{m_j} \boldsymbol{C}^*_i \psi_\tau(e_{ijk_1}(\tau)) \Bigg] \\
&\times E\Bigg[\Bigg\{ \sum_{D_i=1} \sum_{k_2=1}^{m'_j} e_{ij'k_2}(\tau) e^{+}_{ij'k_2}(\tau)/N_1 - \sum_{D_i=0} \sum_{k_2=1}^{m'_j} e_{ij'k_2}(\tau) e^{+}_{ij'k_2}(\tau)/N_0 \Bigg\} (1-\tau)^{-1} \\
&- \{\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0)\} U_f^{-1} \sum_{i=1}^{n} m_j'^{-1} \sum_{k_2=1}^{m'_j} \boldsymbol{C}^*_i \psi_\tau(e_{ij'k_2}(\tau)) \Bigg] \\
&= E(A_1B_1 - A_1B_2 - A_1B_3 - A_2B_1 + A_2B_2 + A_2B_3 - A_3B_1 + A_3B_2 + A_3B_3) - \\
&\{E(A_1)E(B_1) - E(A_1)E(B_2) - E(A_1)E(B_3) - E(A_2)E(B_1) + E(A_2)E(B_2) + \\
&+E(A_2)E(B_3) - E(A_3)E(B_1) + E(A_3)E(B_2) + E(A_3)E(B_3))
\end{aligned}
$$

Where $A_1 = (1-\tau)^{-1} \sum_{D_i=1} \sum_{k_1=1}^{m_j} e_{ijk_1}(\tau) e^{+}_{ijk_1}(\tau)/N_1$

$A_2 = (1-\tau)^{-1} \sum_{D_i=0} \sum_{k_1=1}^{m_j} e_{ijk_1}(\tau) e^{+}_{ijk_1}(\tau)/N_0$

$A_3 = (\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0)) U_f^{-1} \sum_{i=1}^{n} m_j^{-1} \sum_{k_1=1}^{m_j} \boldsymbol{C}^*_i \psi_\tau(e_{ijk_1}(\tau))$

$$B_1 = (1-\tau)^{-1}\sum_{D_i=1}\sum_{k_2=1}^{m'_j} e_{ij'k_2}(\tau)e^+_{ij'k_2}(\tau)/N_1$$
$$B_2 = (1-\tau)^{-1}\sum_{D_i=0}\sum_{k_2=1}^{m'_j} e_{ij'k_2}(\tau)e^+_{ij'k_2}(\tau)/N_0$$
$$B_3 = (\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0))U_f^{-1}\sum_{i=1}^n m_j'^{-1}\sum_{k_2=1}^{m'_j}\boldsymbol{C}^*_i\psi_\tau(e_{ij'k_2}(\tau))$$

$$\begin{aligned}
& E(A_1B_1) - E(A_1)E(B_1) \\
= \;& E\left\{(1-\tau)^{-2}/N_1^2\sum_{D_i=1}\sum_{k_1=1}^{m_j} e_{ijk_1}(\tau)e^+_{ijk_1}(\tau)\sum_{D_i=1}\sum_{k_2=1}^{m'_j} e_{ij'k_2}(\tau)e^+_{ij'k_2}(\tau)\right\} \\
& -(1-\tau)^{-2}/N_1^2\sum_{D_i=1}\sum_{k_1=1}^{m_j} E\left\{e_{ijk_1}(\tau)e^+_{ijk_1}(\tau)\right\}\sum_{D_i=1}\sum_{k_2=1}^{m'_j} E\left\{e_{ij'k_2}(\tau)e^+_{ij'k_2}(\tau)\right\} \\
= \;& (1-\tau)^{-2}/N_1^2\Bigg[\sum_{i_1=i_2}^{D_i=1}\sum_{k_1=1}^{m_j}\sum_{k_2=1}^{m'_j} E\left\{e_{ijk_1}(\tau)e^+_{ij1k_1}(\tau)e_{ij'k_2}(\tau)e^+_{ij'k_2}(\tau)\right\} \\
& +\sum_{i_1\neq i_2}^{D_i=1}\sum_{k_1=1}^{m_j} E\left\{e_{i_1jk_1}(\tau)e^+_{ij1k_1}(\tau)\right\}\sum_{k_2=1}^{m'_j} E\left\{e_{i_2j'k_2}(\tau)e^+_{ij'k_2}(\tau)\right\}\Bigg] \\
& -(1-\tau)^{-2}/N_1^2\left[\sum_{i_1}\sum_{i_2}\sum_{k_1=1}^{m_j} E\{e_{i_1jk_1}(\tau)e^+_{ij1k_1}(\tau)\}\sum_{k_2=1}^{m'_j} E\{e_{i_2j'k_2}(\tau)e^+_{ij'k_2}(\tau)\}\right] \\
= \;& (1-\tau)^{-2}/N_1^2\Bigg[\sum_{i_1=i_2}^{D_i=1}\sum_{k_1=1}^{m_j}\sum_{k_2=1}^{m'_j} E\{e_{ijk_1}(\tau)e^+_{ij1k_1}(\tau)e_{ij'k_2}(\tau)e^+_{ij'k_2}(\tau)\} \\
& -\sum_{i_1=i_2}^{D_i=1}\sum_{k_1=1}^{m_j} E\{e_{ijk_1}(\tau)e^+_{ij1k_1}(\tau)\}\sum_{k_2=1}^{m'_j} E\{e_{ij'k_2}(\tau)e^+_{ij'k_2}(\tau)\}\Bigg]
\end{aligned}$$

$$\begin{aligned}
& E(A_1B_2) - E(A_1)E(B_2) \\
= \;& (1-\tau)^{-2}/(N_0N_1)E\left\{\sum_{D_i=1}\sum_{k_1=1}^{m_j} e_{ijk_1}(\tau)e^+_{ijk_1}(\tau)\sum_{D_i=0}\sum_{k_2=1}^{m'_j} e_{ij'k_2}(\tau)e^+_{ij'k_2}(\tau)\right\} - \\
& (1-\tau)^{-2}/(N_0N_1)\sum_{D_i=1}\sum_{k_1=1}^{m_j} E\left[e_{ijk_1}(\tau)e^+_{ijk_1}(\tau)\right]\sum_{D_i=0}\sum_{k_2=1}^{m'_j} E\left\{e_{ij'k_2}(\tau)e^+_{ij'k_2}(\tau)\right\} \\
= \;& 0
\end{aligned}$$

$$
\begin{aligned}
& E(A_1B_3) - E(A_1)E(B_3) \\
= \; & E\Big\{(1-\tau)^{-1}/(N_1)\sum_{D_i=1}\sum_{k_1=1}^{m_j} e_{ijk_1}(\tau)e^{+}_{ijk_1}(\tau) \\
& \times(\bar{\boldsymbol{C}}^T_\tau(1)-\bar{\boldsymbol{C}}^T_\tau(0))U_f^{-1}\sum_{i=1}^{n} m_j'^{-1}\sum_{k_2=1}^{m'_j}\boldsymbol{C}^*_i\psi_\tau(e_{ij'k_2}(\tau))\Big\} \\
& -(1-\tau)^{-1}/(N_1)\sum_{D_i=1}\sum_{k_1=1}^{m_j} E\{e_{ijk_1}(\tau)e^{+}_{ijk_1}(\tau)\} \\
& \times\{\bar{\boldsymbol{C}}^T_\tau(1)-\bar{\boldsymbol{C}}^T_\tau(0)\}U_f^{-1}\sum_{i=1}^{n} m_j'^{-1}\sum_{k_2=1}^{m'_j} E\{\boldsymbol{C}^*_i\psi_\tau(e_{ij'k_2}(\tau))\} \\
= \; & (1-\tau)^{-1}/(N_1)\{\bar{\boldsymbol{C}}^T_\tau(1)-\bar{\boldsymbol{C}}^T_\tau(0)\}U_f^{-1} \\
& \times\Big[\sum_{D_i=1}\sum_{k_1=1}^{m_j}\sum_{k_2=1}^{m'_j} E\left\{e_{ijk_1}(\tau)e^{+}_{ijk_1}(\tau)m_j'^{-1}\boldsymbol{C}^*_i\psi_\tau(e_{ij'k_2}(\tau))\right\} \\
& +\sum_{i_1\neq i_2}^{D_i=1}\sum_{k_1=1}^{m_j} E\{e_{ijk_1}(\tau)e^{+}_{ij1k_1}(\tau)\}\sum_{k_2=1}^{m'_j} E\{m_j'^{-1}\boldsymbol{C}^*_i\psi_\tau(e_{ij'k_2}(\tau))\}\Big] \\
& -(1-\tau)^{-1}/(N_1)\sum_{D_i=1}\sum_{k_1=1}^{m_j} E\{e_{ijk_1}(\tau)e^{+}_{ijk_1}(\tau)\} \\
& \times\{\bar{\boldsymbol{C}}^T_\tau(1)-\bar{\boldsymbol{C}}^T_\tau(0)\}U_f^{-1}\sum_{i=1}^{n} m_j'^{-1}\sum_{k_2=1}^{m'_j} E\{\boldsymbol{C}^*_i\psi_\tau(e_{ij'k_2}(\tau))\} \\
= \; & (1-\tau)^{-1}/(N_1)\{\bar{\boldsymbol{C}}^T_\tau(1)-\bar{\boldsymbol{C}}^T_\tau(0)\}U_f^{-1} \\
& \times\Big[\sum_{D_i=1}\sum_{k_1=1}^{m_j}\sum_{k_2=1}^{m'_j} E\left\{e_{ijk_1}(\tau)e^{+}_{ijk_1}(\tau)m_j'^{-1}\boldsymbol{C}^*_i\psi_\tau(e_{ij'k_2}(\tau))\right\} \\
& -\sum_{D_i=1}\sum_{k_1=1}^{m_j}\sum_{k_2=1}^{m'_j} E\{e_{ijk_1}(\tau)e^{+}_{ijk_1}(\tau)\}E\{m_j'^{-1}\boldsymbol{C}^*_i\psi_\tau(e_{ij'k_2}(\tau))\}\Big]
\end{aligned}
$$

$$
\begin{aligned}
& E(A_2B_1) - E(A_2)E(B_1) \\
= \; & (1-\tau)^{-2}/(N_0N_1)E\{\sum_{D_i=0}\sum_{k_1=1}^{m_j} e_{ijk_1}(\tau)e^{+}_{ijk_1}(\tau)\sum_{D_i=1}\sum_{k_2=1}^{m'_j} e_{ij'k_2}(\tau)e^{+}_{ij'k_2}(\tau)\} - \\
& (1-\tau)^{-2}/(N_0N_1)\sum_{D_i=0}\sum_{k_1=1}^{m_j} E\{e_{ijk_1}(\tau)e^{+}_{ijk_1}(\tau)\}\sum_{D_i=1}\sum_{k_2=1}^{m'_j} E\{e_{ij'k_2}(\tau)e^{+}_{ij'k_2}(\tau)\} \\
= \; & 0
\end{aligned}
$$

$$
\begin{aligned}
& E(A_2B_2) - E(A_2)E(B_2) \\
= \; & E\left\{ (1-\tau)^{-2}/N_0^2 \sum_{D_i=0} \sum_{k_1=1}^{m_j} e_{ijk_1}(\tau) e^+_{ijk_1}(\tau) \sum_{D_i=0} \sum_{k_2=1}^{m'_j} e_{ij'k_2}(\tau) e^+_{ij'k_2}(\tau) \right\} \\
& -(1-\tau)^{-2}/N_0^2 \sum_{D_i=0} \sum_{k_1=1}^{m_j} E\left\{ e_{ijk_1}(\tau) e^+_{ijk_1}(\tau) \right\} \sum_{D_i=1} \sum_{k_2=1}^{m'_j} E\left\{ e_{ij'k_2}(\tau) e^+_{ij'k_2}(\tau) \right\} \\
= \; & (1-\tau)^{-2}/N_0^2 \Bigg[\sum_{i_1=i_2}^{D_i=0} \sum_{k_1=1}^{m_j} \sum_{k_2=1}^{m'_j} E\left\{ e_{ijk_1}(\tau) e^+_{ij1k_1}(\tau) e_{ij'k_2}(\tau) e^+_{ij'k_2}(\tau) \right\} \\
& + \sum_{i_1 \neq i_2}^{D_i=0} \sum_{k_1=1}^{m_j} E\left\{ e_{i_1jk_1}(\tau) e^+_{ij1k_1}(\tau) \right] \sum_{k_2=1}^{m'_j} E\left[e_{i_2j'k_2}(\tau) e^+_{ij'k_2}(\tau) \right\} \Bigg] \\
& -(1-\tau)^{-2}/N_0^2 \left[\sum_{i_1} \sum_{i_2} \sum_{k_1=1}^{m_j} E\{ e_{i_1jk_1}(\tau) e^+_{ij1k_1}(\tau) \} \sum_{k_2=1}^{m'_j} E\{ e_{i_2j'k_2}(\tau) e^+_{ij'k_2}(\tau) \} \right] \\
= \; & (1-\tau)^{-2}/N_0^2 \Bigg[\sum_{i_1=i_2}^{D_i=0} \sum_{k_1=1}^{m_j} \sum_{k_2=1}^{m'_j} E\{ e_{ijk_1}(\tau) e^+_{ij1k_1}(\tau) e_{ij'k_2}(\tau) e^+_{ij'k_2}(\tau) \} \\
& - \sum_{i_1=i_2}^{D_i=0} \sum_{k_1=1}^{m_j} E\{ e_{ijk_1}(\tau) e^+_{ij1k_1}(\tau) \} \sum_{k_2=1}^{m'_j} E\{ e_{ij'k_2}(\tau) e^+_{ij'k_2}(\tau) \} \Bigg]
\end{aligned}
$$

$$
\begin{aligned}
& E(A_2B_3) - E(A_2)E(B_3) \\
= \; & E\Bigg[(1-\tau)^{-1}/(N_0) \sum_{D_i=0} \sum_{k_1=1}^{m_j} e_{ijk_1}(\tau) e^+_{ijk_1}(\tau) \\
& \times \{\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0)\} U_f^{-1} \sum_{i=1}^{n} m_j'^{-1} \sum_{k_2=1}^{m'_j} \boldsymbol{C}^*_i \psi_\tau(e_{ij'k_2}(\tau))\Bigg] \\
& -(1-\tau)^{-1}/(N_0) \sum_{D_i=0} \sum_{k_1=1}^{m_j} E\{e_{ijk_1}(\tau) e^+_{ijk_1}(\tau)\} \\
& \times \{\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0)\} U_f^{-1} \sum_{i=1}^{n} m_j'^{-1} \sum_{k_2=1}^{m'_j} E\{\boldsymbol{C}^*_i \psi_\tau(e_{ij'k_2}(\tau))\} \\
= \; & (1-\tau)^{-1}/(N_0) \{\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0)\} U_f^{-1} \\
& \times \Bigg[\sum_{D_i=0} \sum_{k_1=1}^{m_j} \sum_{k_2=1}^{m'_j} E\left\{e_{ijk_1}(\tau) e^+_{ijk_1}(\tau) m_j'^{-1} \boldsymbol{C}^*_i \psi_\tau(e_{ij'k_2}(\tau))\right\} \\
& - \sum_{D_i=0} \sum_{k_1=1}^{m_j} \sum_{k_2=1}^{m'_j} E\{e_{ijk_1}(\tau) e^+_{ijk_1}(\tau)\} E\{m_j'^{-1} \boldsymbol{C}^*_i \psi_\tau(e_{ij'k_2}(\tau))\}\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
& E(A_3B_1) - E(A_3)E(B_1) \\
= \; & E\Bigg[\{\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0)\} U_f^{-1} \sum_{i=1}^{n} m_j^{-1} \sum_{k_1=1}^{m_j} \boldsymbol{C}^*_i \psi_\tau(e_{ijk_1}(\tau)) \\
& \times (1-\tau)^{-1} \sum_{D_i=1} \sum_{k_2=1}^{m'_j} e_{ij'k_2}(\tau) e^+_{ij'k_2}(\tau)/N_1\Bigg] \\
& -(1-\tau)^{-1}/(N_1) \sum_{D_i=1} \sum_{k_1=1}^{m_j} E\{e_{ij'k_2}(\tau) e^+_{ij'k_2}(\tau)\} \\
& \times \{\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0)\} U_f^{-1} \sum_{i=1}^{n} m_j'^{-1} \sum_{k_1=1}^{m_j} E\{\boldsymbol{C}^*_i \psi_\tau(e_{ijk_1}(\tau)))\} \\
= \; & (1-\tau)^{-1}/(N_1) \{\bar{\boldsymbol{C}}^T_\tau(1) - \bar{\boldsymbol{C}}^T_\tau(0)\} U_f^{-1} \\
& \times \Bigg[\sum_{D_i=1} \sum_{k_1=1}^{m_j} \sum_{k_2=1}^{m'_j} E\{e_{ij'k_2}(\tau) e^+_{ij'k_2}(\tau) m_j^{-1} \boldsymbol{C}^*_i \psi_\tau(e_{ijk_1}(\tau))\} \\
& - E\{e_{ij'k_2}(\tau) e^+_{ij'k_2}(\tau)\} E\{m_j^{-1} \boldsymbol{C}^*_i \psi_\tau(e_{ijk_1}(\tau))\}\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
& E(A_3B_2) - E(A_3)E(B_2) \\
= \; & E\Bigg[\{\bar{\boldsymbol{C}}_\tau^T(1) - \bar{\boldsymbol{C}}_\tau^T(0)\}U_f^{-1}\sum_{i=1}^{n} m_j^{-1}\sum_{k_1=1}^{m_j}\boldsymbol{C}_i^*\psi_\tau(e_{ijk_1}(\tau)) \\
& \times(1-\tau)^{-1}\sum_{D_i=0}\sum_{k_2=1}^{m'_j} e_{ij'k_2}(\tau)e^+_{ij'k_2}(\tau)/N_0\Bigg] \\
& -(1-\tau)^{-1}/(N_1)\sum_{D_i=0}\sum_{k_1=1}^{m_j} E\{e_{ij'k_2}(\tau)e^+_{ij'k_2}(\tau)\} \\
& \times\{\bar{\boldsymbol{C}}_\tau^T(1) - \bar{\boldsymbol{C}}_\tau^T(0)\}U_f^{-1}\sum_{i=1}^{n} m_j'^{-1}\sum_{k_2=1}^{m'_j} E\{\boldsymbol{C}_i^*\psi_\tau(e_{ijk_1}(\tau))\} \\
= \; & (1-\tau)^{-1}/(N_0)\{\bar{\boldsymbol{C}}_\tau^T(1) - \bar{\boldsymbol{C}}_\tau^T(0)\}U_f^{-1} \\
& \times\Bigg[\sum_{D_i=0}\sum_{k_1=1}^{m_j}\sum_{k_2=1}^{m_j} E\{e_{ij'k_2}(\tau)e^+_{ij'k_2}(\tau)m_j'^{-1}\boldsymbol{C}_i^*\psi_\tau(e_{ijk_1}(\tau))\} \\
& -E\{e_{ij'k_2}(\tau)e^+_{ij'k_2}(\tau)\}E\{m_j'^{-1}\boldsymbol{C}_i^*\psi_\tau(e_{ijk_1}(\tau))\}\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
& E(A_3B_3) - E(A_3)E(B_3) \\
= \; & E\Bigg[\{\bar{\boldsymbol{C}}_\tau^T(1) - \bar{\boldsymbol{C}}_\tau^T(0)\}U_f^{-1}\sum_{i=1}^{n} m_j^{-1}\sum_{k_1=1}^{m_j}\boldsymbol{C}_i^*\psi_\tau(e_{ijk_1}(\tau)) \\
& \times\{\bar{\boldsymbol{C}}_\tau^T(1) - \bar{\boldsymbol{C}}_\tau^T(0)\}U_f^{-1}\sum_{i=1}^{n} m_j'^{-1}\sum_{k_2=1}^{m'_j}\boldsymbol{C}_i^*\psi_\tau(e_{ij'k_2}(\tau))\Bigg] \\
& -\{\bar{\boldsymbol{C}}_\tau^T(1) - \bar{\boldsymbol{C}}_\tau^T(0)\}U_f^{-1}\sum_{i=1}^{n} m_j^{-1}\sum_{k_1=1}^{m_j} E\{\boldsymbol{C}_i^*\psi_\tau(e_{ijk_1}(\tau))\} \\
& \times\{\bar{\boldsymbol{C}}_\tau^T(1) - \bar{\boldsymbol{C}}_\tau^T(0)\}U_f^{-1}\sum_{i=1}^{n} m_j'^{-1}\sum_{k_2=1}^{m'_j} E\{\boldsymbol{C}_i^*\psi_\tau(e_{ij'k_2}(\tau))\} \\
= \; & \{\bar{\boldsymbol{C}}_\tau^T(1) - \bar{\boldsymbol{C}}_\tau^T(0)\}U_f^{-1}\Bigg[\sum_{i=1}^{n} m_j^{-1}m_j'^{-1}\sum_{k_1=1}^{m_j}\sum_{k_2=1}^{m'_j}\boldsymbol{C}_i^*\boldsymbol{C}_i^{*T}\psi_\tau(e_{ijk_1}(\tau))\psi_\tau(e_{ij'k_2}(\tau)) \\
& -\sum_{k_1=1}^{m_j} E\{m_j^{-1}\boldsymbol{C}_i^*\psi_\tau(e_{ijk_1}(\tau))\}\sum_{k_2=1}^{m'_j} E\{m_j'^{-1}\boldsymbol{C}_i^*\psi_\tau(e_{ij'k_2}(\tau))\}\Bigg]U_f^{-1}\{\bar{\boldsymbol{C}}_\tau(1) - \bar{\boldsymbol{C}}_\tau(0)\}
\end{aligned}
$$

Thus, we have derived the covariance matrix of Σ. When Σ is positive definite, we can use $\Sigma^{-1/2}$ to standardize X into $X^* = \Sigma^{-1/2}X \sim N_n(0, \mathbf{I_{j\times j}})$,

and $T_{\tau,p}^{TTS^*}$ in $X^* = (T_{\tau,1}^{TTS^*}, ..., T_{\tau,j}^{TTS^*})$ follows a standard normal distribution.

Theorem 4.1.2

If $T_{\tau,p}^{TTS^*}$ follow standard normal distribution, we have

$$P_{TTS} = \sqrt{j}\bar{T}_{\tau}^{TTS^*}, \text{ with } \bar{T}_{\tau}^{TTS^*} = \frac{1}{j}\sum_{p=1}^{j} T_{\tau,p}^{TTS^*}, \tag{4.9}$$

and

$$P_{TTS} \sim N(0,1) \tag{4.10}$$

with j being the number of genes in pathway of interest.

Proof of Theorem 4.1.2

According to Lemma 4.1.2 and under null hypothesis, we are able to have $T_{\tau,p}^{TTS^*} \sim N(0,1)$. Let $\bar{T}_{\tau}^{TTS^*} = \frac{1}{j}\sum_{p=1}^{j} T_{\tau,p}^{TTS^*}$. Then the moment generating function of $P_{TTS} = \sqrt{j}\bar{T}_{\tau}^{TTS^*}$ is given by

$$\begin{aligned}
M_{P_{TTS}}(t) &= E\left[\exp\left\{t(\frac{\sqrt{j}}{j}\sum_{p=1}^{j} T_{\tau,p}^{TTS^*})\right\}\right] \\
&= \exp\left\{t\left(\frac{\sqrt{j}}{j}\sum_{p=1}^{j}\mu\right) + \frac{t^2}{2}\left(\frac{(\sqrt{j})^2}{j^2}\sum_{p=1}^{j}\sigma^2\right)\right\} \\
&= \exp\left\{t\left(\frac{1}{\sqrt{j}}j\mu\right) + \frac{t^2}{2}\left(\frac{1}{j}(j\sigma^2)\right)\right\} \\
&= \exp\left\{\sqrt{j}\mu t + \frac{t^2}{2}\sigma^2\right\}
\end{aligned}$$

Using the moment generating function, we have shown that P_{TTS} follows a normal distribution with mean $\sqrt{j}\mu$ and variance σ^2, where μ and σ^2 are the mean and variance of $T_{\tau,p}^{TTS^*}$. According to Lemma 4.1.2 that $T_{\tau,p}^{TTS^*} \sim N(0,1)$, P_{TTS} has mean $\sqrt{j}\mu = 0$ and variance $\sigma^2 = 1$. Hence, we have proved $P_{TTS} \sim N(0,1)$.

4.2 Simulation

We conducted simulation studies to investigate the statistical validity and power of the proposed pathway test, P_{TTS}. We compared P_{TTS} to popular pathway and gene set analysis methods, including the Gene Set Enrichment Analysis (called $GSEA$) by Subramanian et al. (2005) using R package by Vremo et al. (2013), the Generally Applicable Gene set Enrichment (called $GAGE$) by Luo et al. (2009) using R package by Luo and Brouwer (2013), and the Fishers combined probability test adapted to pathway analysis (called $Fisher$) by Vremo et al. (2013). We generated exon-level gene expression data in Log2-RPKM format from the following model to fit our model, and converted the measurement to gene-level raw counts to fit other analysis methods.

$$Z_{ijk} = 5 + \gamma C_i + \delta e^{+}_{ran} I(D_i = 1) + e_{ijk}, \text{where} \tag{4.11}$$

Z_{ijk} is the intensity value of exon k of gene j for subject sample i, C_i indicates the covariate value and $C_i \sim N(2.5, 0.5^2)$. D_i indicates the disease status, normal tissue or cancer, of the patient sample i. The corresponding error terms are denoted by e_{ijk}s, and $e_{ijk} = e_{exon,ijk} + e_{gene,ij}$. The inter-exon error term $e_{exon,ijk}$ are normally distributed with unit variance and compound symmetry correlation structure $cor(e_{exon,ijk}, e_{exon,ijk'}) = 0.8$. And gene-wise error term $e_{gene,ij}$ are normally distributed with unit variance and compound symmetry correlation structure $cor(e_{gene,ij}, e_{gene,ij'}) = 0.5$. e_{ran} follows an standard normal distribution and e^{+}_{ran} is the indicator function. Using this setting, we only apply group effect δ on cancer group expression data which are above the 50th quantile. We used certain number of genes to form the pathway and investigated the following three scenarios.

Scenario 1 (null hypothesis): $\delta = 0$ for all pathways of 10 and 30 gene.

Scenario 2 (alternative hypothesis 1): For pathways of 10 genes, 75% of the pathways are none DE with $\delta = 0$. And 25% of the pathway have 5 or 8 DE genes in each pathway. The group effect $\delta \sim uniform(1, 2)$ while the rest of the genes have $\delta = 0$.

Scenario 3 (alternative hypothesis 2): For pathways of 30 genes, 75% of the pathways are none DE with $\delta = 0$. And 25% of the pathway have 15, 20, or 24 DE genes in each pathway. The group effect $\delta \sim uniform(1, 2)$ while the rest of the genes have $\delta = 0$.

We used Monte Carlo to generate $1,000$ pathway samples for scenario 1 and $1,200$ pathway samples for scenario 2 and 3. Each generated gene has a gene length of 30 exon regions. To implement the other pathway analysis methods, we followed the standard procedure recommended by the authors. We first performed DE analysis use *Limma*, *edgeR*, and *DESeq*2 for each individual gene, then used the outputs of these test as inputs for the pathway analysis methods. We used *Limma*'s test statistics for *GSEA*, *Limma*'s p-values to perform *Fisher*, and the logFCs of *Limma*, *edgeR*, and *DESeq*2 to conduct *GAGE* (called $GAGE_{Limma}$, $GAGE_{edgeR}$, and $GAGE_{DESeq2}$ respectively). For the proposed test, we used $\tau = 0.5$ for testing all scenario at the nominal levels of 5% and calculated the average false positive rates (FPRs) and true positive rates (TPRs).

In scenario 1, the FPRs are shown in Table 4.1. We observed that $GAGE_{Limma}$, $GAGE_{edgeR}$, and $GAGE_{DESeq2}$ have relatively conservative FPRs. *GSEA*, which uses a sample based permutation methods, is sensitive to the noise and shows inflated FPRs. *Fisher* shows modestly inflated FPRS. In contrast, FPRs of our proposed method P_{TTS} is able to converge to the nominal value.

In scenario 2, the FPRs are shown in the top of Table 4.2. We observed that $GAGE_{Limma}$, $GAGE_{edgeR}$, $GAGE_{DESeq2}$, and *Fisher* have relatively conservative FPRs. *GSEA* are sensitive to noise and show inflated FPRs while P_{TTS} can maintain the FPRs around the nominal value. When the group effect is present, the cancer group ($D_i = 1$) has a heavier right tail and larger variance than the normal group ($D_i = 0$) for DE genes. The difference between the two groups is relatively small at the median and gradually increases in the upper quantiles.

Table 4.1
FPRs at the nominal levels of 5% for scenario 1. The values in the table are percentages.

Scenario 1		FPR					
Pathway Gene Number	Sample Size	P_{TTS}	$GAGE_{Limma}$	$GAGE_{edgeR}$	$GAGE_{DESeq2}$	$GSEA$	$Fisher$
10	60	6.70	3.90	2.30	2.60	28.50	9.00
30	60	7.70	3.80	1.90	2.20	26.80	10.90
10	80	5.90	3.30	2.50	2.40	28.90	8.80
30	80	7.20	3.20	1.50	1.70	24.70	6.80
10	100	6.50	4.00	2.00	1.90	26.60	8.60
30	100	5.70	3.00	1.90	1.90	23.90	9.00
10	150	5.40	4.40	2.60	2.60	26.40	8.60
30	150	6.50	3.20	2.30	2.40	27.40	8.30
10	200	4.00	2.90	1.50	1.50	27.70	8.70
30	200	5.40	3.40	2.50	2.40	31.30	8.10

The TPRs are shown in bottom of Table 4.2. P_{TTS} has the best performance comparing with all other methods. The advantage of P_{TTS} is more prominent when analyzing smaller sample sizes (e.g., 60) and fewer number of DE genes within the pathway (e.g., 5), which are often encountered in practice. $GSEA$ has the second best TPRs but the result is less trustworthy considering its abnormal FPRs. $GAGE_{Limma}$, $GAGE_{edgeR}$, and $GAGE_{DESeq2}$ shows weaker FPRs. As $GAGE$ assumes independent correlation of the genes, it does not make use of the gene-wise correlation and loses its power. $Fisher$ performs better than these three methods but is still worse than P_{TTS}.

In scenario 3, the FPRs are shown in the top of Table 4.3. We observed that $GAGE_{Limma}$, $GAGE_{edgeR}$, $GAGE_{DESeq2}$, and $Fisher$ have relatively conservative

Table 4.2
FPRs and TPRs at the nominal levels of 5% for scenario 2. The values in the table are percentages.

Scenario 2	Nominal Level		FPR					
Pathway Gene Number	Sample Size	DE Genes	P_{TTS}	$GAGE_{Limma}$	$GAGE_{edgeR}$	$GAGE_{DESeq2}$	$GSEA$	$Fisher$
10	60	5	6.11	0.67	0.56	0.56	19.11	0.11
10	60	8	7.00	0.78	0.56	0.56	23.78	0.44
10	80	5	5.89	0.67	0.11	0.11	20.67	0.44
10	80	8	6.00	0.78	0.11	0.11	22.00	0.44
10	100	5	6.78	0.56	0.56	0.56	15.56	0.11
10	100	8	6.67	0.56	0.56	0.56	21.11	0.22
10	150	5	5.78	0.00	0.11	0.11	15.00	0.00
10	150	8	5.22	0.00	0.11	0.11	21.44	0.00
Scenario 2			TPR					
Pathway Gene Number	Sample Size	DE Genes	P_{TTS}	$GAGE_{Limma}$	$GAGE_{edgeR}$	$GAGE_{DESeq2}$	$GSEA$	$Fisher$
10	60	5	87.00	23.67	18.67	20.00	68.00	70.00
10	60	8	98.33	68.33	57.33	55.33	88.67	79.33
10	80	5	91.33	31.00	23.67	24.00	67.00	77.33
10	80	8	99.67	76.67	62.67	61.33	89.33	80.67
10	100	5	96.33	32.33	25.33	26.00	67.33	76.67
10	100	8	100.00	81.67	73.00	72.00	94.00	86.67
10	150	5	98.67	40.33	35.00	34.00	73.33	79.00
10	150	8	100.00	88.33	82.33	80.33	96.67	91.67

FPRs. $GSEA$ is sensitive to noise and shows inflated FPRs while P_{TTS} can maintain the FPRs around the nominal value.

The TPRs are shown in bottom of Table 4.3. $Fisher$ has the best performance comparing with all other methods. Our methods P_{TTS} has the second best TPRs. All the methods perform equally for larger sample size and more DE genes in the pathway.

Table 4.3
FPRs and TPRs at the nominal levels of 5% for scenario 3. The values in the table are percentages.

Scenario 3			FPR					
Pathway Gene Number	Sample Size	DE Genes	P_{TTS}	$GAGE_{Limma}$	$GAGE_{edgeR}$	$GAGE_{DESeq2}$	$GSEA$	$Fisher$
30	60	15	8.22	1.89	0.89	1.00	46.33	0.44
30	60	20	7.56	1.89	0.78	0.78	52.22	0.33
30	60	24	8.00	2.00	0.89	0.89	57.44	0.56
30	80	15	7.44	1.78	0.78	0.78	45.00	0.22
30	80	20	7.11	0.78	0.44	0.67	51.22	0.11
30	80	24	7.33	0.78	0.67	0.78	59.00	0.22
30	100	15	5.56	0.67	0.67	0.67	45.00	0.11
30	100	20	5.89	0.67	0.56	0.78	53.11	0.11
30	100	24	5.56	0.67	0.44	0.56	62.00	0.11
30	150	15	6.56	0.33	0.56	0.67	44.33	2.22
30	150	20	6.78	0.33	0.56	0.67	57.33	0.00
30	150	24	6.56	0.22	0.56	0.56	64.78	2.22
Scenario 3			TPR					
Pathway Gene Number	Sample Size	DE Genes	P_{TTS}	$GAGE_{Limma}$	$GAGE_{edgeR}$	$GAGE_{DESeq2}$	$GSEA$	$Fisher$
30	60	15	92.33	87.67	84.00	84.00	92.00	97.00
30	60	20	98.33	98.67	93.00	92.67	97.33	97.33
30	60	24	100.00	97.67	93.33	93.67	99.67	96.00
30	80	15	98.33	92.00	95.00	95.00	97.00	99.33
30	80	20	99.67	99.33	97.67	96.33	97.33	98.33
30	80	24	100.00	98.67	96.67	96.00	99.67	98.00
30	100	15	98.67	100.00	98.00	98.00	96.00	99.00
30	100	20	100.00	99.67	98.67	98.33	99.00	99.33
30	100	24	100.00	100.00	98.67	98.33	100.00	99.00
30	150	15	99.33	100.00	99.33	99.00	97.00	99.67
30	150	20	100.00	100.00	100.00	100.00	99.67	99.67
30	150	24	100.00	100.00	100.00	100.00	100.00	99.67

5. An application on non-small cell lung cancer data to detect differential expressed pathway

5.1 Introduction

To detect differential expressed non-small cell lung cancer (NSCLC) pathway, we used lung adenocarcinoma data accessible at the TCGA public data portal, with the RNA-seq data profiled from 50 cancer and 50 normal tissue samples at the exon-level and gene-level. The gene expression data were normalized into Log2-RPKM following standard protocols. Then we eliminated the non-expressed genes in both groups prior to our downstream analysis. As ancillary clinical information, we also considered gender and smoking status in our study. The objective was to utilize biological knowledge on NSCLC pathways to form gene sets of interest, and then detect the gene sets that are differentially expressed between cancer and normal tissue samples. In particular, we used the biological knowledge on the NSCLC from KEGG database (Kanehisa and Goto, 2000). We formed 7 pathway gene sets and 1 whole NSCLC gene set which includes all the associated genes of NSCLC. The complete pathway information and gene lists are list in Table 5.1.

We applied the proposed test, P_{TTS}, to each gene at 50th quantile and used a 10% false discovery rate (FDR) adjustment to control for multiple testing. P_{TTS} first conducts exon-level TTS test on each individual gene in the pathway, then combine the individual TTS test statistics and compute a pathway test statistics. For comparison methods, we first applied standard gene-level DE analysis methods including *edgeR*, $DESeq2$, and *Limma*. Then we followed the standard procedure recommended by the authors and used the outputs of these test as inputs for the

pathway analysis methods. We used *Limma*'s test statistics for $GSEA$, *Limma*'s p-values to perform $Fisher$, and the logFCs of *Limma*, *edgeR*, and $DESeq2$ to conduct $GAGE$ (called $GAGE_{Limma}$,$GAGE_{edgeR}$, and $GAGE_{DESeq2}$ respectively).

Table 5.1
Pathway and gene sets related to non-small cell lung cancer

Pathway Name	Gene
Ras signaling pathway	$CCND1$, $KRAS$, $RASSF1$, $RASSF5$, $STK4$
ErbB signaling pathway	$AKT1$, $AKT2$, $AKT3$, BAD, $CASP9$, $FOXO3$, $KRAS$, $PDPK1$, $PIK3CA$, $PIK3CB$, $PIK3CD$, $PIK3R1$, $PIK3R2$, $PIK3R3$
$MAPK$ signaling pathway	$ARAF$, $BRAF$, $CCND1$, EGF, $EGFR$, $ERBB2$, $GRB2$, $HRAS$, $KRAS$, $MAP2K1$, $MAP2K2$, $MAPK1$, $MAPK3$, $NRAS$, $RAF1$, $SOS1$, $SOS2$, $TGFA$
Calcium signaling pathway	EGF, $EGFR$, $ERBB2$, $PLCG1$, $PLCG2$, $PRKCA$, $PRKCB$, $TGFA$
$PI3K-Akt$ signaling pathway	$AKT1$, $AKT2$, $AKT3$, BAD, $CASP9$, EGF, $EGFR$, $ERBB2$, $FOXO3$, $PDPK1$, $PIK3CA$, $PIK3CB$, $PIK3CD$, $TGFA$
Cell Cycle	$CCND1$, $CDK4$, $CDK6$, $CDKN2A$, $E2F1$, $E2F2$, $E2F3$, $RB1$
RAR/RXR signaling Pathway	$RARB$, $RXRA$, $RXRB$, $RXRG$,
Whole NSCLC gene set	$AKT1$, $AKT2$, $AKT3$, $ARAF$, BAD, $BRAF$, $CASP9$, $CCND1$, $CDK4$, $CDK6$, $CDKN2A$, $E2F1$, $E2F2$, $E2F3$, EGF, $EGFR$, $ERBB2$, $FHIT$, $FOXO3$, $GRB2$, $HRAS$, $KRAS$, $MAP2K1$, $MAP2K2$, $MAPK1$, $MAPK3$, $NRAS$, $PDPK1$, $PIK3CA$, $PIK3CB$, $PIK3CD$, $PIK3CG$, $PIK3R1$, $PIK3R2$, $PIK3R3$ $PIK3R5$, $PLCG1$, $PLCG2$, $PRKCA$, $PRKCB$, $RAF1$ $RARB$, $RASSF1$,$RASSF5$, $RB1$, $RXRA$, $RXRB$, $RXRG$ $SOS1$, $SOS2$, $STK4$, $TGFA$, $TP53$

5.2 Results

We included gender and smoking status, defined as current smoker, reformed smoker, and nonsmoker, as covariates in the analysis. We first compared the DE analysis results. Among 53 genes associated with NSCLC pathways, TTS detected 34 genes at $\tau = 0.5$; *Limma* detected 39 genes, *edgeR* detected 29 genes, while $DESeq2$ detected 41 genes. The Venn diagrams in Figure 5.1 show the number of overlapping gene among the four methods. We observed that 85% of the genes detected by TTS were also detected by *Limma*; 76% of the genes selected by TTS were also detected by *edgeR*; and 85% of the genes selected by TTS were also detected by $DESeq2$ at $\tau = 0.5$. The complete p-values of the four methods for the 53 genes associated with NSCLC pathways are shown in Table 5.2. TTS have similar performance and the identified DE genes tend to overlap with other standard DE methods.

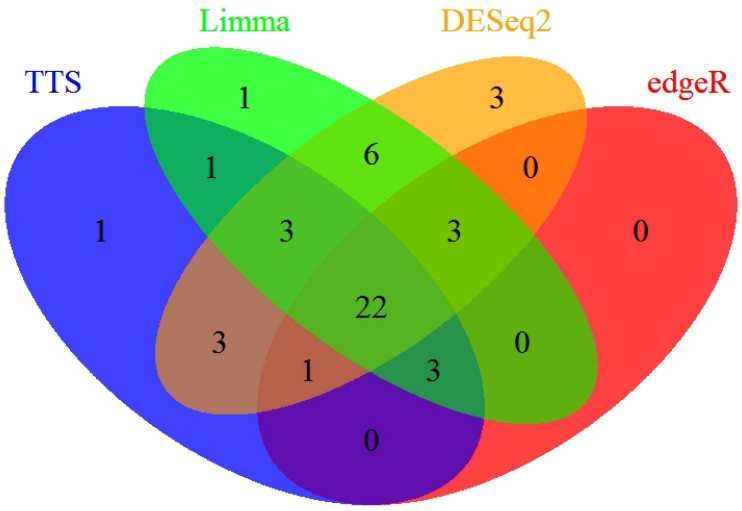

Figure 5.1. Venn diagram of number of overlapping genes among TTS, *edgeR*, $DESeq2$, and *Limma*, for $\tau = 0.5$.

Table 5.2
P-values of genes associated with non-small cell lung cancer pathways.

Gene	*TTS*	*Limma*	*edgeR*	*DESeq2*	Gene	*TTS*	*Limma*	*edgeR*	*DESeq2*
AKT1	0.282	0.115	0.639	0.001	*PDPK1*	0.564	0.044	0.300	0.048
AKT2	0.002	0.003	0.005	0.000	*PIK3CA*	0.501	0.393	0.953	0.689
AKT3	0.051	0.000	0.000	0.000	*PIK3CB*	0.000	0.008	0.015	0.001
ARAF	0.250	0.018	0.339	0.000	*PIK3CD*	0.586	0.250	0.623	0.507
BAD	0.001	0.008	0.091	0.006	*PIK3CG*	0.373	0.002	0.142	0.107
BRAF	0.000	0.000	0.004	0.000	*PIK3R1*	0.000	0.000	0.000	0.000
CASP9	0.528	0.613	0.565	0.599	*PIK3R2*	0.000	0.000	0.000	0.000
CCND1	0.979	0.084	0.735	0.552	*PIK3R3*	0.013	0.000	0.000	0.000
CDK4	0.000	0.000	0.000	0.000	*PIK3R5*	0.000	0.000	0.000	0.000
CDK6	0.002	0.041	0.907	0.903	*PLCG1*	0.494	0.604	0.395	0.495
CDKN2A	0.000	0.000	0.000	0.000	*PLCG2*	0.020	0.000	0.003	0.001
E2F1	0.000	0.000	0.000	0.000	*PRKCA*	0.000	0.917	0.090	0.119
E2F2	0.000	0.000	0.000	0.000	*PRKCB*	0.876	0.000	0.005	0.003
E2F3	0.000	0.000	0.000	0.000	*RAF1*	0.237	0.001	0.329	0.000
EGF	0.000	0.000	0.000	0.001	*RARB*	0.181	0.004	0.108	0.021
EGFR	0.042	0.331	0.006	0.000	*RASSF1*	0.000	0.000	0.000	0.000
ERBB2	0.000	0.000	0.000	0.002	*RASSF5*	0.000	0.000	0.000	0.000
FHIT	0.002	0.236	0.114	0.000	*RB1*	0.001	0.000	0.033	0.000
FOXO3	0.001	0.000	0.000	0.000	*RXRA*	0.000	0.000	0.000	0.677
GRB2	0.223	0.130	0.528	0.089	*RXRB*	0.917	0.548	0.949	0.000
HRAS	0.188	0.535	0.347	0.000	*RXRG*	0.000	0.000	0.000	0.000
KRAS	0.000	0.000	0.000	0.000	*SOS1*	0.035	0.022	0.089	0.000
MAP2K1	0.013	0.485	0.454	0.015	*SOS2*	0.054	0.000	0.053	0.000
MAP2K2	0.000	0.022	0.124	0.014	*STK4*	0.219	0.001	0.046	0.000
MAPK1	0.689	0.005	0.322	0.000	*TGFA*	0.000	0.000	0.000	0.056
MAPK3	0.000	0.000	0.002	0.000	*TP53*	0.000	0.000	0.000	0.619
NRAS	0.000	0.069	0.068	0.029					

Then we focused on the results of the pathway and gene set analysis. Among the 7 pathway gene sets and 1 whole gene set, P_{TTS} is able to detect 7 of these and only misses *ErbB* signaling pathway. *GSEA* and *Fisher* can detect Cell Cycle and *RAR/RXR* signaling pathway and misses all other pathways. $GAGE_{edgeR}$, $GAGE_{Limma}$, and $GAGE_{DESeq2}$ can detect none of these pathways.

Table 5.3
The FDR adjusted P-values of 7 pathway gene sets and 1 whole gene set associated with non-small cell lung cancer.

	Ras signaling pathway	*ErbB* signaling pathway	*MAPK* signaling pathway	Calcium signaling pathway	*PI3K − Akt* signaling pathway	Cell Cycle signaling pathway	*RAR/RXR* signaling pathway	Whole NSCLC gene set
P_{TTS}	3.23E-05	4.44E-01	3.75E-11	5.78E-12	5.88E-04	6.57E-30	6.69E-26	1.83E-07
$GAGE_{edgeR}$	7.63E-01	9.36E-01	8.76E-01	4.84E-01	7.75E-01	2.29E-01	4.02E-01	8.87E-01
$GAGE_{Limma}$	6.69E-01	8.81E-01	9.08E-01	5.59E-01	8.03E-01	1.42E-01	2.68E-01	6.92E-01
$GAGE_{DESeq2}$	7.60E-01	9.42E-01	8.83E-01	4.79E-01	7.73E-01	2.11E-01	3.84E-01	8.79E-01
GSEA	4.10E-01	6.12E-01	8.97E-01	6.18E-01	9.50E-01	1.62E-02	3.35E-02	5.67E-01
Fisher	2.00E-01	6.83E-01	9.19E-01	7.32E-01	8.97E-01	6.56E-02	3.73E-02	3.80E-01

To understand the results of P_{TTS}, we first plotted the correlation structure of Calcium signaling pathway and *ErbB* signaling pathway for cancer samples and normal samples. P_{TTS} is able to detect Calcium signaling pathway because we observed a clear pattern of compound symmetry correlation structure for inter-exon regions within the gene and most gene-wise correlations. Only $ERBB2$ tends to have non compound symmetry correlation with $PLCG2$ and $PRKCB$ in cancer sample and EGF tends to have non compound symmetry correlation with $ERBB2$ and $PLCG1$ in normal samples. As for *ErbB* signaling pathway which P_{TTS} fails to detect, we also observed a clear pattern of compound symmetry correlation structure for all inter-exon regions within genes. As for gene-wise correlation, we observe that $AKT1$, $CASP9$, and $PIK3R2$ have non compound symmetry correlation with other most of the genes in this pathway. Since our model assumption of compound symmetry correlation structure fails, P_{TTS} is unable to correctly detect *ErbB* signaling pathway.

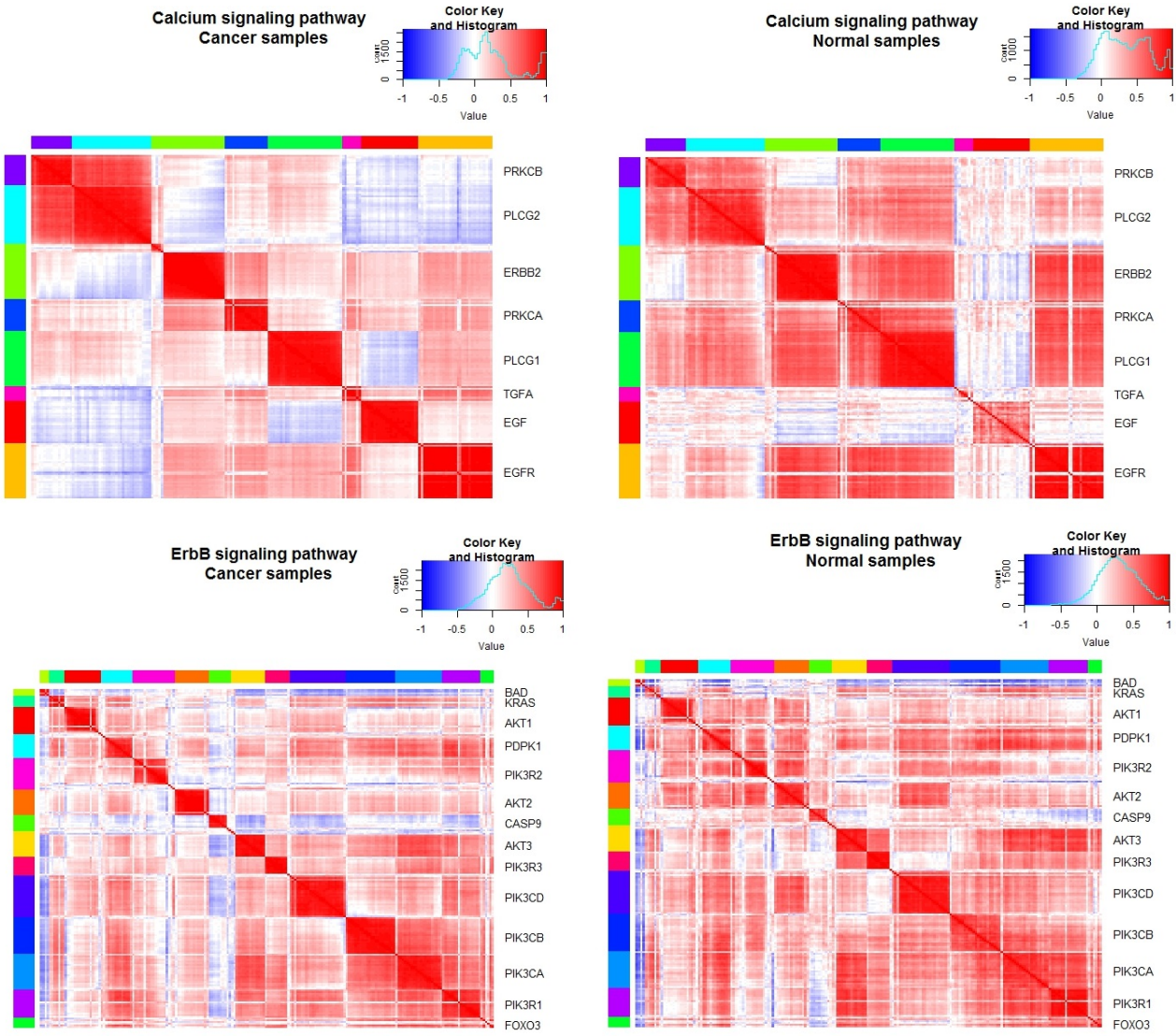

Figure 5.2. Correlation Heatmap of the Calcium signaling pathway and *ErbB* signaling pathway for cancer samples on the left and normal samples on the right.

Then we focused on the results of *GAGE*. We listed the logFCs of NSCLC genes from three DE analysis methods in Table 5.4, and plotted the distribution of logFCs in Figure 5.3. The reason that *GAGE* is not able to detect any pathway is because it uses a two sample t-test on the mean difference between LogFCs of the gene set and whole data set. Then it uses gene permutation method to assess

the significance of its test statistics. When the logFCs of the gene set are relatively small comparing to the whole data set, the effect is not significant. Another reason $GAGE$ fails to detect any pathway is because its assumption of independence between observations. As we have shown in Figure 5.2, the genes from the Calcium signaling pathway and the $ErbB$ signaling pathway are correlated. For example, $PIK3R1$ has strong positive correlation with $PIK3R3$, $PIK3CA$, $PIK3CB$, $PIK3CD$, $PDPK1$, $KRAS$, $FOXO3$, and $AKT3$ in the $ErbB$ signaling pathway for both cancer and normal samples. Ignoring the correlation between genes yields poor result for $GAGE$.

Table 5.4
LogFcs of genes associated with non-small cell lung cancer pathways.

Gene	*Limma*	*edgeR*	*DESeq2*	Gene	*Limma*	*edgeR*	*DESeq2*
AKT1	-0.11	-0.06	0.28	*PDPK1*	-0.17	-0.13	-0.16
AKT2	0.24	0.36	0.33	*PIK3CA*	-0.08	-0.01	-0.04
AKT3	-0.96	-0.69	-0.69	*PIK3CB*	0.24	0.32	0.29
ARAF	-0.14	-0.12	-0.89	*PIK3CD*	-0.15	-0.07	-0.09
BAD	-0.25	-0.22	-0.25	*PIK3CG*	-0.65	-0.28	-0.30
BRAF	0.54	0.59	0.56	*PIK3R1*	-0.88	-0.85	-0.87
CASP9	-0.05	0.09	0.06	*PIK3R2*	0.52	0.55	0.52
CCND1	-0.31	-0.06	-0.10	*PIK3R3*	-0.82	-0.68	-0.70
CDK4	0.93	1.23	1.20	*PIK3R5*	-1.34	-1.08	-1.10
CDK6	-0.33	0.02	-0.02	*PLCG1*	0.05	0.11	0.07
CDKN2A	2.60	3.60	0.47	*PLCG2*	-0.57	-0.42	-0.44
E2F1	1.01	1.21	1.18	*PRKCA*	0.02	0.26	0.23
E2F2	2.15	2.23	2.20	*PRKCB*	-0.75	-0.49	-0.50
E2F3	1.48	1.60	1.56	*RAF1*	-0.14	-0.13	1.56
EGF	1.77	2.72	-0.54	*RARB*	-0.43	-0.31	0.32
EGFR	0.19	0.59	2.66	*RASSF1*	-0.81	-0.82	-0.84
ERBB2	0.50	0.71	0.22	*RASSF5*	-0.59	-0.48	1.60
FHIT	0.18	0.43	-3.15	*RB1*	-0.35	-0.27	-0.30
FOXO3	-0.56	-0.56	-0.42	*RXRA*	-0.88	-0.81	-0.02
GRB2	-0.10	-0.09	-0.11	*RXRB*	-0.04	0.01	-3.12
HRAS	0.06	0.14	-1.09	*RXRG*	-4.12	-3.17	-0.44
KRAS	0.54	0.73	0.69	*SOS1*	0.20	0.22	-0.28
MAP2K1	0.06	0.10	0.17	*SOS2*	-0.29	-0.25	-2.90
MAP2K2	0.16	0.20	-0.32	*STK4*	-0.29	-0.25	-0.28
MAPK1	-0.16	-0.13	-1.05	*TGFA*	1.16	1.62	0.30
MAPK3	-0.46	-0.39	-0.42	*TP53*	0.47	0.59	0.05
NRAS	0.18	0.23	0.21				

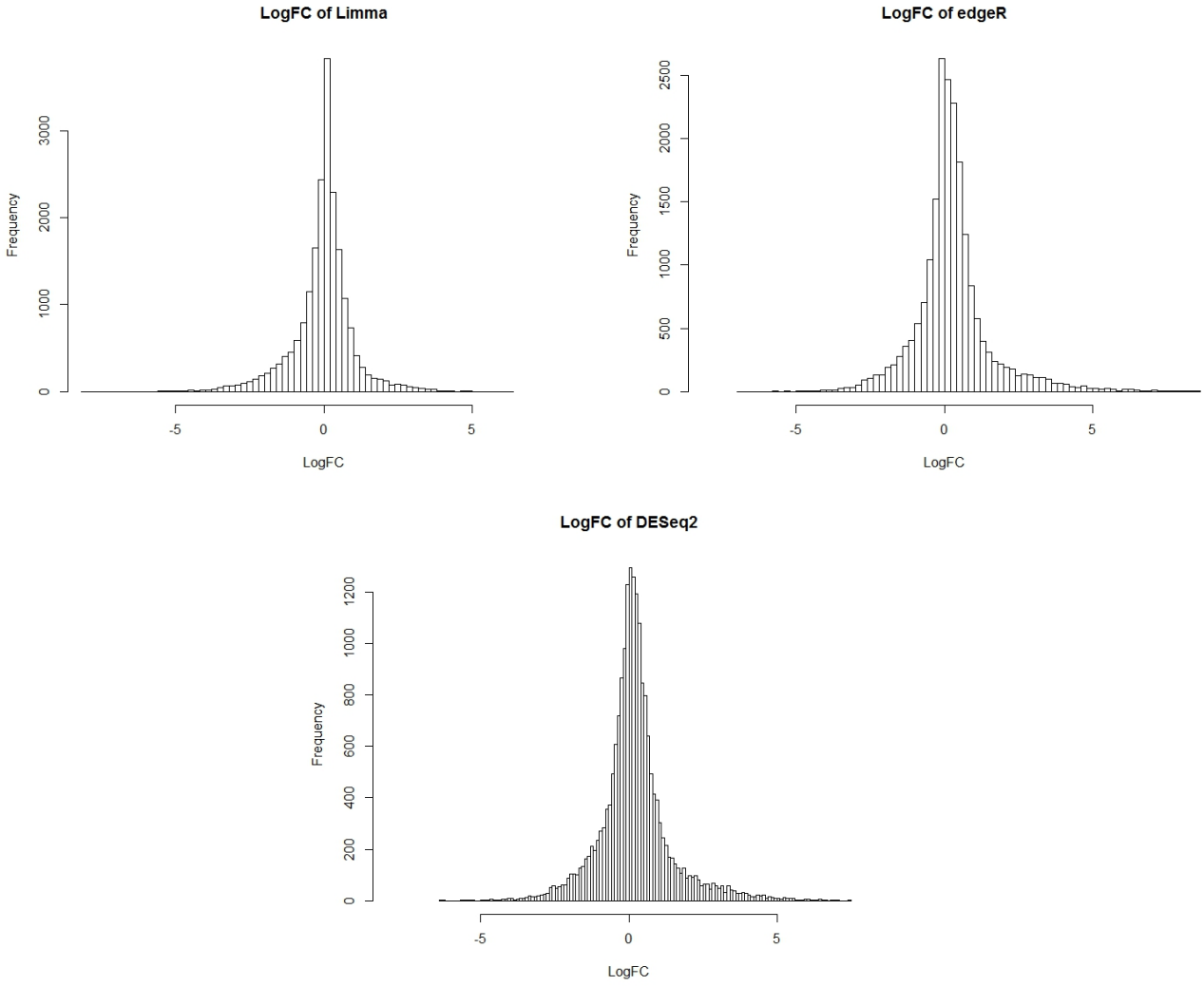

Figure 5.3. Distribution of LogFCs of *Limma*, *edgeR*, and *DESeq*2.

We then looked at the results of *GSEA* and *Fisher*. *GSEA* used a sample based permutation test in order to produce a null distribution for the test statistics. The construction of null distribution strongly depends on the distribution of DE analysis results (*Limma*'s output) and the size of the gene set. The KolmogorovSmirnov-like statistic is also known to be low on power. When the distribution of *Limma*'s test statistics are widely spread as shown in the left of Figure 5.4 and test statistics of

Limma are modest for the genes of interest, *GSEA* loses it power and only detects two pathways.

Fisher constructs the null distributions by a gene sampling based permutation approach. For each gene set, *Fisher* randomly took a group of genes of same size and calculated the gene set statistic. The fraction of random generated gene set statistics that are equal or larger than the original gene set statistics is the final P-value. Hence *Fisher* is also very sensitive to the distribution of the DE analysis results (*Limma*'s output). When the p-values from *Limma* are modest for the genes of interest and the distribution of the p-values are concentrated near 0 as shown in the right of Figure 5.4, *Fisher* has trouble construct good null distributions and only detects two pathways.

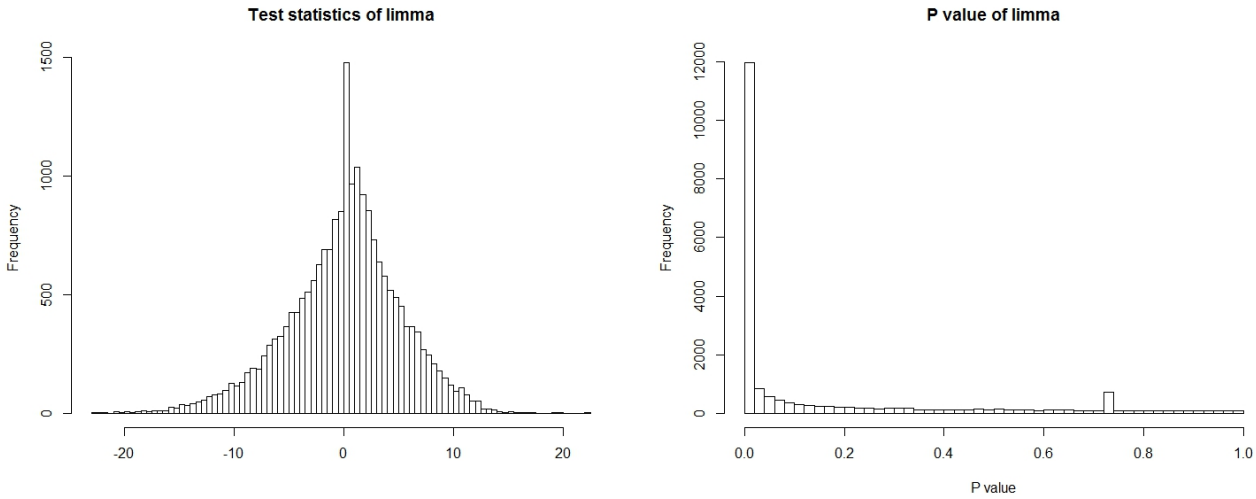

Figure 5.4. Distribution of test statistics and P-values of *Limma*.

In summary, P_{TTS} shows better performance than *GAGE*, *GSEA*, and *Fisher* due to its ability to utilize all the information in the upper quantile region and its robustness to model distributions and individual outliers in the DE analysis step by using *TTS* method. Then in the pathway analysis, P_{TTS} is able to account for gene-wise correlation for the gene set and hence gain more information from the data. Our

proposed test is also a good supplement method to use along with standard pathway analysis methods, as it is able to include pathways that are missed by $GAGE$, $GSEA$, and $Fisher$. P_{TTS} loses it power advantage when the gene-wise correlation structure is not exchangeable as the it violates the assumption of our model.

Overall, our proposed method offers a powerful and robust supplement for detecting differentially expressed pathway by utilizing the information in the region of interest and account for inter-exon and gene-wise correlation.

6. Discussion

6.1 Discussion and future work

We have proposed a new test based on quantile regression that can detect differential gene expression in RNA-seq data. This covariate-adjusted test utilizes the information of quantiles in a tail region of the distribution instead of a single quantile level to make substantial improvement in power. The intrinsic correlation among exons within a gene can be directly accounted for in the proposed method. The quantile-based test is also robust to a heavy tailed distribution in RNA-seq data. Simulation results and real data analysis of TCGA lung adenocarcinoma data demonstrate the merit of the proposed method. The method has been further extend to conduct pathway analysis for RNA-seq data. The proposed pathway test incorporates biological knowledge of pathways to generate gene set of interest, then utilizes the test statistics of differential gene expression from our tail-based test and computes the pathway test statistics. By accounting for intrinsic gene wise correlation, this method is a powerful and robust tool to detect differential expressed pathway in RNA-seq data. Simulation results and real data analysis on NSCLC pathway using KEGG pathway database demonstrate the advantage of the proposed method over other popular pathway analysis methods.

In this paper, we focus on the compound symmetry correlation structure among exons within a gene, which has been empirically shown to be sensible for RNA-seq data. In further investigations, we plan to broaden the study to account for more flexible correlation structures for other applications. We also want to explore the possibility to borrow information across genes in biological pathways to improve

test efficiency. In the lung cancer study, we find that the outliers in the tail region sometimes cause the quantile difference to overturn in the extreme tail region. In future investigations, we will explore how to handle outliers of this type.

7. Appendix

Tables 7.1, 7.2, 7.3, and 7.4 list genes that are detected by TTS but not by the QRS_C, LME, $Limma$, $edgeR$, and $DESeq2$ with the supporting medical literature.

Table 7.1
List of genes detected by TTS but missed by QRS_c nad LME

Test method	Gene list
QRS_c	$ADAMTS9$, $C3ORF21$, $MBD4$, $ZMAT3$, $FOXP1$, $GSK3B$
	$PLD1$, $SIAH2$, $C3orf33$, $EHHADH$, $IQCB1$
	$RPL14$, $BTLA$, $TP63$, $CCR5$, $DOCK3$
	$CTNNB1$, $IGF2BP2$, $MYD88$, $LIPH$
	$PFKFB4$, $PIK3CB$, $VPRBP$, $TLR9$, VHL
	$LRRN1$, $PAK2$, $PPP1R2$, $EAF2$
	TF, $VGLL4$, $RASSF1$, $FHIT$, $BAP1$, $FLNB$
Reference	(Kumar et al., 2012; Zhang et al., 2012; Shin et al., 2006)
	(Wen et al., 2012; Feng et al., 2012; Zheng et al., 2007)
	(Chen et al., 2012; Mller et al., 2014; Hu et al., 2015)
	(Comtesse et al., 2007; de Miguel et al., 2014; Shriver et al., 1998)
	(Thommen et al., 2015; Wang et al., 2011; Cheng et al., 2016)
	(Zhou et al., 2015; Shigemitsu et al., 2001; Bell et al., 2013)
	(Coste et al., 2010; Seki et al., 2014; Minchenko et al., 2014)
	(Wee et al., 2008; Wang et al., 2013; Belmont et al., 2014)
	(Zhou et al., 2012; Dmitriev et al., 2012; Kikuchi et al., 2012)
	(Takakura et al., 2001; Xiao et al., 2008; Regina et al., 2008)
	(Zhang et al., 2014; Pelosi et al., 2010; Zchbauer-Mller et al., 200)
	(Carbone et al., 2013; Bandaru et al., 2014)
LME	$ADAMTS9$, $BAP1$, $C3orf33$
	$CCR5$, $CTNNB1$, $EHHADH$, $FHIT$, $FLNB$
	$GSK3B$, $IGF2BP2$, $IQCB1$
	$LIPH$, $NKIRAS1$, $PAK2$, $PPP1R2$
	$RPL14$, $SENP2$, $SIAH2$, $TP63$
	$UBA3$, $VPRBP$, $VGLL4$
Reference	(Kumar et al., 2012; Carbone et al., 2013; Hu et al., 2015)
	(Cheng et al., 2016; Shigemitsu et al., 2001; Comtesse et al., 2007)
	(Zchbauer-Mller et al., 200; Bandaru et al., 2014; Zheng et al., 2007)
	(Bell et al., 2013; de Miguel et al., 2014; Seki et al., 2014)
	(Braga et al., 2015; Kikuchi et al., 2012; Takakura et al., 2001)
	(Shriver et al., 1998; Wang et al., 2013; Mller et al., 2014; Wang et al., 2011)
	(Li et al., 2014; Wang et al., 2013; Zhang et al., 2014)

Table 7.2
List of genes detected by TTS but missed by *Limma* and *edgeR*

Test method	Gene list
Limma	$ADAMTS9$, $BAP1$, $C3orf33$
	$CCR5$, $FHIT$, $GSK3B$
	$IGF2BP2$, $LIPH$, $PAK2$
	$PPP1R2$, $RABL3$, $RBM5$
	$SETD2$, TF, $TP63$
Reference	(Kumar et al., 2012; Comtesse et al., 2007; Hu et al., 2015)
	(Cheng et al., 2016; Zchbauer-Mller et al., 200; Zheng et al., 2007)
	(Bell et al., 2013; Seki et al., 2014; Kikuchi et al., 2012)
	(Takakura et al., 2001; Zhang et al., 2016; Sutherland et al., 2010)
	(Walter et al., 2017; Regina et al., 2008; Wang et al., 2011)
edgeR	$BAP1$, $CACNA2D3$, $CAMK1$
	$CCR5$, $CD86$, $FBXL2$
	$GSK3B$, $LIPH$, $NKIRAS1$
	$PPP1R2$, $SETD2$, $SLC6A20$
Reference	(Comtesse et al., 2007; Li et al., 2013; Liu et al., 2015)
	(Cheng et al., 2016; Wroblewski et al., 2001; Chen et al., 2012)
	(Zheng et al., 2007; Seki et al., 2014; Braga et al., 2015)
	(Takakura et al., 2001; Walter et al., 2017; Tsou et al., 2007)

Table 7.3
List of genes detected by TTS but missed by $DESeq2$, part 1

Gene list	Reference
$ABCC5$, $ABHD5$, $ACTL6A$, $AGTR1$, $ALDH1L1$, $ATP11B$	(Pelosi et al., 2006; Ou et al., 2014; Sun et al., 2017; Guo et al., 2015; Oleinik et al., 2011; Qian et al., 2015)
$ATP1B3$, ATR, $B3GALNT1$, $BAP1$, $BCHE$, $BTLA$	(Mesri et al., 2013; Beumer et al., 2015; Umeyama et al., 2014; Comtesse et al., 2007; Brass et al., 1997; Thommen et al., 2015)
$C3orf1$, $C3orf21$, $C3orf33$, $CACNA2D2$, $CACNA2D3$, $CAMK1$	(Wu et al., 2014; Zhang et al., 2012; Hu et al., 2015; Carbone et al., 2003; Li et al., 2013; Liu et al., 2015)
$CBLB$, $CCDC37$, $CCR5$, $CD86$, $CDC25A$, $CDCP1$	(Li et al., 2016; Tessema et al., 2015; Cheng et al., 2016; Wroblewski et al., 2001; He et al., 2005; Chiu et al., 2015)
$CHL1$, $CLDN18$, $COPB2$, $CSTA$, $CTDSPL$, $CTNNB1$	(Senchenko et al., 2011; Micke et al., 2014; Erdogan et al., 2009; Butler et al., 2011; Senchenko et al., 2010; Shigemitsu et al., 2001)
$CX3CR1$, $DCBLD2$, $DCUN1D1$, $DLEC1$, $DOCK3$, $DTX3L$,	(Schmall et al., 2015; Butler et al., 2011; Yoo et al., 2012; Kwong et al., 2006; Zhou et al., 2015; Thang et al., 2015)
$DVL3$, $EAF2$, $EHHADH$, $EIF2A$, $EIF4A2$, $EIF4G1$	(Wei et al., 2008; Xiao et al., 2008; Comtesse et al., 2007; He et al., 2011; Shaoyan et al., 2013; Cao et al., 2016)
$EIF5A2$, $EPHB3$, $ETV5$, $FAM107A$, $FBXL2$, $FGD5$	(Xu et al., 2017; Ji et al., 2011; Zhang et al., 2017; Pastuszak-Lewandoska et al., 2015; Chen et al., 2012; Dmitriev et al., 2012)
$FHIT$, $FLNB$, $FNDC3B$, $FOXP1$, $FXR1$, $GATA2$	(Zchbauer-Mller et al., 200; Bandaru et al., 2014; Cai et al., 2012; Feng et al., 2012; Comtesse et al., 2007; Kumar et al., 2012)
$GORASP1$, $GSK3B$, $HDAC11$, $HES1$, $HYAL1$, $HYAL2$	(Dmitriev et al., 2012; Zheng et al., 2007; Koeneke et al., 2015; Baumgart et al., 2015; Wang et al., 2008)
$IGF2BP2$, $IL17RD$, $IL17RE$, $IQCB1$, $IQSEC1$, $KIAA1524$	(Bell et al., 2013; Wu et al., 2016; de Miguel et al., 2014; Dmitriev et al., 2012; De et al., 2014)
$LEPREL1$, $LIMD1$, $LIPH$, $LMCD1$, LPP, $LRIG1$,	(Sheng et al., 2016; Sharp et al., 2008; Seki et al., 2014; Chang et al., 2012; Kuriyama et al., 2016; Kvarnbrink et al., 2015)
$LRRN1$, LTF, $LZTFL1$, $MAGI1$, $MASP1$, $MBD4$	(Dmitriev et al., 2012; Iijimai et al., 2006; Wei et al., 2016; Dorr et al., 2015; Kang et al., 2009; Shin et al., 2006)
$MCM2$, $METTL6$, $MINA$, MME, $MYD88$, $MYLK$	(Ramnath et al., 2001; Tan et al., 2011; Thakur et al., 2015; Leithner et al., 2014; Coste et al., 2010; Tan et al., 2014)
$NEK10$, $NEK4$, $NISCH$, $NKIRAS1$, $NPRL2$	(Moniz et al., 2011; Nguyen et al., 2012; Ostrow et al., 2011; Braga et al., 2015; Ueda et al., 2006)

Table 7.4
List of genes detected by TTS but missed by $DESeq2$, part 2

Gene list	Reference
$OPA1$, $P2RY14$, $PAK2$, $PDCD6IP$	(Roberts et al., 2013; Wu et al., 2012; Kikuchi et al., 2012; Li et al., 2014)
$PFKFB4$, $PIK3CB$, $PLD1$, $PLS1$	(Minchenko et al., 2014; Wee et al., 2008; Chen et al., 2012; Erdogan et al., 2009)
$POLQ$, $PPP1R2$, $PTH1R$, $PTPRG$	(Wood et al., 2016; Takakura et al., 2001; Montgrain et al., 2015; Pitterle et al., 1998)
$RABL3$, $RAP2B$, $RASSF1$, $RFC4$	(Zhang et al., 2016; Peng et al., 2016; Pelosi et al., 2010; Erdogan et al., 2009)
$RNF7$, $RPL14$, $RPL22L1$, $RUVBL1$	(Lazar et al., 2013; Shriver et al., 1998; O'Leary et al., 2013; Yuan et al., 2016)
$RYBP$, $SATB1$, $SEMA3B$, $SENP2$	(Voruganti et al., 2015; Selinger et al., 2011; Loginov et al., 2015; Wang et al., 2013)
$SETD2$, $SIAH2$, $SLC4A7$, $SLC6A20$	(Walter et al., 2017; Mller et al., 2014; Gorbatenko et al., 2014; Tsou et al., 2007)
$SLCO2A1$, $SMARCC1$, $SPCS1$	(Zhu et al., 2015; DelBove et al., 2011; Too et al., 2012)
$TBL1XR1$, TF, $TFRC$	(Liu et al., 2007; Regina et al., 2008; Jiang et al., 2010)
$TGFBR2$, $THPO$, $THRB$, $TIGIT$	(Xu et al., 2007; Lazar et al., 2013; Buchhagen, DL., 1996; Zhang et al., 2016)
TKT, $TLR9$, $TNFSF10$, $TP63$	(Xu et al., 2016; Belmont et al., 2014; He et al., 2012; Wang et al., 2011)
$TRAIP$, $TRIM59$, $UBA3$, $UBE2E2$	(Soo et al., 2016; Zhan et al., 2015; Li et al., 2014; Dmitriev et al., 2012)
$VGLL4$, VHL, $VPRBP$, $WWTR1$	(Zhang et al., 2014; Zhou et al., 2012; Wang et al., 2013; Noguchi et al., 2014)
XPC, $ZMAT3$, $ZMYND10$	(Zhang et al., 2015; Wen et al., 2012; Guo et al., 2015)

REFERENCES

Al-Shahrour F, Daz-Uriarte R, Dopazo J. (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* **20**(4), 578–580.

Al-Shahrour F, Daz-Uriarte R, Dopazo J. (2004). Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics* **21**(13), 2988–2993.

Auer, P. L., Doerge, R. W. (2010). Statistical Design and Analysis of RNA Sequencing Data. *Genetics* **185,** 405–416.

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

Bloom, J. S., Khan, Z., Kruglyak, L., Singh, M., Caudy, A. A. (2009). Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics* **10,** 221–231.

Bullard, J. H., Purdom, E., Hansen, K. D., Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11,** 94–107.

Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M., Sherlock, G. (2004). GO::TermFinderopen source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes *Bioinformatics* **20**(18), 3710.

Chu. C., Fang, Hua., Yang, Y., Chen, E., Cowley, AW Jr., Liang, M., Liu, P., Lu, Y. (2015) deGPS is a powerful tool for detecting differential expression in RNA-sequencing studies. *BMC Genomic* **16,** 1.

Dennis, G. Jr, Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., Lempicki, R.A.(2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol* **4,** R60.

Gutenbrunner, C., Jureckova, J., Koenker, R., Portnoy, S. (1993). Tests of linear hypotheses based on regression rank scores. *J Nonparametr Stat* **2,** 307-331.

Hardcastle, T. J., Kelly, K. A. (2010). BaySeq: Empirical Bayesian Methods For Identifying Differential Expression In Sequence Count Data. *BMC Bioinformatics* **11,** 422–436.

He, X. M., Hsu, Y. H., HU, M. X. (2010). Detection of treatment effects by covariate-adjusted expected shortfall. *Ann Appl Stat* **4,** 2114–2125.

He, X. M., Shao, Q. M. (1996). A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *Ann Stat* **24,** 2608–2630.

He, X. M., Zhu, Z. Y., Fung, W. K. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika* **89,** 579–590.

Hsu, Y. H. (2010). Applications of quantile regression to estimation and detection of some tail characteristics. Ph.D. Dissertation, University of Illinois at Urbana-Champaign.

Irizarry RA, Wang C, Zhou Y, Speed TP. (2009).Gene set enrichment analysis made simple. *Stat Methods Med Res* **18**(6), 565–575.

Jiang, H., Wong, W. H. (2009). Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* **25,**1026–1032.

Kanehisa, M., Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**(1), 27–30.

Khatri, P., Drghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* **21**(18), 3587–3595.

Koenker, R., Bassett, G. (1978). Regression quantiles. *Econometrica* **46,** 33–50.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Lader, A., Ramoni, M., Zetter, B., Kohane, I., Kwiatkowski, D.(2004). Identification of a transcriptional profile associated with in vitro invasion in non-small cell lung cancer cell lines. *Cancer Biol Ther* **3,** 624–631.

Laiho, A., Elo, L.L(2014). A note on an exon-based strategy to identify differentially expressed genes in RNA-Seq experiments. *PloS one* **9**(12), e115964.

Laird, N. M., Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38,** 963–974.

Lin, W., Sun, F. (2012). CEDER: Accurate detection of differentially expressed genes by combining significance of exons using RNA-Seq *IEEE/ACM Trans Comput Biol Bioinform* **9,** 1281-1292.

Liu, X., Milo, M., Lawrence, ND., Rattray, M. (2006). Probe-level measurement error improves accuracy in detecting differential gene expression. *Bioinformatics* **22,** 2107-2113.

Love, MI., Huber, W., Anders. S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550

Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D., Woolf, P. J. (2009). GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* **10**, 161

Luo, W., Brouwer, C. (2013). Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* **29**(14), 1830–1831

Maere,S., Heymans,K. Kuiper,M. (2005). Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21,** 3448–3449.

Marileila, V. G. (2010). Chromosomal and genomic changes in lung cancer. *Cell Adh Migr* **4,** 100–106.

McCarthy, D. J., Chen, Y., Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* **40,** 4288–4297.

Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrle M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* **34**(3), 267–273.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Method* **5,** 621–628.

Owen, A. B. (2001). *Empirical Likelihood.*Chapman & Hall/CRC.

Pavlidis, P., Qin, J., Arango, V., Mann, JJ. (2004). Using the Gene Ontology for Microarray Data Mining: A Comparison of Methods and Application to Age Effects in Human Prefrontal Cortex. *Neurochemical Research* **29**(6), 1213–1222.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43,** 7.

Rivals, I., Personnaz, L., Taing, L., Potier, MC. (2007). Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics* **23**(4), 401–407. Bioinformatics. 2007 Feb 15;23(4):401-7. Epub 2006 Dec 20.

Robinson, MD., McCarthy, DJ., Smyth., GK. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 1.

Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. *London: Chapman and Hall.*

Shanker S, Paulson A, Edenberg HJ, Peak A, Perera A, Alekseyev YO, Beckloff N, Bivens NJ, Donnelly R, Gillaspy AF, Grove D, Gu W, Jafari N, Kerley-Hamilton JS, Lyons RH, Tepper C, Nicolet CM. (2015). Evaluation of commercially available RNA amplification kits for RNA sequencing using very low input amounts of total RNA. *J Biomol Tech* **26**(1), 4–18.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA.* **102**(42), 15545–15550.

Subramanian, J., Govindan, R. (2007). Lung cancer in never smokers: a review. *J Clin Oncol* **25,** 561–570.

Vremo, L., Nielsen, J., Nookaew, I. (2013). Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Res* **41**(8), 4378–4391.

Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ.(2005). Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci U S A.* **102**(38), 13544–13549.

Wan, L., Sun, F. (2012). CEDER: Accurate detection of differentially expressed genes by combining significance of exons using RNA-Seq. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **9,** 1281 - 1292.

Wang, H., He. X. (2008). An enhanced quantile approach for assessing differential gene expressions. *Biometrics* **64,** 449–457.

Appendix Reference

Bandaru, S., Zhou, AX., Rouhi, P., Zhang, Y., Bergo, MO., Cao, Y., Akyrek, LM. (2014). Targeting filamin B induces tumor growth and metastasis via enhanced activity of matrix metalloproteinase-9 and secretion of VEGF-A. *Oncogenesis* **3,** e119.

Baumgart, A., Mazur, PK., Anton, M., Rudelius, M., Schwamborn, K., Feuchtinger, A., Behnke, K., Walch, A., Braren, R., Peschel, C., Duyster, J., Siveke, JT., Dechow, T. (2015). Opposing role of Notch1 and Notch2 in a Kras(G12D)-driven murine non-small cell lung cancer model. *Oncogene* **34**(5) 578–588.

Bell, J. L., Wchter, K., Mhleck, B., Pazaitis, N., Khn, M., Lederer, M., Httelmaier, S. (2013). Insulin-like growth factor 2 mRNA-binding proteins (IGF2BPs): post-transcriptional drivers of cancer progression? *Cell Mol Life Sci* **70**(15), 2657–2675.

Belmont, L., Rabbe, N., Antoine, M., Cathelin, D., Guignabert, C., Kurie, J., Cadranel, J., Wislez, M. (2014). Expression of TLR9 in tumor-infiltrating mononuclear cells enhances angiogenesis and is associated with a worse survival in lung cancer. *Int J Cancer* **134,** 765–777.

Beumer, JH., Fu, KY., Anyang, BN., Siegfried, JM., Bakkenist, CJ. (2015). Functional analyses of ATM, ATR and Fanconi anemia proteins in lung carcinoma : ATM, ATR and FA in lung carcinoma. *BMC Cancer* **15,** 649.

Braga, EA., Loginov, VI., Pronina, IV., Khodyrev, DS., Rykov, SV., Burdennyy, AM., Friedman, MV., Kazubskaya, TP., Kubatiev, AA., Kushlinskii, NE. (2015).

Upregulation of RHOA and NKIRAS1 genes in lung tumors is associated with loss of their methylation as well as with methylation of regulatory miRNA genes. *Biochemistry* **80**(4), 483–494.

Brass, N., Rcz, A., Heckel, D., Remberger, K., Sybrecht, GW., Meese, EU. (1997). Amplification of the genes BCHE and SLC2A2 in 40% of squamous cell carcinoma of the lung. *Cancer Res* **57**(11), 2290–2294.

Buchhagen, DL. (1996). Frequent involvement of chromosome 3p alterations in lung carcinogenesis: allelotypes of 215 established cell lines at six chromosome 3p loci. *J Cell Biochem* **24**,198–209.

Butler MW, Fukui T, Salit J, Shaykhiev R, Mezey JG, Hackett NR, Crystal RG. (2011). Modulation of cystatin A expression in human airway epithelium related to genotype, smoking, COPD, and lung cancer. *Cancer Res* **71**(7), 2572–2581.

Cai, C., Rajaram, M., Zhou, X., Liu, Q., Marchica, J., Li, J., Powers, R. S. (2012). Activation of multiple cancer pathways and tumor maintenance function of the 3q amplified oncogene FNDC3B. *Cell Cycle* **11**(9), 1773-1781.

Cao, Y., Wei, M., Li, B., Liu, Y., Lu, Y., Tang, Z., Lu, T., Yin, Y., Qin, Z., Xu, Z. (2016). Functional role of eukaryotic translation initiation factor 4 gamma 1 (EIF4G1) in NSCLC. *Oncotarget* **7**(17), 24242-24251.

Carbone, M., Yang, H., Pass, H. I., Krausz, T., Testa, J. R., Gaudino, G. (2013). BAP1 and Cancer. *Nature Reviews. Cancer* **13**(3), 153–159.

Carboni GL, Gao B, Nishizaki M, Xu K, Minna JD, Roth JA, Ji L. (2003). CACNA2D2-mediated apoptosis in NSCLC cells is associated with alterations of the intracellular calcium signaling and disruption of mitochondria membrane integrity. *Oncogene* **22**(4), 615–626.

Dorr, C., Janik, C., Weg, M., Been, R. A., Bader, J., Kang, R., Ng, B., Foran, L., Landman, SR., OSullivan, M.G., Steinbach, M., Sarver, A.L., Silverstein, K., Largaespada, DA., Starr, T. K. (2015). Transposon Mutagenesis Screen Identifies Potential Lung Cancer Drivers and CUL3 as a Tumor Suppressor. *Molecular Cancer Research: MCR* **13**(8), 1238-1247.

Chang, CY., Lin, SC., Su, WH., Ho, CM., Jou, YS. (2012). Somatic LMCD1 mutations promoted cell migration and tumor metastasis in hepatocellular carcinoma. *Oncogene* **31**(21), 2640-2652.

Chen, BB., Glasser, JR., Coon, TA., Mallampalli, RK. (2012). F-box protein FBXL2 exerts human lung tumor suppressor-like activity by ubiquitin-mediated degradation of cyclin D3 resulting in cell cycle arrest. *Oncogene* **31**(20), 2566–2579.

Chen, Q., Hongu, T., Sato, T., Zhang, Y., Ali, W., Cavallo, J. A., van der Velden, A., Tian, H., Di Paolo, G., Nieswandt, B., Kanaho, Y., Frohman, M. A. (2012). Key roles for the lipid signaling enzyme phospholipase d1 in the tumor microenvironment during tumor angiogenesis and metastasis. *Sci Signal* **5,** ra79.

Cheng, Z., Shi, Y., Yuan, M., Xiong, D., Zheng, J., Zhang, Z. (2016). Chemokines and their receptors in lung cancer progression and metastasis. *Journal of Zhejiang University. Science. B* **17**(5), 342351.

Chiu, KL., Kuo, TT., Kuok, QY., Lin, YS., Hua, CH., Lin, CY., Su, PY., Lai. LC., Sher, YP. (2015). ADAM9 enhances CDCP1 protein expression by suppressing miR-218 for lung tumor metastasis. Scientific. *Scientific Reports* **5**, 16426.

Comtesse, N., Keller, A., Diesinger, I., Bauer, C., Kayser, K., Huwer, H., Lenhof, H. P., Meese, E. (2007). Frequent overexpression of the genes FXR1, CLAPM1 and EIF4G located on amplicon 3q26-27 in squamous cell carcinoma of the lung. *Int J Cancer* **120,** 2538–2544.

Coste, I., Le Corf, K., Kfoury, A., Hmitou, I., Druillennec, S., Hainaut, P., Eychene, A., Lebecque, S., Renno, T. (2010). Dual function of MyD88 in RAS signaling and inflammation, leading to mouse and human cell transformation. *J Clin Invest* **120,** 3663–3667.

de Miguel, F. J., Sharma, R. D., Pajares, M. J., Montuenga,L. M., Rubio, A., Pio, R. (2014). Identification of alternative splicing events regulated by the oncogenic factor SRSF1 in lung cancer. *Cancer Res* **74,** 1105–1115.

De, P., Carlson, J., Leyland-Jones, B., Dey, N. (2014). Oncogenic nexus of cancerous inhibitor of protein phosphatase 2A (CIP2A): An oncoprotein with many hands. *Oncotarget* **5**(13), 4581–4602.

DelBove, J., Rosson, G., Strobeck, M., Chen, J., Archer, T. K., Wang, W., Knudesn, E. S., Weissman, B. E. (2011). Identification of a core member of the SWI/SNF complex, BAF155/SMARCC1, as a human tumor suppressor gene. *Epigenetics* **6**(12), 1444–1453.

Dmitriev, A. A., Kashuba, V. I., Haraldson, K., Senchenko, V. N., Pavlova, T. V., Kudryavtseva, A. V., Anedchenko, E. A., Krasnov, G. S., Pronina, I. V., Loginov, V. I., Kondratieva, T. T., Kazubskaya, T. P., Braga, E. A., Yenamandra, S. P., Ignatjev, I., Ernberg, I., Klein, G., Lerman, M. I., Zabarovsky, E. R. (2012). Genetic and epigenetic analysis of non-small cell lung cancer with NotI-microarrays. *Epigenetics* **7,** 502–513.

Erdogan, E., Klee, EW., Thompson, EA., Fields, AP.(2009). Meta-analysis of oncogenic protein kinase Ciota signaling in lung adenocarcinoma. *Clin Cancer Res* **15**(5), 1527–1533.

Feng, H., Lopez, GY., Kim, CK., Alvarez, A., Duncan, CG., Nishikawa, R., Nagane, M., Su, AJ., Auron, PE., Hedberg, ML., Wang, L., Raizer, JJ., Kessler, JA., Parsa, AT., Gao, WQ., Kim, SH., Minata, M., Nakano, I., Grandis, JR., McLendon, RE., Bigner, DD., Lin, HK., Furnari, FB., Cavenee, WK., Hu, B., Yan, H., Cheng, SY. (2012). EGFR phosphorylation of DCBLD2 recruits TRAF6 and stimulates AKT-promoted tumorigenesis. *J Clin Invest.* **124**(9), 3741–3756.

Feng, J., Zhang, X., Zhu, H., Wang, X., Ni, S., Huang, J. (2012). High expression of FoxP1 is associated with improved survival in patients with non-small cell lung

cancer. *Am J Clin Pathol* **138,** 230–235.

Gorbatenko, A., Olesen, C. W., Boedtkjer, E., Pedersen, S. F. (2014). Regulation and roles of bicarbonate transporters in cancer. *Frontiers in Physiology* **5,** 130.

Guo, S., Yan, F., Xu, J., Bao, Y., Zhu, J., Wang, X., Wu, J., Li, Y., Pu, W., Liu, Y., Jiang, Z., Ma, Y., Chen, X., Xiong, M., Jin, L., Wang, J. (2015). Identification and validation of the methylation biomarkers of non-small cell lung cancer (NSCLC). *Clinical Epigenetics* **7**(1) 3.

He, N., Li, C., Zhang, X., Sheng, T., Chi, S., Chen, K., Wang, Q., Vertrees, R., Logrono, R., Xie, J. (2005). Regulation of lung cancer cell growth and invasiveness by beta-TRCP. *Mol Carcinog* **42**(1), 18–28.

He, W., Wang, Q., Xu, J., Xu, X., Padilla, M. T., Ren, G., Gou, X., Lin, Y. (2012). Attenuation of TNFSF10/TRAIL-induced apoptosis by an autophagic survival pathway involving TRAF2- and RIPK1/RIP1-mediated MAPK8/JNK activation. *Autophagy* **8**(12), 1811–1821.

He, Y., Correa, AM., Raso, MG., Hofstetter, WL., Fang, B., Behrens, C., Roth, JA., Zhou, Y., Yu, L., Wistuba, II., Swisher, SG., Pataer, A. (2011). The role of PKR/eIF2 signaling pathway in prognosis of non-small cell lung cancer. *PLoS One* **6**(11), e24855.

Hu, F., Yang, S., Lv. S., Peng. Y., Meng. L., Gou. L., Zhang. X. (2015). Analysis of AC3-33 gene expression in multiple organ cancer and adjacent normal tissue by RNA in situ hybridization. *Oncol Lett* **9,** 2795–2798.

Iijima, H., Tomizawa, Y., Iwasaki, Y., Sato, K., Sunaga, N., Dobashi, K., Saito, R., Nakajima, T., Minna, JD., Mori, M. (2006). Genetic and epigenetic inactivation of LTF gene at 3p21.3 in lung cancers. *Int J Cancer* **118**(4), 797–801.

Ji, XD., Li, G., Feng, YX., Zhao, JS., Li, JJ., Sun, ZJ., Shi, S., Deng, YZ., Xu, JF., Zhu, YQ., Koeffler, HP., Tong, XJ., Xie, D.(2011). EphB3 is overexpressed in non-small-cell lung cancer and promotes tumor metastasis by enhancing cell survival and migration. *Cancer Res* **71**(3) 1156–1166.

Jiang, XP., Elliott, RL., Head, JF. (2010). Manipulation of iron transporter genes results in the suppression of human and mouse mammary adenocarcinomas. *Anticancer Res* **30**(3) 759–765.

Kang, J. U., Koo, S. H., Kwon, K. C., Park, J. W., Kim, J. M. (2009). Identification of novel candidate target genes, including EPHB3, MASP1 and SST at 3q26.2q29 in squamous cell carcinoma of the lung. *BMC Cancer* **9,** 237.

Kikuchi, T., Hassanein, M., Amann, J. M., Liu, Q. F., Slebos, R. J., Rahman, J., Kaufman, J. M., Zhang, X. Q., Hoeksema, M. D., Harris, B. K., Li, M., Shyr, Y., Gonzalez, A. L., Zimmerman, L. J., Liebler, D. C., Massion, P. P., Carbone, D. P. (2012). In-depth Proteomic Analysis of Nonsmall Cell Lung Cancer to Discover Molecular Targets and Candidate Biomarkers. *Mol Cell Proteomics* **11,** 916–932.

Koeneke, E., Witt, O., Oehme, I. (2015). HDAC Family Members Intertwined in the Regulation of Autophagy: A Druggable Vulnerability in Aggressive Tumor Entities. *Cells* **4**(2), 135-168.

Kumar, MS., Hancock, DC., Molina-Arcas, M., Steckel, M., East, P., Diefenbacher, M., Armenteros-Monterroso, E., Lassailly, F., Matthews, N., Nye, E., Stamp, G., Behrens, A., Downward, J. (2012). The GATA2 transcriptional network is requisite for RAS oncogene-driven non-small cell lung cancer. *Cell* **149**(3), 642–655.

Kumar, S., Rao, N., Ge, R. (2012). Emerging Roles of ADAMTSs in Angiogenesis and Cancer.*Cancers* **4**(4), 12521299.

Kuriyama, S., Yoshida, M., Yano, S., Aiba, N., Kohno, T., Minamiya, Y., Goto, A., Tanaka, M. (2016). LPP inhibits collective cell migration during lung cancer dissemination. *Oncogene* **35**(8), 952–964.

Kvarnbrink, S., Karlsson, T., Edlund, K., Botling, J., Lindquist, D., Jirstrm, K., Micke, P., Henriksson, R., Johansson, M., Hedman, H. (2015). LRIG1 is a prognostic biomarker in non-small cell lung cancer. *Acta Oncol* **54**(8), 1113–1119.

Kwong, J., Lee, JY., Wong, KK., Zhou, X., Wong, DT., Lo, KW., Welch, WR., Berkowitz, RS., Mok, SC. (2006). Candidate tumor-suppressor gene DLEC1 is frequently downregulated by promoter hypermethylation and histone hypoacetylation in human epithelial ovarian cancer. *Neoplasia* **8**(4), 268–278.

Lazar, V., Suo, C., Orear, C., van den Oord, J., Balogh, Z., Guegan, J., Job, B., Meurice, G., Ripoche, H., Calza, S., Hasmats, J., Lundeberg, J., Lacroix, L., Vielh, P., Dufour, F., Lehti, J., Napieralski, R., Eggermont, A., Schmitt, M., Cadranel, J., Besse, B., Girard, P., Blackhall, F., Validire, P., Soria, JC., Dessen, P., Hansson, J., Pawitan, Y. (2013). Integrated molecular portrait of non-small cell lung cancers. *BMC Med Genomics* **6**, 53.

Leithner, K., Wohlkoenig, C., Stacher, E., Lindenmann, J., Hofmann, NA., Gall, B., Guelly, C., Quehenberger, F., Stiegler, P., Smolle-Jttner, FM., Philipsen, S., Popper, HH., Hrzenjak, A., Olschewski, A., Olschewski, H. (2014). Hypoxia increases membrane metallo-endopeptidase expression in a novel lung cancer ex vivo model - role of tumor stroma cells. *BMC Cancer* **14**, 40.

Li, L.H., Wang, M.S., Yu, G. Y., Chen, P., Li, H., Wei, D.P., Zhu, J., Xie, L., Jia, H. X., Shi, J. Y., Li, C. J., Yao, W. T., Wang, Y. C., Gao, Q., Jeong, L. S., Lee, H. W., Yu, J. H., Hu, F. Q., Mei, J., Wang, P., Chu, Y. W., Qi, H., Yang, M., Dong, Z. M., Sun, Y., Hoffman, R. M., Jia, L. J. (2014). Overactivated Neddylation Pathway as a Therapeutic Target in Lung Cancer. *J Natl Cancer Inst* **106,** dju083.

Li, P., Wang, X., Liu, Z., Liu, H., Xu, T., Wang, H., Gomez, DR., Nguyen, QN., Wang, LE., Teng, Y., Song, Y., Komaki, R., Welsh, JW., Wei, Q., Liao, Z. (2016). Single Nucleotide Polymorphisms in CBLB, a Regulator of T-Cell Response, Predict Radiation Pneumonitis and Outcomes After Definitive Radiotherapy for Non-Small-Cell Lung Cancer. *Clin Lung Cancer* **17**(4), 253–262

Li, Y., Zhu, CL., Nie, CJ., Li, JC., Zeng, T., Zhou, J., Chen, J., Chen, K., Li, F., Liu, HB., Qin, YR., Guan, XY. (2013). Investigation of Tumor Suppressing Function of CACNA2D3 in Esophageal Squamous Cell Carcinoma. *PLoS ONE* **8**(4), e60027.

Liu, SG., Yuan, SH., Wu, HY., Huang, CS., Liu, J. (2014). The programmed cell death 6 interacting protein insertion/deletion polymorphism is associated with non-small cell lung cancer risk in a Chinese Han population. *Tumour Bio* **35**(9), 8679–8683.

Liu, X., Yu, X., Xie, J., Zhan, M., Yu, Z., Xie, L., Zeng, H., Zhang, F., Chen, G., Yi, X., Zheng, J. (2015). ANGPTL2/LILRB2 signaling promotes the propagation of lung cancer cells. *Oncotarget* **6**(25), 21004–21015.

Liu, Y., Sun, W., Zhang, K., Zheng, H., Ma, Y., Lin, D., Zhang, X., Feng, L., Lei, W., Zhang, Z., Guo, S., Han, N., Tong, W., Feng, X., Gao, Y., Cheng, S. (2007). Identification of genes differentially expressed in human primary lung squamous cell carcinoma. *Lung Cancer* **56**(3), 307–317.

Loginov, VI., Dmitriev, AA., Senchenko, VN., Pronina, IV., Khodyrev, DS., Kudryavtseva, AV., Krasnov, GS., Gerashchenko, GV., Chashchina, LI., Kazubskaya, TP., Kondratieva, TT., Lerman, MI., Angeloni, D., Braga, EA., Kashuba, VI. (2015). Tumor Suppressor Function of the SEMA3B Gene in Human Lung and Renal Cancers. *PLoS One* **10**(5), e0123369.

Mesri, M., Birse, C., Heidbrink, J., McKinnon, K., Brand, E., Bermingham, CL., Feild, B., Fitzhugh, W., He, T., Ruben, S., Moore, PA. (2013). Identification and

characterization of angiogenesis targets through proteomic profiling of endothelial cells in human cancer tissues. *PLoS One* **8**(11), e78885.

Micke, P., Mattsson, JS., Edlund, K., Lohr, M., Jirstrm, K., Berglund, A., Botling, J., Rahnenfuehrer, J., Marincevic, M., Pontn, F., Ekman, S., Hengstler, J., Wll, S., Sahin, U., Treci, O. (2014). Aberrantly activated claudin 6 and 18.2 as potential therapy targets in non-small-cell lung cancer. *Int J Cancer* **135**(9), 2206–2214.

Minchenko, O. H., Tsuchihara, K., Minchenko, D. O., Bikfalvi, A., Esumi, H. (2014). Mechanisms of regulation of PFKFB expression in pancreatic and gastric cancer cells. *World J Gastroenterol* **14,** 13705–12717.

Moniz, L. S., Stambolic, V. (2011). Nek10 Mediates G2/M Cell Cycle Arrest and MEK Autoactivation in Response to UV Irradiation . *Molecular and Cellular Biology* **31**(1), 30–42.

Montgrain, PR., Phun, J., Vander Werff, R., Quintana, RA., Davani, AJ., Hastings, RH. (2015). Parathyroid-hormone-related protein signaling mechanisms in lung carcinoma growth inhibition. *Springerplus* **4**, 268.

Mller, S., Chen, Y., Ginter, T., Schfer, C., Buchwald, M., Schmitz, L. M., Klitzsch, J., Schtz, A., Haitel, A., Schmid, K., Moriggl, R., Kenner, L., Friedrich, K., Haan, C., Petersen, I., Heinzel, T., Krmer, O. H. (2014). SIAH2 antagonizes TYK2-STAT3 signaling in lung carcinoma cells. *Oncotarget* **30,** 3184–3196.

Nguyen, C. L., Possemato, R., Bauerlein, E. L., Xie, A., Scully, R., Hahn, W. C. (2012). Nek4 Regulates Entry into Replicative Senescence and the Response to DNA Damage in Human Fibroblasts. *Molecular and Cellular Biology* **32**(19), 3963-3977.

Noguchi, S., Saito, A., Horie, M., Mikami, Y., Suzuki, HI., Morishita, Y., Ohshima, M., Abiko, Y., Mattsson, JS., Knig, H., Lohr, M., Edlund, K., Botling, J., Micke, P., Nagase, T. (2014). An integrative analysis of the tumorigenic role of TAZ in human non-small cell lung cancer. *Clin Cancer Res* **20**(17), 4660–4672.

Oguri T, Achiwa H, Sato S, Bessho Y, Takano Y, Miyazaki M, Muramatsu H, Maeda H, Niimi T, Ueda R. (2006). The determinants of sensitivity and acquired resistance to gemcitabine differ in non-small cell lung cancer: a role of ABCC5 in gemcitabine sensitivity. *Mol Cancer Ther* **5**(7), 1800–1806.

O'Leary, MN., Schreiber, KH., Zhang, Y., Duc, A-CE., Rao, S., Hale, JS., Academia, EC., Shah, SR., Morton, JF., Holstein, CA., Martin, DB., Kaeberlein, M., Ladiges, WC., Fink, PJ., Mackay, VL., Wiest, DL., Kennedy, BK. (2013) The Ribosomal Protein Rpl22 Controls Ribosome Composition by Directly Repressing Expression of Its Own Paralog, Rpl22l1. *PLoS Genet* **9**(8), e1003708.

Oleinik, NV., Krupenko, NI., Krupenko, SA. (2011). Epigenetic Silencing of ALDH1L1, a Metabolic Regulator of Cellular Proliferation, in Cancers. *Genes Cancer* **2**(2), 130–139.

Ostrow, KL., Michailidi, C., Guerrero-Preston, R., Hoque, MO., Greenberg, A., Rom, W., Sidransky, D. (2011). Cigarette smoke induces methylation of the tumor suppressor gene NISCH. *Epigenetics* **8**(4), 383–388.

Ou, J., Miao, H., Ma, Y., Guo, F., Deng, J., Wei, X., Zhou, J., Xie, GF., Shi, H., Xue, B., Liang, H., Yu, L. (2014). Loss of Abhd5 Promotes Colorectal Tumor Development and Progression by Inducing Aerobic Glycolysis and Epithelial-Mesenchymal Transition. *Cell Reports* **9**(5), 1798–1811.

Pitterle DM, Jolicoeur EM, Bepler G. (1998). Hot spots for molecular genetic alterations in lung cancer. *In Vivo* **12**(6), 643–658.

Pastuszak-Lewandoska D, Czarnecka KH, Migdalska-Sk M, Nawrot E, Domaska D, Kiszakiewicz J, Kordiak J, Antczak A, Grski P, Brzeziaska-Lasota E. (2015). Decreased FAM107A Expression in Patients with Non-small Cell Lung Cancer. *Adv Exp Med Biol* **852**, 39–48.

Pelosi, G., Fumagalli, C., Trubia, M., Sonzogni, A., Rekhtman, N., Maisonneuve, P., Galetta, D., Spaggiari, L., Veronesi, G., Scarpa, A., Malpeli, G., Viale, G. (2010). Dual role of RASSF1 as a tumor suppressor and an oncogene in neuroendocrine tumors of the lung. *Anticancer Res* **30,** 4269–4281.

Peng, YG., Zhang, ZQ., Chen, YB., Huang, JA. (2016). Rap2b promotes proliferation, migration, and invasion of lung cancer cells. *J Recept Signal Transduct Res* **36**(5), 459–464.

Qian J, Hassanein M, Hoeksema MD, Harris BK, Zou Y, Chen H, Lu P, Eisenberg R, Wang J, Espinosa A, Ji X, Harris FT, Rahman SM, Massion PP. (2015). The RNA binding protein FXR1 is a new driver in the 3q26-29 amplicon and predicts poor prognosis in human cancers. *Proc Natl Acad Sci U S A*. **112**(11), 3469–3474

Qian, J., Zou, Y., Wang, J., Zhang, B., Massion, P. P. (2015). Global gene expression profiling reveals a suppressed immune response pathway associated with 3q amplification in squamous carcinoma of the lung. *Genomics Data* **5**, 272–274.

Ramnath, N., Hernandez, FJ., Tan, DF., Huberman, JA., Natarajan, N., Beck, AF., Hyland, A., Todorov, IT., Brooks, JS., Bepler, G. (2001). MCM2 is an independent predictor of survival in patients with non-small-cell lung cancer. *J Clin Oncol* **19**(22), 4259–4266.

Regina, S., Rollin, J., Blchet, C., Iochmann, S., Reverdiau, P., Gruel, Y. (2008). Tissue factor expression in non-small cell lung cancer: relationship with vascular endothelial growth factor expression, microvascular density, and K-ras mutation. *J Thorac Oncol* **3,** 689–697.

Roberts, E. R., Thomas, K. J. (2013). The role of mitochondria in the development and progression of lung cancer. *Computational and Structural Biotechnology Journal* **6,** e201303019.

Schmall, A., Al-Tamari, HM., Herold, S., Kampschulte, M., Weigert, A., Wietelmann, A., Vipotnik, N., Grimminger, F., Seeger, W., Pullamsetti, SS., Savai, R.

(2015). Macrophage and cancer cell cross-talk via CCR2 and CX3CR1 is a fundamental mechanism driving lung cancer. *Am J Respir Crit Care Med* **191**(4), 437–447.

Seki, Y., Yoshida, Y., Ishimine, H., Shinozaki-Ushiku, A., Ito, Y., Sumitomo, K., Nakajima, J., Fukayama, M., Michiue, T., Asashima, M., Kurisaki, A. (2014). Lipase member H is a novel secreted protein selectively upregulated in human lung adenocarcinomas and bronchioloalveolar carcinomas. *Biochem Biophys Res Commun* **443**(4), 1141–1147.

Selinger, CI., Cooper, WA., Al-Sohaily, S., Mladenova, DN., Pangon, L., Kennedy, CW., McCaughan, BC., Stirzaker, C., Kohonen-Corish, MR. (2011). Loss of special AT-rich binding protein 1 expression is a marker of poor survival in lung cancer. *J Thorac Oncol* **6**(7), 1179-1189.

Senchenko, V., Anedchenko, E., Kondratieva, T., Krasnov, G., Dmitriev, A., Zabarovska, V., Pavlova, T., Kashuba, V., Lerman, M., Zabarovsky, E. (2010). Simultaneous down-regulation of tumor suppressor genes RBSP3/CTDSPL, NPRL2/G21 and RASSF1A in primary non-small cell lung cancer. *BMC Cancer* **10**, 75.

Senchenko, V., Krasnov, G., Dmitriev,A., Kudryavtseva, A., Anedchenko, E., Braga, E., Pronina, I., Kondratieva, T., Ivanov, S., Zabarovsky, E., Lerman, M. (2011). Loss of special AT-rich binding protein 1 expression is a marker of poor survival in lung cancer. *PLoS ONE* **6**(3), e15612.

Shaoyan, X., Juanjuan, Y., Yalan, T., Ping, H., Jianzhong, L., Qinian, W. (2013). Downregulation of EIF4A2 in non-small-cell lung cancer associates with poor prognosis. *Clin Lung Cancer* **14**(6), 658–665.

Sharp, TV., Al-Attar, A., Foxler, DE., Ding, L., de A Vallim, TQ., Zhang, Y., Nijmeh, HS., Webb, TM., Nicholson, AG., Zhang, Q., Kraja, A., Spendlove, I., Osborne, J., Mardis, E., Longmore, GD. (2008). The chromosome 3p21.3-encoded gene, LIMD1, is a critical tumor suppressor involved in human lung cancer development. *Proc Natl Acad Sci U S A.* **105**(50), 19932–19937.

Sheng, X., Wang, Z. (2016). Protein arginine methyltransferase 5 regulates multiple signaling pathways to promote lung cancer cell proliferation. *BMC Cancer* **16,** 567.

Shigemitsu, K., Sekido, Y., Usami, N., Mori, S., Sato, M., Horio, Y., Hasegawa, Y., Bader, S. A., Gazdar, A. F., Minna, J. D., Hida, T., Yoshioka, H., Imaizumi, M., Ueda, Y., Takahashi, M., Shimokata, K. (2001). Genetic alteration of the beta-catenin gene (CTNNB1) in human lung cancer and malignant mesothelioma and identification of a new 3p21.3 homozygous deletion. *Oncogene* **12,** 4249–4257.

Shin, M. C., Lee, S. J., Choi, J. E., Cha, S. I., Kim, C. H., Lee, W. K., Kam, S., Kang, Y. M., Jung, T. H., Park, J. Y. (2006). Glu346Lys Polymorphism in the Methyl-CpG Binding Domain 4 Gene and the Risk of Primary Lung Cancer. *Jpn J Clin Oncol* **36,** 483–488.

Shriver, S. P., Shriver, M. D., Tirpak, D. L., Bloch, L. M., Hunt, J. D., Ferrell, R. E., Siegfried, J. M. (1998). Trinucleotide repeat length variation in the human ribosomal

protein L14 gene (RPL14): localization to 3p21.3 and loss of heterozygosity in lung and oral cancers. *Mutat Res* **406,** 9–23.

Soo Lee, N., Jin Chung, H., Kim, H.-J., Yun Lee, S., Ji, J.-H., Seo, Y., Han, S.-H., Choi, M., Yun, M., Lee, S.-G., Myung, K., Kim, Y., Kang, H., Kim, H. (2016). TRAIP/RNF206 is required for recruitment of RAP80 to sites of DNA damage. *Nature Communications* **7,** 10463.

Sun, W., Wang, W., Lei, J., Li, H., Wu, Y. (2017). Actin-like protein 6A is a novel prognostic indicator promoting invasion and metastasis in osteosarcoma. *Oncol Rep* **37**(4), 2405–2417.

Sutherland, LC., Wang, K., Robinson, AG. (2010). RBM5 as a putative tumor suppressor gene for lung cancer. *J Thorac Oncol* **5**(3), 294–298.

Takakura, S., Kohno, T., Manda, R., Okamoto, A., Tanaka, T., Yokota, J. (2001). Genetic alterations and expression of the protein phosphatase 1 genes in human cancers. *Int J Oncol* **18,** 817–824.

Tan, XL., Moyer, AM., Fridley, BL., Schaid, DJ., Niu, N., Batzler, AJ., Jenkins, GD., Abo, RP., Li, L., Cunningham, JM., Sun, Z., Yang, P., Wang, L. (2011). Genetic variation predicting cisplatin cytotoxicity associated with overall survival in lung cancer patients receiving platinum-based chemotherapy. *Clinical Cancer Research* **17**(17), 5801-5811.

Tan, X., Chen, M. (2014). MYLK and MYL9 expression in non-small cell lung cancer identified by bioinformatics analysis of public expression data. *Tumour Biol* **35**(12), 12189–12200.

Tessema, M., Yingling, CM., Picchi, MA., Wu, G., Liu, Y., Weissfeld, JL., Siegfried, JM., Tesfaigzi, Y., Belinsky, SA. (2015). Epigenetic Repression of CCDC37 and MAP1B Links Chronic Obstructive Pulmonary Disease to Lung Cancer. *J Thorac Oncol* **10**(8), 1181–1188.

Thakur, C., Chen, F. (2015). Current understanding of mdig/MINA in human cancers. *Genes & Cancer* **6**(7-8), 288–302.

Thang, ND., Yajima, I., Kumasaka, M. Y., Iida, M., Suzuki, T., Kato, M. (2015). Deltex-3-like (DTX3L) stimulates metastasis of melanoma through FAK/PI3K/AKT but not MEK/ERK pathway. *Oncotarget* **6**(16), 14290–14299.

Thommen, D., Schreiner, J., Muller, P., Herzig, P., Roller, A., Belousov, A., Umana, P., Pisa, P., Klein, C., Bacac, M., Fischer, O., Moersig, W., Prince, S., Levitsky, V., Karanikas, V., Lardinois, D., Zippelius, A. (2015). Progression of Lung Cancer Is Associated with Increased Dysfunction of T Cells Defined by Coexpression of Multiple Inhibitory Receptors. *Cancer Immunol Res* **3**(12), 1344–1355.

Too, I. H. K., Ling, M. H. T. (2012). Signal Peptidase Complex Subunit 1 and Hydroxyacyl-CoA Dehydrogenase Beta Subunit Are Suitable Reference Genes in Human Lungs. *ISRN Bioinformatics* **2012,** 790452.

Tsou, J. A., Galler, J. S., Wali, A., Ye, W., Siegmund, K. D., Groshen, S., Laird, P., Turla, S., Koss, MN., Pass, HI., Laird-Offringa, I. A. (2007). DNA methylation profile of 28 potential marker loci in malignant mesothelioma. *Lung Cancer* **58**(2), 220–230.

Ueda, K., Kawashima, H., Ohtani, S., Deng, WG., Ravoori, M., Bankson, J., Gao, B., Girard, L., Minna, JD., Roth, JA., Kundra, V., Ji, L. (2006). The 3p21.3 tumor suppressor NPRL2 plays an important role in cisplatin-induced resistance in human non-small-cell lung cancer cells. *Cancer Res* **66**(19), 9682–9690.

Umeyama, H., Iwadate, M., Taguchi, YH. (2014). TINAGL1 and B3GALNT1 are potential therapy target genes to suppress metastasis in non-small cell lung cancer. *BMC Genomics* **15**, 1344–1355.

Voruganti, S., Xu, F., Qin, JJ., Guo, Y., Sarkar, S., Gao, M., Zheng, Z., Wang, MH., Zhou, J., Qian, B., Zhang, R., Wang, W. (2015). RYBP predicts survival of patients with non-small cell lung cancer and regulates tumor cell growth and the response to chemotherapy. *Cancer Lett* **369**(2), 386–395.

Walter, DM., Venancio, OS., Buza, EL., Tobias, JW., Deshpande, C., Gudiel, AA., Kim-Kiselak, C., Cicchini, M., Yates, TJ., Feldser, DM. (2017). Systematic In Vivo Inactivation of Chromatin-Regulating Enzymes Identifies Setd2 as a Potent Tumor Suppressor in Lung Adenocarcinoma. *Cancer Res* **77**(7) 1719–1729.

Wang, B. S., Liu, Y. Z., Yang, Y., Zhang, Y., Hao, J. J., Yang, H., Wang, X. M., Zhang, Z. Q., Zhan, Q. M., Wang, M. R. (2013). Autophagy negatively regulates cancer cell proliferation via selectively targeting VPRBP. *Clin Sci* **124,** 203–214.

Wang, F., Grigorieva, E. V., Li, J., Senchenko, V. N., Pavlova, T. V., Anedchenko, E. A., Kudryavtseva, A., Tsimanis, A., Angeloni, D., Lerman, M., Kashuba, Vladimir., Klein, G., Zabarovsky, E. R. (2008). HYAL1 and HYAL2 Inhibit Tumour Growth In Vivo but Not In Vitro. *PLoS ONE* **3**(8) e3031.

Wang, J., Qian, J., Hoeksema, MD., Zou, Y., Espinosa, AV., Rahman, SM., Zhang, B., Massion, PP. (2013). Integrative genomics analysis identifies candidate drivers at 3q26-29 amplicon in squamous cell carcinoma of the lung. *Clin Cancer Res* **19,** 5580–5590.

Wang, Y., Broderick, P., Matakidou, A., Vijayakrishnan, J., Eisen, T., Houlston, R. S. (2011). Variation in TP63 is associated with lung adenocarcinoma in the UK population. *Cancer Epidemiol Biomarkers Prev* **20,** 1453–1462.

Wee, S., Wiederschain, D., Maira, S. M., Loo, A., Miller, A., Beaumont, R., Stegmeier, F., Yao, Y. M., Lengauer, C. (2008). PTEN-deficient cancers depend on PIK3CB. *PNAS* **105,** 13057–13062.

Wei, Q., Zhao, Y., Yang, ZQ., Dong, QZ., Dong, XJ., Han, Y., Zhao, C., Wang, EH. (2008).Dishevelled family proteins are expressed in non-small cell lung cancer and function differentially on tumor progression. *Lung Cancer* **62**(2), 181–192.

Wei, Q., Chen, ZH., Wang, L., Zhang, T., Duan, L., Behrens, C., Wistuba, II., Minna, JD., Gao, B., Luo, JH., Liu, ZP. (2016). LZTFL1 suppresses lung tumorigenesis by maintaining differentiation of lung epithelial cells. *Oncogene* **35**(20), 2655–2663.

Wen, Y., Gamazon, E. R., Bleibel, W. K., Wing, C., Mi, S., McIlwee, B. E., Delaney, S. M., Duan, S., Im, H. K., Dolan, M. E. (2012) An eQTL-based method identifies CTTN and ZMAT3 as pemetrexed susceptibility markers. *Hum Mol Genet* **21,** 1470–1480.

Wood, R. D., Doubli, S.(2016). DNA polymerase (POLQ), double-strand break repair, and cancer. *DNA Repair* **44**, 22–32.

Wroblewsk,i JM., Bixby, DL., Borowski, C., Yannelli, JR. (2001). Characterization of human non-small cell lung cancer (NSCLC) cell lines for expression of MHC, costimulatory molecules and tumor-associated antigens. *Lung Cancer* **33**(2-3), 181–194.

Wu, F., Xu, J., Huang, Q., Han, J., Duan, L., Fan, J., Lv, Z., Guo, M., Hu, G., Chen, L., Zhang, S., Tao, X., Ma, W., Jin, Y. (2016). The Role of Interleukin-17 in Lung Cancer. *Mediators of Inflammation* **2016,** 84940790.

Wu, H., Wang, W., Xu, H. (2014). Depletion of C3orf1/TIMMDC1 Inhibits Migration and Proliferation in 95D Lung Carcinoma Cells. *International Journal of Molecular Sciences* **15**(11), 20555-20571.

Wu, X., Zang, W., Cui, S., Wang, M. (2012). Bioinformatics analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *Eur Rev Med Pharmacol Sci* **16**(11), 1582–1587.

Xiao, w., Zhang, Q., Habermacher, G., Yang, X., Zhang, A. Y., Cai, X., Hahn, J., Liu, J., Pins, M., Doglio, L., Dhir, R., Gingrich, J., Wang, Z. (2008). U19/Eaf2 knockout causes lung adenocarcinoma, B-cell lymphoma, hepatocellular carcinoma and prostatic intraepithelial neoplasia. *Oncogene* **6,** 1536–1544.

Xu, IM., Lai, RK., Lin, SH., Tse, AP., Chiu, DK., Koh, HY., Law, CT., Wong, CM., Cai, Z., Wong, CC., Ng, IO. (2016). Transketolase counteracts oxidative stress to drive cancer development. *Proc Natl Acad Sci U S A.* **113**(6), E725-34.

Xu, G., Shao, G., Pan, Q., Sun, L., Zheng, D., Li, M., Li, N., Shi, H., Ni, Y. (2017). MicroRNA-9 regulates non-small cell lung cancer cell invasion and migration by targeting eukaryotic translation initiation factor 5A2. *American Journal of Translational Research* **9**(2), 478488.

Xu, JB., Bao, Y., Liu, X., Liu, Y., Huang, S., Wang, JC.(2007). Defective expression of transforming growth factor beta type II receptor (TGFBR2) in the large cell variant of non-small cell lung carcinoma. *Lung Cancer* **58**(1), 36–43.

Yoo, J., Lee, SH., Lym, KI., Park, SY., Yang, SH., Yoo, CY., Jung, JH., Kang, SJ., Kang, CS. (2012). Immunohistochemical Expression of DCUN1D1 in Non-small Cell Lung Carcinoma: Its Relation to Brain Metastasis. *Cancer Res Treat* **44**(1), 57–62.

Yuan, XS., Wang, ZT., Hu, YJ., Bao, FC., Yuan, P., Zhang, C., Cao, JL., Lv, W., Hu, J. (2016) Downregulation of RUVBL1 inhibits proliferation of lung adenocarcinoma cells by G1/S phase cell cycle arrest via multiple mechanisms. *Tumor Biology* **37**(12), 16015–16027.

Zhan, W., Han, T., Zhang, C., Xie, C., Gan, M., Deng, K., Fu, M., Wang, J.(2015). TRIM59 Promotes the Proliferation and Migration of Non-Small Cell Lung Cancer Cells by Upregulating Cell Cycle Related Proteins. *PLOS ONE* **10**(11), e0142596.

Zhang L, Wang J, Wei F, Wang K, Sun Q, Yang F, Jin H, Zheng Y, Zhao H, Wang L, Yu W, Zhang X, An Y, Yang L, Zhang X, Ren X. (2016). Profiling the dynamic expression of checkpoint molecules on cytokine-induced killer cells from non-small-cell lung cancer patients. *Oncotarget* **7**(28), 43604–43615.

Zhang, W., Gao, Y., Li, P., Shi, Z., Guo, T., Li, F., Han, X., Feng, Y., Zheng, C., Wang, Z., Li, F., Chen, H., Zhou, Z., Zhang, L., Ji, H. (2014). VGLL4 functions as a new tumor suppressor in lung cancer by negatively regulating the YAP-TEAD transcriptional complex. *Cell Res* **24,** 331–343.

Zhang, W., Sun, J., Luo, J. (2016). High Expression of Rabl3 is Associated with Poor Survival of Patients with Non-Small Cell Lung Cancer via Repression of MAPK8/9/10-Mediated Autophagy. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research* **22,** 1582-1588.

Zhang, X., He, N., Gu, D., Wickliffe, J., Salazar, J., Boldogh, I., Xie, J. (2015). Genetic Evidence for XPC-KRAS Interactions During Lung Cancer Development. *J Genet Genomics* **42**(10), 589–596.

Zhang, Y., Gu, C., Shi, H., Zhang, A., Kong, X., Bao, W., Deng, D., Ren, L., Gu, D. (2012). Association between C3orf21, TP63 polymorphisms and environment and NSCLC in never-smoking Chinese population. *Gene* **497,** 93–97.

Zhang, Z., Newton, K., Kummerfeld, S. K., Webster, J., Kirkpatrick, D. S., Phu, L., Eastham-Anderson, J., Liu, J., Lee, W., Wu, J., Li, H., Junttila, M., Dixit, V. M. (2017). Transcription factor Etv5 is essential for the maintenance of alveolar type II cells. *Proc Natl Acad Sci U S A.* **114**915), 3903–3908.

Zheng, H., Saito, H., Masuda, S., Yang, X., Takano, Y. (2007) Phosphorylated GSK3beta-ser9 and EGFR are good prognostic factors for lung carcinomas. *Anticancer Res* **27,** 3561–3569.

Zhou, Q., Chen, T., Ibe, J. C., Raj, J. U., Zhou, G. (2012). Knockdown of von Hippel-Lindau protein decreases lung cancer cell proliferation and colonization. *FEBS Lett* **21,** 1510–1515.

Zhu, Q., Liang, X., Dai, J., Guan, X. (2015). Prostaglandin transporter, SLCO2A1, mediates the invasion and apoptosis of lung cancer cells via PI3K/AKT/mTOR pathway. *Int J Clin Exp Pathol* **8**(8), 9175–9181.

Zhu, X., Gao, G., Chu, K., Yang, X., Ren, S., Li, Y., Wu, H., Huang, Y., Zhou, C. (2015). Inhibition of RAC1-GEF DOCK3 by miR-512-3p contributes to suppression of metastasis in non-small cell lung cancer. *Int J Biochem Cell Biol* **61,** 103114.

Zchbauer-Mller, S., Wistuba, II., Minna, JD., Gazdar, AF. (2000). Fragile histidine triad (FHIT) gene abnormalities in lung cancer. *Clin Lung Cancer* **2,** 141–145.

8. Vita

Jiong Chen was born in Shanghai, China on June 16th, 1988, the son of Han Li and Xizhi Chen. After completing his high school work at Shanghai Foreign Language School, Shanghai, China in 2006, he went to the United States and entered Hamilton College in Clinton, NY. He received the degree of Bachelor of Arts with major in Mathematics and Economics from Hamilton in May, 2010. Then he entered Columbia University in New York, NY where he received the degree of Master of Arts with major in Statistics in Dec, 2011. In June of 2012 he entered the University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences to pursue the degree of Doctor of Philosophy in Biostatistics.

Permanent address:
Rm 1302. No 20, Lane 979,
Chang An Road, Shanghai, China, 200070