Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2021, Vol. 30(1) 261–276

DOI: 10.1177/0962280220951907
journals.sagepub.com/home/smm
SAGE

A tail-based test to detect differential expression in RNA-sequencing data

Jiong Chen[1,*], Xinlei Mi[2,*], Jing Ning[3], Xuming He[4] and Jianhua Hu[2]

Abstract

RNA sequencing data have been abundantly generated in biomedical research for biomarker discovery and other studies. 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.

Keywords

Correlated data, differential expression analysis, quantile regression, RNA sequencing, robust tail-based test

1 Introduction

RNA sequencing (RNA-seq), also called 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.[1] Utilizing such data, researchers are able to extract rich genomic information from biological systems and advance our knowledge about various diseases, including cancer.

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, have RNA-seq profiling data available for a large

[1]Data Science, LinkedIn, Mountain View, CA, USA
[2]Department of Biostatistics, Columbia University, New York, NY, USA
[3]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[4]Department of Statistics, University of Michigan at Ann Arbor, Ann Arbor, MI, USA

*Co-first authors.

Corresponding author:
Jianhua Hu, Department of Biostatistics, Columbia University, New York, NY 10032, USA.
Email: jh3992@cumc.columbia.edu

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.[2–9]

Several methods have been developed to detect differential gene expression in RNA-seq experiments. Jiang and Wong[10] 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.[11] and McCarthy et al.[12] 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[2] and edgeR[26] 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[13] 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[14] detected the group difference while addressing the correlation structure within each gene. However, the normality assumption is usually not satisfied, even with data transformation,[15] 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.[15] and Chu et al.[16]

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.[17] Furthermore, Wang and He[31] 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 for RNA-seq data commonly use gene-level read counts by summarizing exon-level sequenced reads to gene-level data. These methods lose potentially useful information about the exon-level expression distribution.[18] In this paper, we propose a new tail-based test at the level of exon-level expression data. In the new test, we accumulate the information on all the quantiles of a tail region and account for the inter-exon correlations. This is motivated by previous research on microarray expression data that show that statistical testing on probe-level data can improve the detection of differential gene expression over that on gene-level data.[19] The idea of using quantile aggregation was initially proposed by He et al.[20] that focused on detecting treatment effects on independent observations of a response variable in clinical studies. RNA degradation renders the read counts unevenly across the different exon regions and often causes biases towards the 3′ end.[21] Hence, we focus on the upper tails in the test since high gene expression intensities are particularly biologically 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. Figure 1 shows high correlations between exons in gene *FHIT*. Besides, Figure 2 is the histogram of median correlations for all genes, revealing high inter-exon correlations for the most genes. The proposed test is capable of adjusting for covariates and accounting for the inter-exon correlations within a gene. In this paper, the choice of quantile τ is a user-specified value (e.g. 0.5 or 0.75) as our empirical investigation shows these are effective starting points to accumulate upper quantile distribution information.

This paper is organized as follows. In Section 2, we introduce the model and notations and present the tail-based test and its limiting distribution under the null hypothesis. In Section 3, we perform Monte Carlo simulations on correlated data and make comparisons with several popular methods including edgeR, DESeq2, and Limma. In Section 4, we analyze TCGA lung adenocarcinoma data using the proposed test. We conclude with some brief remarks in Section 5.

2 Method

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

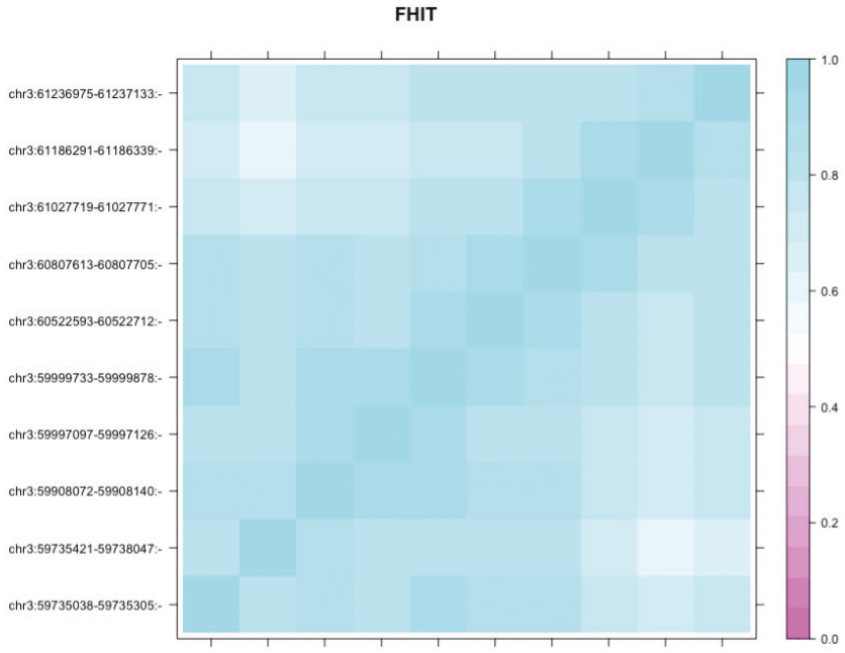

Figure 1. Heatmap of correlation on exon-level expression for gene *FHIT* from TCGA lung adenocarcinoma data.

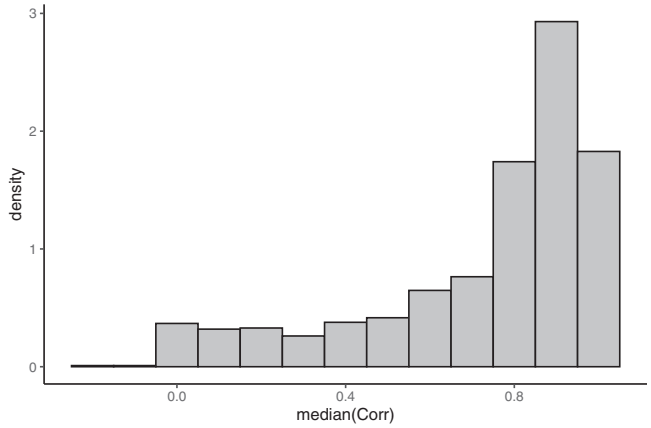

Figure 2. Histogram of median correlation among exons within each gene.

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

$$Q_{\boldsymbol{Z}}(\tau|\mathbf{D}, \mathbf{C}) = \alpha(\tau) + \boldsymbol{D}\delta(\tau) + \mathbf{C}\boldsymbol{\gamma}(\tau) = X\boldsymbol{\beta}(\tau) \tag{1}$$

where $X = (1_{n\times 1}, \boldsymbol{D}_{n\times 1}, \boldsymbol{C}_{n\times K})$ 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

$$Z_{ij} = \alpha(\tau) + D_i\delta(\tau) + \mathbf{C_i}^T\boldsymbol{\gamma}(\tau) + e_{ij}(\tau) \tag{2}$$

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 $(Z_{ij}, D_i, \mathbf{C_i})$. In this paper, we assume compound symmetry correlation structure among exons within a gene, which has been empirically shown to be sensible for RNA-seq data. We obtain the estimate $\hat{\alpha}(\tau), \hat{\delta}(\tau), \hat{\boldsymbol{\gamma}}(\tau)$ at the τth quantile via quantile regression.[22] 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

$$T_\tau^{TTS}(n_1, n_0) = TTS_\tau(1) - TTS_\tau(0) \tag{3}$$

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 $TTS_\tau(d)$ carries the information of covariate-adjusted residuals and represents the average expression intensity above the τth quantile in group d after adjusting for the covariates. As an example, $TTS_{0.5}(.)$ incorporates the information of the whole region above the 50th quantile for a group at $\tau = 0.5$. This summary statistic represents a weighted average of the upper quantile region, rather than a single quantile value that is typically used in the traditional rank score tests. As a result, the test statistic leverages the power to detect the difference between two groups in terms of the distributions above the τ-th quantile.

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

$$T_\tau^{TTS}(n_1, n_0) = \delta(\tau) - (\bar{\boldsymbol{C}}_\tau^T(1) - \bar{\boldsymbol{C}}_\tau^T(0))(\hat{\boldsymbol{\gamma}}(\tau) - \boldsymbol{\gamma}(\tau)) + (\bar{e}_\tau(1) - \bar{e}_\tau(0)) \tag{4}$$

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 $e_{ij} = 0$, 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.[20] We also denote the transformed D and C via Gram–Schmidt orthogonalization as follows

$$D_i^* = D_i - n_d^{-1}\sum_i D_i I(D_i = d) \tag{5}$$

$$\boldsymbol{C}_i^* = \boldsymbol{C}_i - n_d^{-1}\sum_i \boldsymbol{C}_i I(D_i = d) \tag{6}$$

In addition, let

$$\begin{aligned} V_d = &\sum_{D_i=d}\sum_{j=1}^{m_i} var(e_{ij}e_{ij}^+) + \sum_{D_i=d}\sum_{j\neq j'} cov(e_{ij}e_{ij}^+, e_{ij'}e_{ij'}^+) \\ &\text{and } \zeta = P(e_{ij} < 0, e_{ij'} < 0) \end{aligned} \tag{7}$$

Theorem 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

$$T_\tau^{TTS}(n_1, n_0)/s_{n_0,n_1} \to N(0,1) \text{ as } n_1, n_0 \to \infty \tag{8}$$

The proof and notation of s_{n_0,n_1} are in the Section B of the Supplement.

Remark (*a*): A consistent estimate of U_f can be obtained using the kernel density estimate of f_{ij} based on empirical residuals $\hat{e}_{ij}(\tau)$.[23,24] 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,[25] 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,

Remark (*c*): The choices of τ depend on real applications, but as a guideline, we focus on the upper tails in the test since high gene expression intensities are particularly biologically meaningful in the applications. Multiple τs in a range of $50\% - 75\%$ are desirable to be looked at to obtain the list of top promising biomarker candidates for further biological validation. The optimal choice of τ is not the goal here, but any reasonable choice tends to improve on the use of one quantile level which will be demonstrated in the next section.

$$\hat{\zeta} = \left\{ \sum_i m_i(m_i - 1)/2 - K \right\}^{-1} \sum_i \sum_{j \neq j'} \hat{e}^-_{ij} \hat{e}^-_{ij'} \tag{9}$$

$$\begin{aligned} \hat{V}_d = & \sum_{D_i=d} \sum_{j=1}^{m_i} \left(\hat{e}^2_{ij} \hat{e}^+_{ij} \right) - N_d^{-1} \left(\sum_{D_i=d} \sum_{j=1}^{m_i} \hat{e}_{ij} \hat{e}^+_{ij} \right)^2 \\ & + \sum_{D_i=d} \sum_{j \neq j'} \left[\left\{ \sum_{D_i=d} m_i(m_i - 1) \right\}^{-1} \sum_{D_i=d} \sum_{j \neq j'} \hat{e}_{ij} \hat{e}^+_{ij} \hat{e}_{ij'} \hat{e}^+_{ij'} - n_d{}^{-1} \left(\sum_{D_i=d} \sum_j \hat{e}_{ij} \hat{e}^+_{ij} \right)^2 \right] \end{aligned} \tag{10}$$

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

3 Simulation studies

3.1 Comparison with quantile rank score test and linear mixed effect model

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), and the Wald test for coefficient estimates of the linear mixed effect model (called *LME*). We generated exon-level gene expression data from the following model

$$Z_{ij} = 5 + \gamma C_i + \delta_1 I(D_i = 1) + \delta_2 I(e_{ij} > 0) I(D_i = 1) e_{ij} + e_{ij} \tag{11}$$

where Z_{ij} is the intensity value of exon j of a gene for subject sample i. 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 40, 50, 75, 100, and 150 subjects per group and gene lengths of 5, 10 and 30 exon locations within a gene, respectively. In each scenario, we ran 5000 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 1. We observe that *QRS* fails to maintain appropriate type I error rates due to high correlation among the exons. In contrast, *TTS*, QRS_c, and *LME* are able to preserve the type I error rates in various cases.

Table 1. Type I error rates at the nominal levels of 1% and 5% for Scenarios 1 and 2.

Scenario 1, 2	Nominal level	1%				5%			
Gene Length	Sample Size	*TTS*	QRS_c	*QRS*	*LME*	*TTS*	QRS_c	*QRS*	*LME*
	40	1.26	1.10	16.84	0.80	5.66	4.58	29.70	4.94
5	50	1.26	0.96	16.46	0.78	5.60	5.46	30.78	4.94
	75	0.98	0.70	16.12	0.96	4.86	4.74	28.26	4.74
	100	0.86	0.96	16.22	0.74	5.16	5.18	29.00	4.90
	150	0.98	1.06	16.30	1.00	5.10	4.90	29.32	4.88
	40	1.34	1.06	30.20	1.00	5.50	5.38	43.98	5.02
10	50	1.44	1.08	30.44	1.02	5.98	4.94	43.34	4.44
	75	1.22	0.98	30.56	0.78	5.72	5.04	44.40	5.12
	100	1.10	1.12	29.98	1.04	5.18	5.20	43.16	4.76
	150	1.06	1.04	30.26	1.02	5.16	5.06	43.80	5.34
	40	1.44	0.90	55.56	0.98	5.96	5.04	65.58	4.68
30	50	1.52	1.20	53.96	1.16	6.14	5.34	64.04	4.68
	75	1.36	1.24	54.96	1.16	5.62	5.20	64.86	5.02
	100	1.32	1.26	53.72	1.10	5.64	5.26	63.92	5.38
	150	1.30	1.06	55.54	1.06	4.98	4.78	65.14	4.78

Note: Scenarios 1 and 2 have identical type-I error rates. The values in the table are percentages.

The power results are shown for *TTS*, QRS_c, and *LME* in the top panel of Table 2. We did not investigate *QRS* further due to its statistical 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 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. 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 Table 3. 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 $\tau = 0.75$. It is also noteworthy that TTS reaches close to 100% power with larger sample sizes.

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 Tables 4 and 5. The type I error rates of the proposed test, *TTS*, as well as QRS_c and *LME*, are well maintained at the corresponding nominal level in the various setups. The power results displayed in Tables 6 and 7 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.

3.2 Comparison with 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 intensity data in Log2-RPKM format from the following model to fit our model, and converted the measurement to the initial gene-level counts to implement other DE analysis methods.

$$Z_{ij} = \alpha + \gamma C_i + \delta I(e_{ij} > 0) I(D_i = 1) e_{ij} + e_{ij} \tag{12}$$

We investigated the following two scenarios.

Table 2. Power for scenarios 1 at quantiles $\tau = 0.5$ and 0.75 at the significance level of 0.05.

Scenario 1		$\tau = 0.5$			$\tau = 0.75$		
Gene Length	Sample Size	*TTS*	QRS_c	*LME*	*TTS*	QRS_c	*LME*
	40	58.04	59.28	67.46	47.42	52.42	67.46
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
	150	98.54	98.76	99.62	93.44	97.36	99.62
	40	60.28	61.74	68.22	49.38	55.38	68.22
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
	150	98.82	99.06	99.86	94.70	98.34	99.86
	40	61.80	63.68	69.88	51.42	57.36	69.88
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
	150	99.10	99.42	99.74	95.36	98.64	99.74

Note: The values in the table are percentages.

Table 3. Power for scenarios 2 at quantiles $\tau = 0.5$ and 0.75 at the significance level of 0.05.

Scenario 2		$\tau = 0.5$			$\tau = 0.75$		
Gene Length	Sample Size	*TTS*	QRS_c	*LME*	*TTS*	QRS_c	*LME*
	40	76.18	5.98	44.82	95.10	59.14	44.82
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
	150	99.96	5.86	96.10	100.00	98.84	96.10
	40	77.60	6.68	44.50	95.52	62.02	44.50
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
	150	99.98	6.64	96.14	100.00	98.94	96.14
	40	78.24	6.26	45.28	96.26	64.30	45.28
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
	150	99.98	6.12	96.60	100.00	99.18	96.60

Note: The values in the table are percentages.

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 the unit variance and an exchangeable correlation structure with $cor(e_{ij}, e_{ij'}) = 0.8$ and $cor(e_{ij}, e_{i'j'}) = 0$. To study the impact of sample size and gene length (the number of exon locations) on the test, we considered the sample sizes of 40, 60, 80, and 100 subjects per group and gene lengths of 5, 10, and 30, respectively. In each scenario, we ran 5000 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 4. Type I error rates at the nominal levels of 1% and 5% for Scenarios 3.

Scenario 3	Nominal Level	1%				5%			
Gene Length	Sample Size	*TTS*	QRS_c	*QRS*	*LME*	*TTS*	QRS_c	*QRS*	*LME*
	40	1.26	1.00	16.54	0.82	5.66	4.76	29.84	4.88
5	50	1.48	1.02	16.18	0.78	5.82	5.24	30.66	4.80
	75	1.06	0.86	15.76	0.90	5.06	5.12	28.78	4.84
	100	0.94	1.00	17.32	0.84	5.24	5.24	29.00	4.60
	150	0.92	1.06	16.10	0.98	5.26	4.84	29.36	5.02
	40	1.34	1.00	29.40	0.98	5.66	5.42	43.78	4.88
10	50	1.42	1.10	30.34	0.98	5.98	5.34	43.62	4.76
	75	1.24	0.92	31.36	0.70	5.68	5.30	44.46	5.14
	100	1.02	1.06	30.18	1.02	5.22	4.86	44.20	4.78
	150	1.06	1.04	30.20	0.94	5.10	5.10	43.56	5.36
	40	1.40	0.92	55.86	1.00	6.04	5.18	65.60	4.66
30	50	1.56	1.26	54.46	1.20	6.14	5.48	65.06	4.68
	75	1.32	1.32	55.04	1.10	5.58	5.56	64.62	4.98
	100	1.36	1.46	53.80	1.32	5.88	5.26	63.54	5.00
	150	1.24	1.00	55.38	0.94	5.12	4.92	65.00	4.88

Note: The values in the table are percentages.

Table 5. Type I error rates at the nominal levels of 1% and 5% for Scenarios 4.

Scenario 4	Nominal Level	1%				5%			
Gene Length	Sample Size	*TTS*	QRS_c	*QRS*	*LME*	*TTS*	QRS_c	*QRS*	*LME*
	40	1.48	0.96	16.02	0.94	5.92	5.18	29.44	5.64
5	50	1.36	0.84	16.40	0.74	5.94	5.44	29.16	4.94
	75	1.10	0.88	16.42	0.88	4.84	5.20	29.22	4.56
	100	0.84	0.76	16.46	0.96	5.16	5.10	29.30	4.64
	150	1.26	1.18	15.72	1.28	5.18	4.94	28.34	5.06
	40	1.40	1.34	30.30	1.14	5.42	5.02	43.92	4.96
10	50	1.58	1.12	30.50	1.06	6.18	5.40	42.84	5.02
	75	1.08	1.04	31.00	1.04	5.64	5.40	43.10	5.18
	100	1.06	0.98	30.18	1.04	5.28	5.28	44.56	5.34
	150	0.94	1.04	30.30	1.16	5.56	5.24	43.30	5.16
	40	1.46	1.08	55.00	0.90	5.92	5.06	64.96	4.30
30	50	1.52	1.38	55.30	1.14	5.88	5.32	64.66	5.00
	75	1.30	0.86	54.44	0.98	5.12	5.56	64.66	4.82
	100	1.28	1.18	53.42	1.00	5.36	5.08	63.48	4.80
	150	1.04	0.98	55.32	1.06	5.08	4.86	65.04	5.02

Note: The values in the table are percentages.

We calculated the average false positive rates (FPRs) and true positive rates (TPRs) to assess the performance of the aforementioned four methods.

Scenario DE-1. The FPRs are shown in Table 8. We observe that *edgeR* and *DESeq*2 are sensitive to noises and exhibit inflated FPRs. In contrast, *TTS* and *Limma* can appropriately maintain the FPRs close to the nominal value.

Scenario DE-2. In this scenario, the cancer group has a heavier right tail and larger variance than that of the normal group ($D_i = 0$) for DE genes. The difference between the two groups is relatively small at the median and enlarges in the upper quantiles as shown in Figure 3. The result in Table 9 indicates that *edgeR* and *DESeq*2 are again noise-susceptible and result in inflated FPRs, while *TTS* is capable of preserving FPRs at the appropriate levels. In terms of TPRs, *TTS* has the similar performance compared to both *edgeR* and *DESeq*2, while *Limma* is inferior to the others. Overall, *TTS* performs better than *edgeR*, *DESeq*2, and *Limma*.

Table 6. Power for scenarios 3 at quantiles $\tau = 0.5$ and 0.75 at the significance level of 0.05.

Scenario 3		$\tau = 0.5$			$\tau = 0.75$		
Gene Length	Sample Size	*TTS*	QRS_c	*LME*	*TTS*	QRS_c	*LME*
	40	76.74	5.92	44.58	95.84	60.08	44.58
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
	150	99.96	5.82	96.14	100.00	99.00	96.14
	40	77.94	6.64	44.32	96.00	63.04	44.32
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
	150	99.98	6.42	96.12	100.00	99.08	96.12
	40	78.70	6.22	45.00	96.48	64.90	45.00
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
	150	99.98	5.90	96.56	100.00	99.08	96.56

Note: The values in the table are percentages.

Table 7. Power for scenarios 4 at quantiles $\tau = 0.5$ and 0.75 at the significance level of 0.05.

Scenario 4		$\tau = 0.5$			$\tau = 0.75$		
Gene Length	Sample Size	*TTS*	QRS_c	*LME*	*TTS*	QRS_c	*LME*
	40	68.80	5.84	34.66	90.66	48.12	34.66
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
	150	99.88	5.88	90.06	100.00	96.48	90.06
	40	70.30	6.18	34.00	91.36	50.18	34.00
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
	150	99.92	6.52	90.48	100.00	97.32	90.48
	40	71.34	5.84	33.84	92.78	52.84	33.84
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
	150	99.94	5.82	91.30	100.00	97.78	91.30

Note: The values in the table are percentages.

Remark: In scenarios DE-1 and DE-2, *TTS* is able to control FPRs appropriately, while *edgeR* and *DESeq*2 have inflated results. *TTS* is also more powerful than *Limma*.

3.3 Comparison with edgeR, DESeq2, and Limma part 2

In the third set of simulations, we generated the initial gene-level count data, on which we fitted edgeR, DESeq2, and Limma+voom. We also converted the gene-level counts to exon-level Log2-RPKM measurements to implement our methods. The complete analysis results are reported in the section A of the Supplement. The purpose of this investigation is to demonstrate that the proposed test is robust and comparable with *edgeR*, *DESeq*2, and *Limma*, even when the data are not generated from our assumed model.

Table 8. FPRs at the nominal levels of 1% and 5% for scenario *DE-1*.

Scenario *DE-1*	Nominal Level	1%				5%			
Gene Length	Sample 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
	100	0.98	2.04	2.02	0.92	4.96	7.64	7.80	4.94
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
	100	1.28	2.42	2.12	1.22	5.54	7.98	9.10	5.42
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
	100	1.52	2.36	2.70	1.22	6.28	8.72	9.70	5.32

Note: The values in the table are percentages.

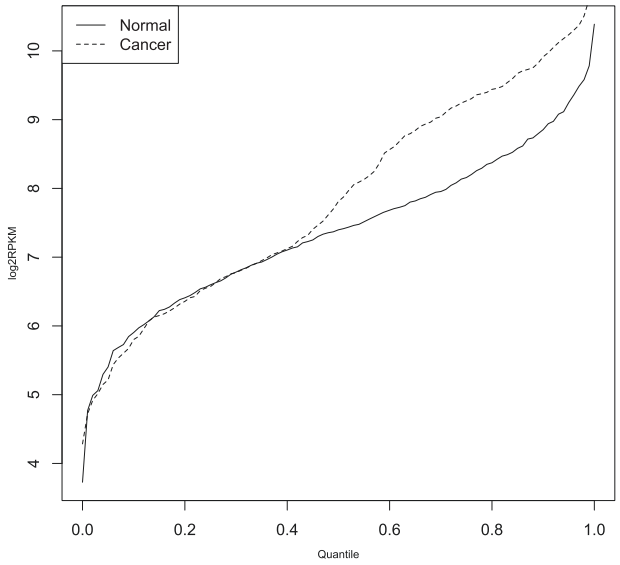

Figure 3. Quantile intensity plots of normal tissue and cancer samples for scenario DE-2.

Table 9. FPRs and TPRs at the nominal level of 5% for scenarios *DE-2*.

Scenario *DE-2*		FPR				TPR			
Gene Length	Sample Size	*TTS*	*edgeR*	*DESeq2*	*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
	100	5.67	10.69	10.56	8.84	100.00	100.00	100.00	98.40
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
	100	5.29	10.40	11.68	8.84	100.00	100.00	99.80	98.80
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
	100	5.22	10.18	11.92	7.69	100.00	100.00	100.00	99.00

Note: The values in the table are percentages.

Remark: In scenarios DE-3 and DE-4, *TTS* again correctly controls FPRs and achieves the similar power as above-described state-of-the-art DE methods.

4 A lung cancer study

We analyzed the lung adenocarcinoma (LUAD) data accessible at the TCGA public data portal, with the RNA-seq data profiled from 50 cancer and 50 normal tissue samples from cancer patients at the exon level. The gene expression data were normalized into Log2-RPKM following standard protocols, then the non-expressed genes in both groups were eliminated[26] 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 LUAD.[27] We applied the proposed test, *TTS*, the quantile rank score test, QRS_c of literature[31] 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.[28] We also applied standard gene-level differential expression analysis methods including likelihood ratio test from edgeR,[29] Wald test from DESeq2,[30] and the ordinary linear model associated t-test from Limma+voom.[13]

We included gender and smoking status, defined as current smoker, reformed smoker, and nonsmoker, as the covariates in the analysis. *TTS* detected 535 and 526 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 4 show the number of the overlapping genes among the three tests. We observed that 76% and 85% of the genes detected by *TTS* were also detected by QRS_c at and 0.75, respectively. Moreover, 84% and 78% 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 4 show the number of the overlapping genes among the four tests. We observed that 93% and 88% of the genes detected by *TTS* were also detected by *Limma* at $\tau = 0.5$ and 0.75, respectively; 94% and 93% of the genes

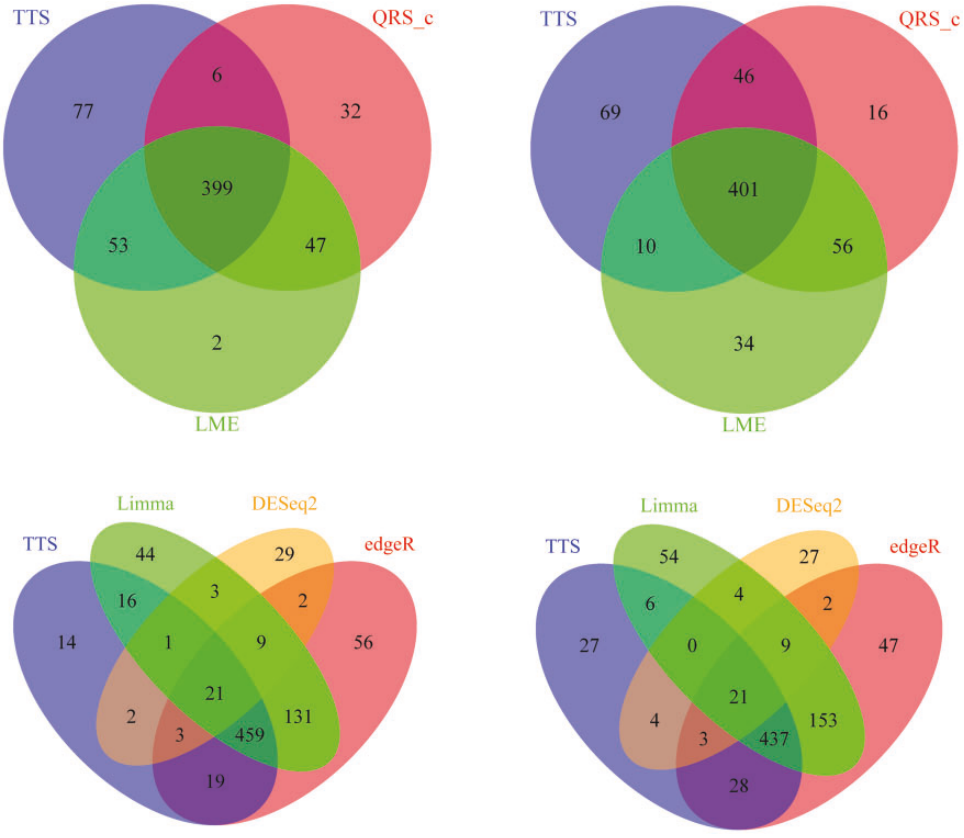

Figure 4. Venn diagram of number of overlapping genes among *TTS*, QRS_c, *LME* at top and *TTS*, *edgeR*, *DESeq2*, *Limma* at bottom, for $\tau = 0.5$ at left and 0.75 at right.

Table 10. *P*-values of the six genes are reported.

Gene	*TTS*	QRS_c	*LME*	*Limma*	*edgeR*	*DESeq2*
FHIT	**3.17e-03**	3.81e-01	5.11e-02	1.16e-01	**3.17e-03**	3.18e-01
RASSF1	**4.23e-22**	7.61e-01	**3.39e-06**	**9.10e-15**	**8.38e-15**	8.40e-01
TUSC2	4.09e-01	3.38e-01	9.19e-01	9.93e-01	5.63e-01	**4.33e-02**
SEMA3B	**2.40e-13**	**2.60e-14**	**3.10e-18**	**1.05e-17**	**4.50e-09**	9.97e-01
SEMA3F	1.21e-01	**2.00e-02**	**3.23e-02**	5.82e-02	1.57e-01	7.85e-01
MLH1	9.90e-01	5.33e-01	2.92e-01	9.31e-01	4.75e-01	7.55e-01

Note: The detected genes with false discovery rates ≤ 0.05 are highlighted in bold.

selected by *TTS* were also detected by *edgeR* at $\tau = 0.5$ and 0.75 respectively; and 5% 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*.[27] For example, *FHIT* is an identified tumor-suppressor gene that has abnormal expression in lung cancer. In Table 10, 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 of 0.03. In contrast, QRS_c detected only *SEMA3B* and *SEMA3F*, where *SEMA3F* was discovered with a FDR of 0.02. *DESeq2* detected only *TUSC2* with the FDR of 0.04.

To understand the discrepancy in the results between the methods, we first compared the results obtained from our methods with those from those 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 5.

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 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, 31 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 Table 3 in the section C of the Supplement. 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 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 16% and 14% of genes detected by QRS_c at $\tau = 0.5$ and 0.75, respectively. For example, with the FDR of 9.03×10^{-7}, QRS_c identified *FAM3D* as being associated with lung adenocarcinoma. In Figure 5, we plot the group difference at various quantiles for *FAM3D*. 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.

Seventeen 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 Table 4 in the section C of the Supplement. 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.

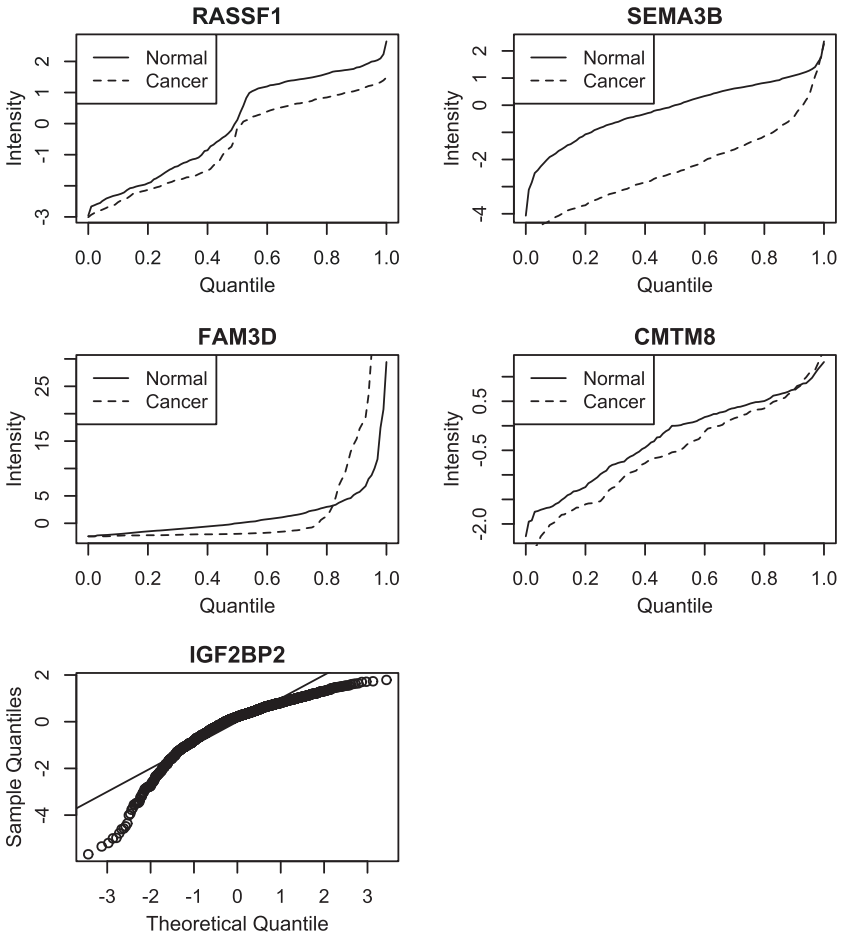

Figure 5. Top two rows are exon-level covariate-adjusted quantile intensity plots of normal tissue and cancer samples for genes *RASSF*1, *SEMA*3*B*, *FAM*3*D*, and *CMTM*8; bottom row is QQ-plot of the standardized residuals obtained from linear mixed model for gene *IGF*2*BP*2.

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 *IGF*2*BP*2 in the bottom row of Figure 5. 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 10% and 18% of genes detected by *LME*. For example, with the respective FDR of 0.0076, *LME* identified *CMTM*8 as being associated with lung adenocarcinoma. In the second row of Figure 5, we plot the group differences at various quantiles for *CMTM*8. 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.

The next, we compared the results of our method with those of the popular DE analysis methods, including *Limma*, *edgeR*, and *DESeq*2. Likely associated with lung cancer according to the medical literature are 11 genes detected by *TTS* at $\tau = 0.5$ but not by *Limma*, 7 genes detected by *TTS* at $\tau = 0.5$ but not by *edgeR*, and 138 genes detected by *TTS* at $\tau = 0.5$ but not by *DESeq*2. The complete list of genes and their literature citations are presented in Tables 5 to 9 in the section C of the Supplement. For example, *GSK*3*B* is involved in the histogenesis of lung carcinomas, and its overexpression indicates worse prognosis in lung carcinoma. *SETD*2 is a potential tumor suppressor in lung adenocarcinoma and its inactivation has led to accelerated tumor progression. *TRIM*59 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 gene *GSK*3*B* in Figure 6. *GSK*3*B* was detected by *TTS* but missed by *edgeR*, *Limma*, and *DESeq*2. *GSK*3*B* show trivial differences between the normal tissue and cancer samples below the median, which also show little

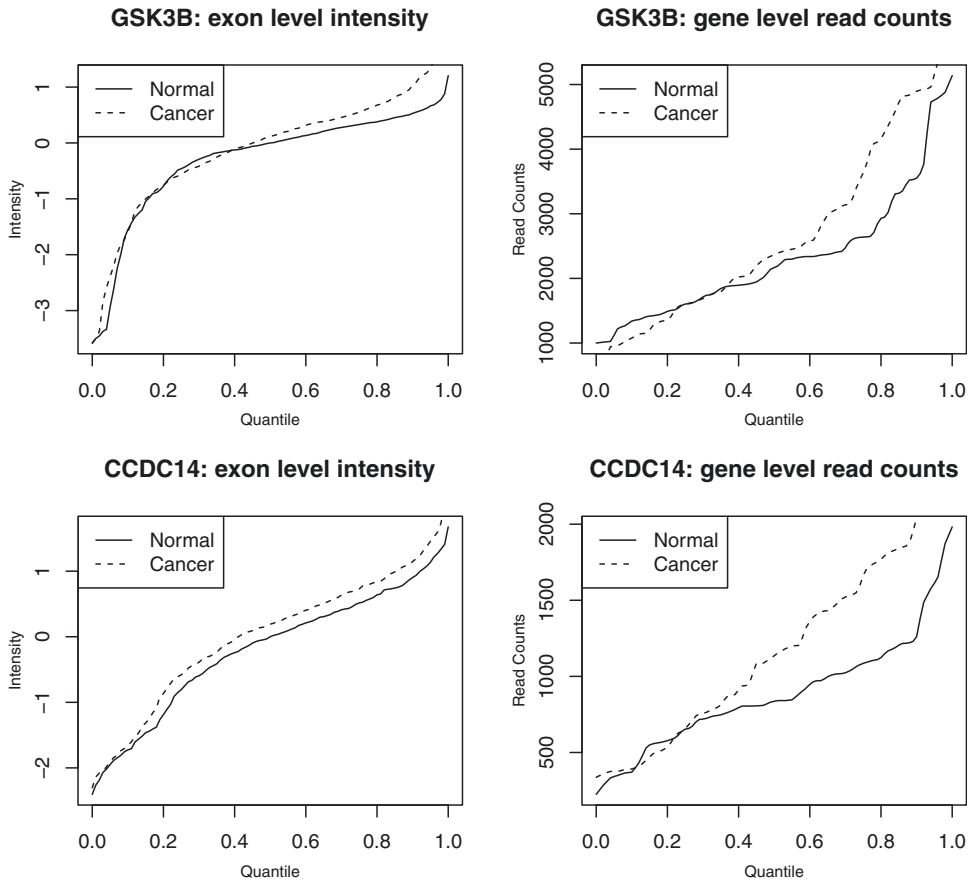

Figure 6. Left column: exon-level covariate-adjusted quantile intensity plots of normal tissue and cancer samples for genes *GSK*3*B*, and *CCDC*14; right column: gene-level read count quantile plot for the corresponding genes.

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 27% and 32% 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 28% and 30% of genes detected by *edgeR* at $\tau = 0.5$ and 0.75, respectively. Genes that were detected by *DESeq*2 but not by *TTS* account for 61% and 60% of genes detected by *DESeq*2 at $\tau = 0.5$ and 0.75, respectively. For example, *Limma*, *edgeR*, and *DESeq*2 identified *CCDC*14 with the respective FDRs of 0.018, 0.033, and 0.008. In Figure 6, we plot the group difference at various quantiles for *CCDC*14 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 the 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 several standard DE methods, as it is able to identify 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 its 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.

5 Conclusion

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. We believe the proposed method can be useful in other applications when we are interested to compare the upper or lower quantile region difference between the two groups.

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 can also perform simultaneous tests of multiple genes by utilizing information across genes in biological pathways and networks 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.

It is worth noting that the proposed method requires independence between D and C, which should be true for randomized trials. Otherwise, the independence needs to be assumed and can possibly be achieved upon removal of the dependence of C on D by a projection. Although the paper focuses on binary D, the central idea of the proposed method applies to other types of D, including a continuous variable.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was partially supported by Grants NIH/NCI 5 P30 CA013696-43, R01AI143886, R01CA219896, and National Science Foundation Awards DMS-1914496 and DMS-1951980.

ORCID iDs

Jiong Chen https://orcid.org/0000-0001-5971-1681
Jing Ning https://orcid.org/0000-0002-5289-331X

Supplemental Material

Supplemental material for this article is available online.

References

1. Auer PL and Doerge RW. Statistical design and analysis of RNA sequencing data. *Genetics* 2010; **185**: 405–416.
2. Subramanian J and Govindan R. Lung cancer in never smokers: a review. *J Clin Oncol* 2007; **25**: 561–570.
3. He XM and Shao QM. A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *Ann Stat* 1996; **24**: 2608–2630.
4. He XM, Zhu ZY and Fung WK. Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika* 2002; **89**: 579–590.
5. Hsu YH. *Applications of quantile regression to estimation and detection of some tail characteristics*. PhD Dissertation, University of Illinois at Urbana-Champaign, IL, USA, 2010.
6. Lader A, Ramoni M, Zetter B, et al. Identification of a transcriptional profile associated with in vitro invasion in non-small cell lung cancer cell lines. *Cancer Biol Ther* 2004; **3**: 624–631.
7. Lin W and Sun F. CEDER: accurate detection of differentially expressed genes by combining significance of exons using RNA-Seq. *IEEE/ACM Trans Comput Biol Bioinform* 2012; **9**: 1281–1292.
8. Owen AB. *Empirical likelihood*. London: Chapman & Hall/CRC, 2001.
9. Wang H, He X. An enhanced quantile approach for assessing differential gene expressions. *Biometrics* 2008; **64**: 449–457.
10. Jiang H and Wong WH. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* 2009; **25**: 1026–1032.
11. Bloom JS, Khan Z, Kruglyak L, et al. Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genom* 2009; **10**: 221–231.

12. McCarthy DJ, Chen Y and Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* 2012; **40**: 4288–4297.
13. Ritchie ME, Phipson B, Wu D, et al Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015; **43**: 7.
14. Laird NM and Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; **38**: 963–974.
15. Bullard JH, Purdom E, Hansen KD, et al. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinform* 2010; **11**: 94–107.
16. Chu C, Fang Hua, Yang Y, et al. deGPS is a powerful tool for detecting differential expression in RNA-sequencing studies. *BMC Genomic* 2015; **16**: 1.
17. Gutenbrunner C, Jureckova J, Koenker R, et al. Tests of linear hypotheses based on regression rank scores. *J Nonparametric Stat* 1993; **2**: 307–331.
18. Laiho A and Elo LL. A note on an exon-based strategy to identify differentially expressed genes in RNA-Seq experiments. *PLoS One* 2014; **9**: e115964.
19. Liu X, Milo M, Lawrence ND, et al. Probe-level measurement error improves accuracy in detecting differential gene expression. *Bioinformatics* 2006; **22**: 2107–2113.
20. He XM, Hsu YH and HU MX. Detection of treatment effects by covariate-adjusted expected shortfall. *Ann Appl Stat* 2010; **4**: 2114–2125.
21. Shanker S, Paulson A, Edenberg HJ, et al. Evaluation of commercially available RNA amplification kits for RNA sequencing using very low input amounts of total RNA. *J Biomol Tech* 2015; **26**: 4–18.
22. Koenker R and Bassett G. Regression quantiles. *Econometrica* 1978; **46**: 33–50.
23. Hardcastle TJ and Kelly KA. BaySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinform* 2010; **11**: 422–436.
24. Koenker R. *Quantile regression*. Cambridge: Cambridge University Press, 2005.
25. Silverman BW. *Density estimation for statistics and data analysis*. London: Chapman and Hall, 1986.
26. Mortazavi A, Williams BA, McCue K, et al. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Method* 2008; **5**: 621–628.
27. Marileila VG. Chromosomal and genomic changes in lung cancer. *Cell Adhes Migrat* 2010; **4**: 100–106.
28. Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 1995; **57**: 289–300.
29. Robinson MD, McCarthy DJ and Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010; **26**: 1.
30. Love MI, Huber W and Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014; **15**: 550.