Quantitative Biology 2020, 8(2): 155–171
https://doi.org/10.1007/s40484-020-0208-3

RESEARCH ARTICLE

A censored-Poisson model based approach to the analysis of RNA-seq data

Xing Chen, Yinglei Lai*

Department of Statistics, The George Washington University, Washington, DC 20052, USA
* Correspondence: ylai@gwu.edu

Received January 6, 2020; Revised March 31, 2020; Accepted May 4, 2020

Background: **With the recent advance of sequencing technology, the collection of RNA expression (RNA-seq) data has been growing rapidly. RNA-seq data are statistically count-type measurements. Poisson distribution is a basic probability distribution for modeling count-type data. With Poisson regression models, various experimental factors, GC content as well as alternative splicing isoforms can be flexibly considered in RNA-seq data analysis. Due to the biochemical and technical limitations of sequencing technology, the biases among RNA-seq data have been recognized.**

Methods: **In this study, an artificial censoring approach has been proposed to an isoform-specific Poisson regression model for analyzing RNA-seq data. Low expression values can be grouped (censored) into one probability category, and high expression values can also be grouped (censored) into another probability category. We have implemented the related Newton-Raphson numeric computing procedure to achieve the maximum likelihood estimation for our censored-Poisson regression model. The related mathematical simplifications have been derived for the consideration of stable and convenient numerical computing.**

Results: **The advantages of our artificial censoring approach have been demonstrated in both simulation studies and application analysis of experimental data.**

Conclusions: **Our proposed artificial censoring approach allows us to focus on the majority of data. As the extreme values (tails) of data are artificially censored, more efficient analysis results can be obtained, even from relatively simple Poisson regression models. Our proposed artificial censoring approach can certainly be considered for other well-developed models or methods for RNA-seq data analysis.**

Keywords: RNA-seq; Poisson models; censored distribution

Author summary: RNA sequencing (RNA-seq) expression data have been increasingly collected for various biomedical studies. Due to the biochemical and technical limitations of sequencing technology, the biases among RNA-seq data have been recognized. We have developed an artificial censoring approach to the analysis of isoform-specific RNA-seq expression data. Low and high expression values can be grouped (censored) into the related probability categories. This approach allows us to focus on the majority of data and to obtain more efficient analysis results. Our proposed artificial censoring approach can also be considered in other RNA-seq data analysis scenarios.

INTRODUCTION

RNA sequencing (RNA-seq) data are essential for us to gain further insights into the molecular functions and regulations related to biomedical studies. High-throughput RNA-seq data have being increasingly collected in biomedical studies. Statistically, RNA-seq data are count-type measurements. Due to the complicated RNA-seq experimental procedure, many factors must be considered in the related data analysis. This is usually achieved by a statistical regression approach. To build an appropriate regression model, it is important to understand the experimental sequencing process for obtaining RNA-seq data.

In this study, we focus on mRNA sequencing data analysis. Before the analysis for a RNA-seq data set, the data preprocessing must be conducted. The following is a brief summary. There are currently two types of short reads from a RNA-seq experiment: single-end and paired-end. After recording short reads from a RNA-seq experiment, it is necessary to perform a preprocessing procedure so that numerical data can be available for a follow-up analysis. The protocol proposed by Trapnell *et al*. [1] is a widely used data preprocessing method. Then, RNA-seq data are made available as count-type measurements for mRNA exons. Other RNA-seq data preprocessing methods have also been made public available [2,3]. Additionally, RNA-seq data normalization/quantification is also important in a genome-wide mRNA expression study, and the reads per kilo-base exons per million reads (RPKM) [4] and RSEM [5,6] are two representative normalization/quantification methods. This is because it is still difficult to obtain direct mRNA expression measurements due to the current technology limitations.

For RNA-seq data analysis, Jiang and Wong [7] were among the earliest to propose a Poisson distribution based statistical method for this purpose. Further Poisson distribution based statistical methods were also developed for analyzing RNA-seq data [8,9]. Poisson distribution is one of the most widely used probability distribution for modeling count-type measurements. Many related mathematical theories and computing implementations have been developed. Alternative splicing is a fundamental molecular process, which makes different versions of transcripts (isoforms) available from a single gene. With exon usage information, it is feasible to perform RNA-seq data analysis with the consideration of mRNA isoforms. GC content is the percentage of nucleobases G and C from a fragment of RNA/DNA sequence (*e.g*., an exon). Its impact on RNA-seq data has been widely studied [10–13]. Poisson distribution based regression models have been widely developed to incorporate different molecular information (*e.g*., isoform-specific exon usage, GC content) into a RNA-seq data analysis. Other related methods, such as Poisson mixture models and negative binomial distribution based regression models, have also been widely used in practice [1,14–20].

Censoring is statistically a situation that the exact value of an observation is not available but a related interval can be specified. Due to the biochemical and technical limitations of sequencing technology, the biases among RNA-seq data have been recognized [11,12,21]. Additionally, we have the following motivation. In a common situation of RNA-seq data analysis, the majority of data in general could be well modeled by a probability distribution, but it was usually difficult to model the extreme values (tails) of data. Notice that, in many analysis situations, the impact on model performance from extreme values could be significant. Sometimes, just like outlier effects, such an impact would result in a clearly reduced model performance. For this important concern, we can consider these observations as censored data, which is an artificial censoring approach. We group low expression values (*e.g*., count lower than a given value as censored) into one probability category, and high expression values (*e.g.*, count higher than a given value as censored) into another probability category. After artificial censoring, the undesirable impact from extreme values could be significantly reduced. The advantage of artificial censoring is that it allows us to focus on the majority of data. (It is true that, when an interval of continuous values is considered as a category, a considerable amount of data information is lost.) As the extreme values (tails) of data are artificially censored, more efficient analysis results can be obtained, even from relatively simple Poisson regression models. (Our proposed artificial censoring approach can certainly be considered for other well-developed models, such as Poisson mixture models and negative binomial models.)

In this study, we first introduce our artificial censoring approach to a Poisson regression model designed for RNA-seq data analysis (with isoform-specific expression considered). The Newton-Raphson method is used in our numerical computing to achieve the maximum likelihood estimation. We have also derived the related mathematical simplifications for the consideration of stable and convenient numerical computing. We have conducted application analysis of experimental data as well as simulation studies to illustrate the advantages of our method. Compared to the traditional non-censoring approach, our artificial censoring approach can achieve more efficient results in RNA-seq data analysis.

RESULTS

An application to TCGA RNA-seq data: Gene *SPDYE6*

This gene is a speedy/RINGO cell cycle regulator family member (E6). It locates on chromosome seven and it has only one transcript/isoform with seven exons. GC content is an important genomic feature. Figure 1 shows the relationship between exon raw counts and GC content of gene *SPDYE6* for normal subjects and tumor subjects. A consideration of quadratic term for GC content would allow us to accommodate possible nonlinear effect to a certain extent. Therefore, we included a quadratic term in our censored-Poisson regression model. The GC content values (percentages) were [0.6286, 0.5482, 0.5233, 0.4746, 0.6147, 0.5799, 0.5232] for the seven exons, and the related exon length values were [105, 394, 86, 59,

Table 1 Quantiles of exon counts for gene *SPDYE6* from normal, tumor and pooled subjects

Sample	15%	20%	25%	30%	80%	85%	90%	95%
Normal	0	1.0	31.0	48.0	881.6	1089.0	1384.0	2022.1
Tumor	0	18.2	50.0	53.3	1600.0	2200.0	2800.9	4099.35
Pooled	0	3.0	41.5	50.0	1101.2	1501.5	2054.6	3114.1

Table 2 Estimation results for gene *SPDYE6*

Sample	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\theta} = \exp(\hat{\beta}_0)$
Normal	−83.84	225.57	−203.92	3.89e–37
Tumor	−83.62	222.58	−199.00	4.83e–37

Normal and tumor subjects were analyzed separately.

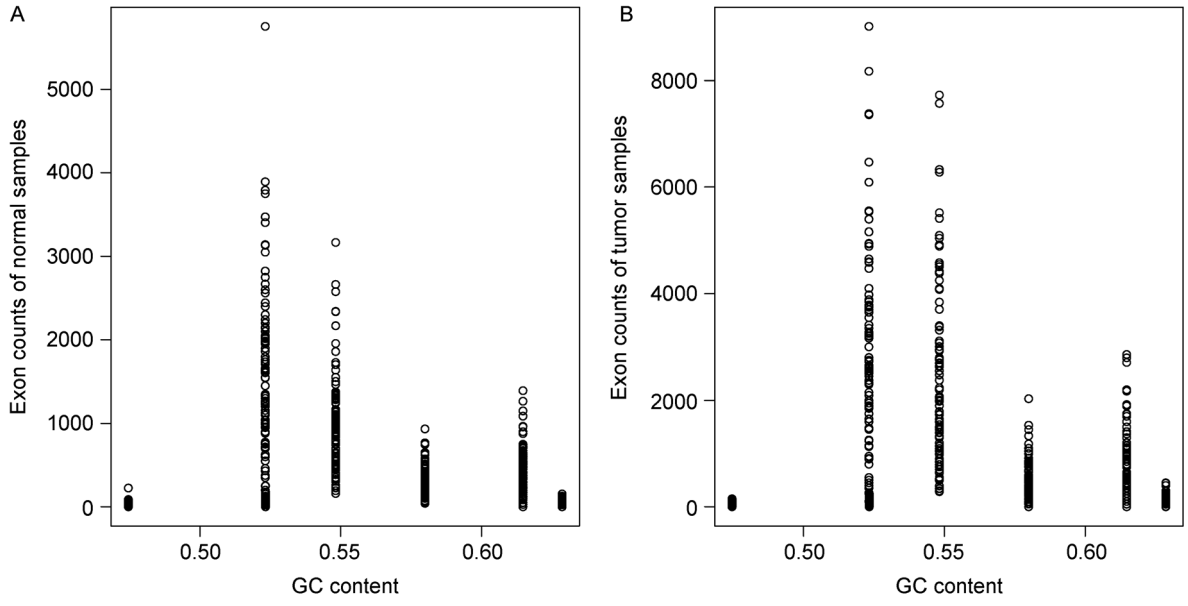

Figure 1. The relationship between exon count and GC content percentage for gene *SPDYE6*. (A) Normal subjects. (B) Tumor subjects.

322, 219, 581]. For this example, we performed our analysis separately for normal subjects vs. tumor subjects. The data were also considered as reference for a simulation study presented later for illustrating the impact of artificial censoring bounds.

Table 1 shows exon count quantiles for normal, tumor and pooled subjects. The range for pooled subjects was (0, 9014). At each quantile, the exon counts from tumor subjects were clearly larger than these from normal subjects. As we performed our analysis separately for normal vs. tumor subjects, we could consider different artificial censoring bounds. For the lower censoring bound, we set 3 for both normal and tumor subjects. For the upper censoring bounds, we set 2,000 for normal subjects and 3,000 for tumor subjects. Figure 2 provides an illustration for the data. In our censored-Poisson regression model, the coefficient β_0 was intercept. β_1 and β_2 were the linear and quadratic effects of GC content, respectively. Table 2 gives the estimation results. In a simulation study presented later, we used the same data as reference.

An application to TCGA RNA-seq data: Gene *TP53*

This gene encodes tumor suppressor protein [22]. It locates on chromosome seventeen and it has many transcripts/isoforms by different exon usages. Furthermore, new transcripts/isoforms can still be possibly discovered. Due to the limited data and computing resources at the time of analysis, we considered the following four alternative splicing isoforms (represented by their exon length matrix, or ELM) for illustrating our method. The rows and columns in an ELM represent isoforms and exons, respectively, and each entry is an

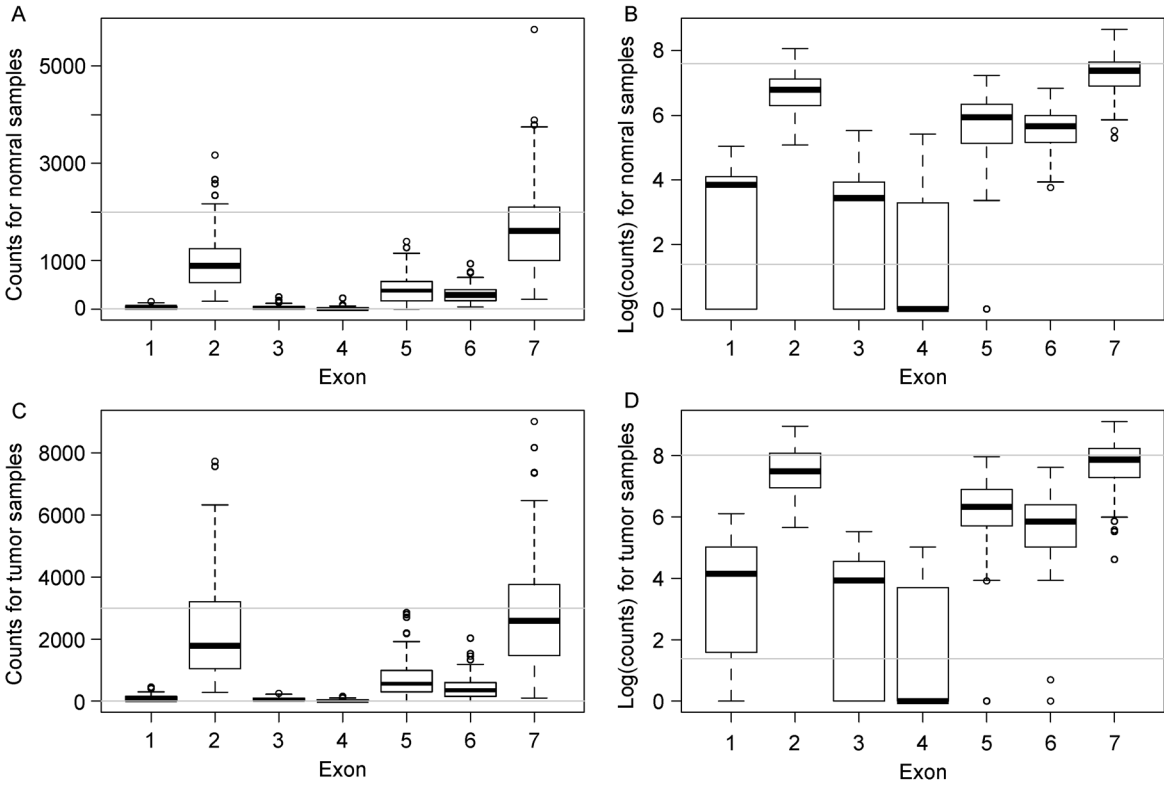

Figure 2. Boxplots of counts from different exons of gene *SPDYE6*. (A) and (B) Normal subjects; (C) and (D) Tumor subjects. Panels (A) and (C) are based on counts, and panels (B) and (D) are based on log(counts + 1). Within each graph, two grey horizontal lines represent upper and lower censoring bounds.

exon length value. (Also, notice that some transcripts/isoforms have only a few exons. They are not included in this analysis because they are usually lowly or even rarely expressed. It is numerically difficult to consider them in the current analysis. Therefore, the following four isoforms were included in this analysis.)

$$ELM = \begin{bmatrix} 236 & 0 & 0 & 0 & 0 & 0 & 0 & 110 & 429 & 441 & 279 & 241 & 103 & 0 \\ 0 & 142 & 0 & 0 & 0 & 74 & 137 & 110 & 429 & 441 & 279 & 241 & 103 & 0 \\ 0 & 0 & 1289 & 107 & 133 & 74 & 137 & 110 & 429 & 441 & 279 & 241 & 103 & 169 \\ 0 & 0 & 1289 & 107 & 0 & 74 & 137 & 110 & 429 & 441 & 279 & 241 & 103 & 169 \end{bmatrix}.$$

For this example, we performed our analysis for normal and tumor subjects together to illustrate a comprehensive analysis of our method, particularly for differential isoform-specific expression analysis. Figure 3 provides an illustration for the data. To choose the artificial censoring bounds, we pooled normal and tumor subjects and found 250 and 30,000 approximately as the 15th and 85th percentiles (set as lower and upper censoring bounds), respectively. (Notice that the counts from exons 3 and 13 were either mostly or all artificially censored.) The range of exon counts from tumor subjects was clearly wider than that from normal subjects.

Before conducting a differential isoform-specific expression analysis, we obtained the isoform-specific estimates by analyzing normal subjects and tumor subjects separately. Table 3 gives the isoform-specific estimation results and their ratios between normal vs. tumor subjects. The estimation results for isoforms 1 and 2 were similar but the estimation results for isoforms 3 and 4 were clearly different. Then, we pooled normal and tumor subjects together for a differential isoform-specific expression analysis. We performed the related likelihood ratio test (LRT) to confirm this differential expression at the (unobserved) isoform level. The LRT was calculated as the ratio between the maximum likelihood under the non-null hypothesis (differential expression) vs. the maximum likelihood under the null hypothesis (non-differential expression). Equation (2) was used for the calculation of maximum likelihood (under non-null or null hypothesis). The results in Table 3 were based on the

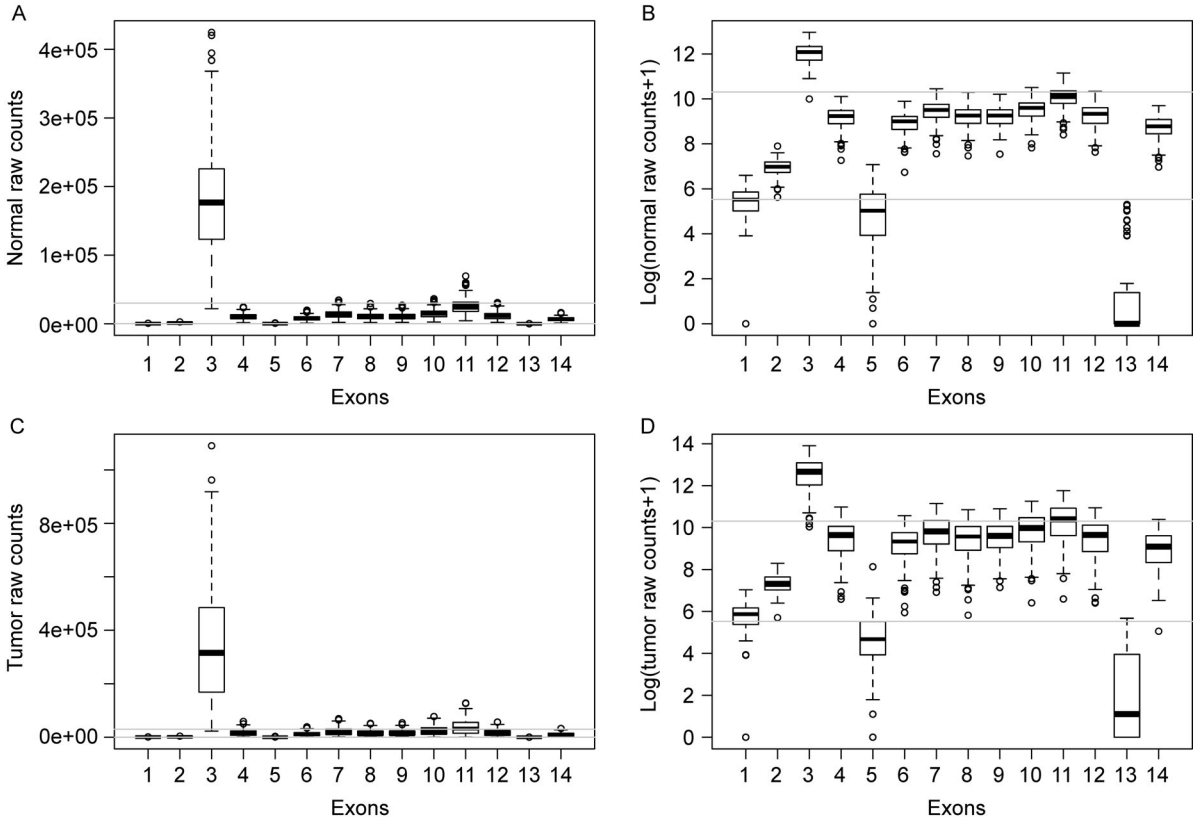

Figure 3. Boxplots of counts from different exons of gene *TP53*. (A) and (B) Normal subjects; (C) and (D) Tumor subjects. Panels (A) and (C) are based on counts, and panels (B) and (D) are based on log(counts + 1). Within each graph, two grey horizontal lines represent upper and lower censoring bounds.

Table 3 Isoform-specific estimation results for gene *TP53* and their ratios between normal vs. tumor subjects

Sample type	Isoform 1	Isoform 2	Isoform 3	Isoform 4
Normal	1.04e–10	1.10e–09	5.37e–10	1.01e–08
Tumor	1.07e–10	9.33e–10	3.31e–10	7.85e–09
Ratio	0.97	1.18	1.62	1.29

non-null hypothesis. To obtain the results for null hypothesis, we pooled the data from normal and tumor subjects and removed the group-specific coefficient in the regression model. The permutation procedure was used to evaluate the significance of LRT. For each round of permutation, we randomly reassigned subjects to normal and tumor groups, and then recalculated the LRT. After 500 rounds of permutations, we obtained an empirical distribution of permuted LRT values, which was used to compare the observed LRT value (based on original data). Figure 4A shows the histogram of 500 permuted LRT values and the vertical grey line for observed LRT value (p-value < 0.05). It clearly demonstrates the statistical significance of differential expression (at the unobserved isoform level). Additionally, we repeated this analysis but without artificial censoring (*e.g.*, 0 and ∞ for lower and upper censoring bounds, respectively). Figure 4B shows the histogram of 500 permuted LRT values and the vertical grey line for observed LRT value (p-value $>$ 0.05). It clearly suggests no differential expression (at the unobserved isoform level). This comparison illustrates the advantage of artificial censoring approach.

A simulation study

Reference data for simulations

As described in Section "An application to TCGA RNA-seq data: Gene *SPDYE6*", the experimental RNA-seq data were used as reference for our simulation study (including GC content percentages and exon length values). We conducted simulations based on the situation of only one isoform to understand the model parameter estimation performance. We also conducted simulations based on the

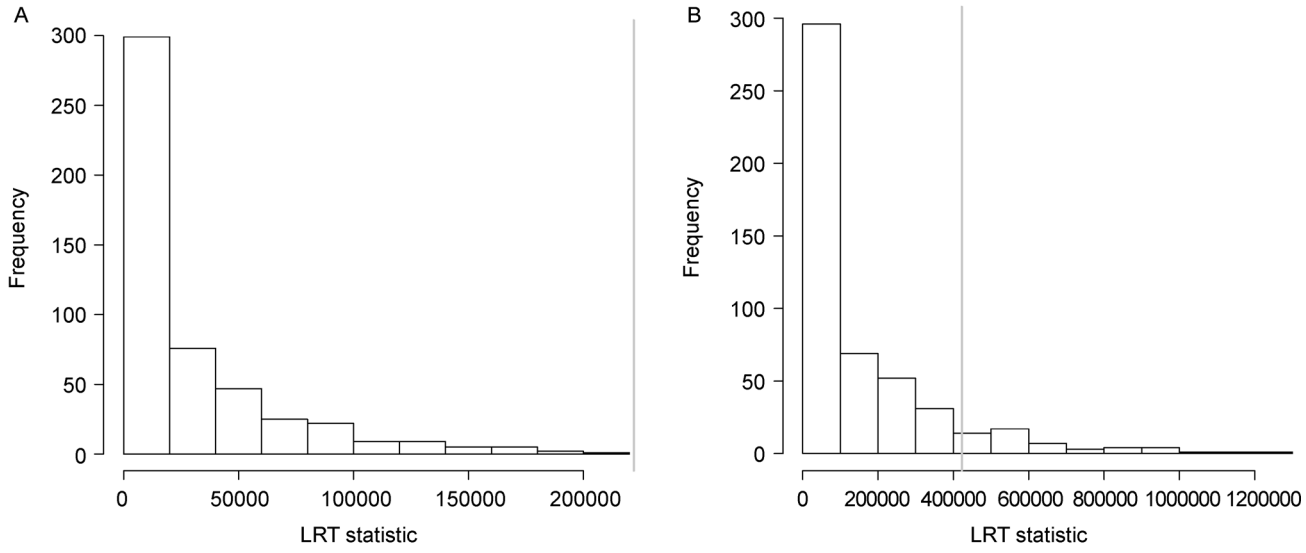

Figure 4. Likelihood ratio test (LRT) results for differential isoform-specific expression analysis. (A) Histogram of 500 permuted LRT values based on the censored-Poisson model (with vertical grey line representing observed LRT value). (B) Histogram of 500 permuted LRT values based on non-censored Poisson model (with vertical grey line representing observed LRT value).

situation of multiple isoforms to understand the isoform-specific estimation performance. Both simulation studies were based on the model specified by Eq. (1).

We compared the estimation results from the generalized linear regression (R package glm) to the estimation results from our censored-Poisson regression model but without artificial censoring (*e.g.*, 0 and ∞ for lower and upper censoring bounds, respectively). They were consistent with the same estimates: $\hat{\beta}_0 = -74$ for the intercept, $\hat{\beta}_{GC} = 199.29$ and $\hat{\beta}_{GC^2} = -183.30$ for the linear and quadratic effects of GC content. These were considered in our simulations as below.

One isoform

In addition to the above coefficient values, we included β_G as the group effect (0.1 for weak differential expression between normal and tumor subjects). Then, our coefficient parameters were $\{\beta_0,\beta_1,\beta_2,\beta_3\} = \{\beta_0,\beta_G,\beta_{GC},\beta_{GC^2}\} = (-74, 0.1, 199.29, -183.30)$.

Based on the whole data as described in Section "TCGA RNA-seq data", Fig. 5 shows the histogram of total volume (all gene/exon counts from each subject) and the fitted normal curve. For the convenience of simulations, we set a normal distribution for the total volume n_m (for each subject) with mean 4.5×10^9 and standard deviation $SD = 1.0 \times 10^9$. For the purpose of a comprehensive simulation study, our simulated data should have low, moderate, and high expressed counts all included so that both upper and lower artificial censoring could be applied. There was a lack of low expressed counts if the simulation setting based on gene *SPDYE6* was not changed. Therefore, we modified two length values for exons 1 and 4 to be 5 and 3, respectively (length vector then modified as [5, 394, 86, 3, 322, 219, 581]). The modification of these length values in our simulations was actually to make our simulated data more comprehensive (or more complicated) so that both low expression counts and high expression counts were available for our simulation analysis. Moreover, we considered a Poisson distribution for each exon length value (length vector as the Poisson distribution means).

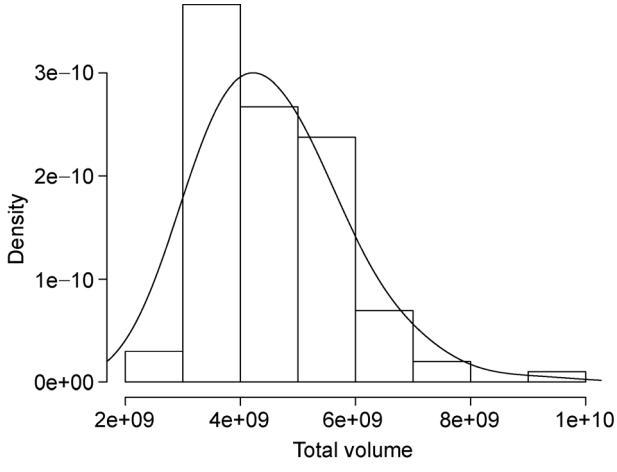

Figure 5. Histogram of total volume (all gene/exon counts from each subject) and fitted normal curve.

The GC content percentage for each exon was also randomly simulated following a uniform distribution U [0.4746, 0.6286].

After the above simulations, we included some contaminations. We added a random Poisson number with mean 250 to high expressed counts (>5000) and subtracted a random Poisson number with mean 4 to low expressed counts (<15). (Negative simulated counts were adjusted to zero.) We repeated simulations and analysis for 1,000 times. For each round, we simulated data for 100 normal subjects and 100 tumor subjects. We considered different censoring strategies: censor exactly at (15, 5000), censor more at (18, 4400), censor few at (8, 5670), as well as no censor. To compare different results, we used the absolute deviation of estimators: $|\hat{\beta}-\beta|$. Figure 6 shows the results. The absolute deviations based on "no censor" were clearly overall larger among different censoring strategies. It was not surprising that "censor exactly" was the best choice, but "censor more" was also a comparable choice. The absolute deviations based on "censor few" were overall between these based on "no censor" and "censor more" (consistently observed for different parameter estimates).

Multiple isoforms

For this scenario, we need to set values for different θ's instead of one β_0 value. Based on the modified exon length vector (5, 394, 86, 3, 322, 219, 581) from gene *SPDYE6*, we assume three artificial isoforms (just for the purpose of simulations) with the ELM as below. (Again, the modification of two exon length values in our

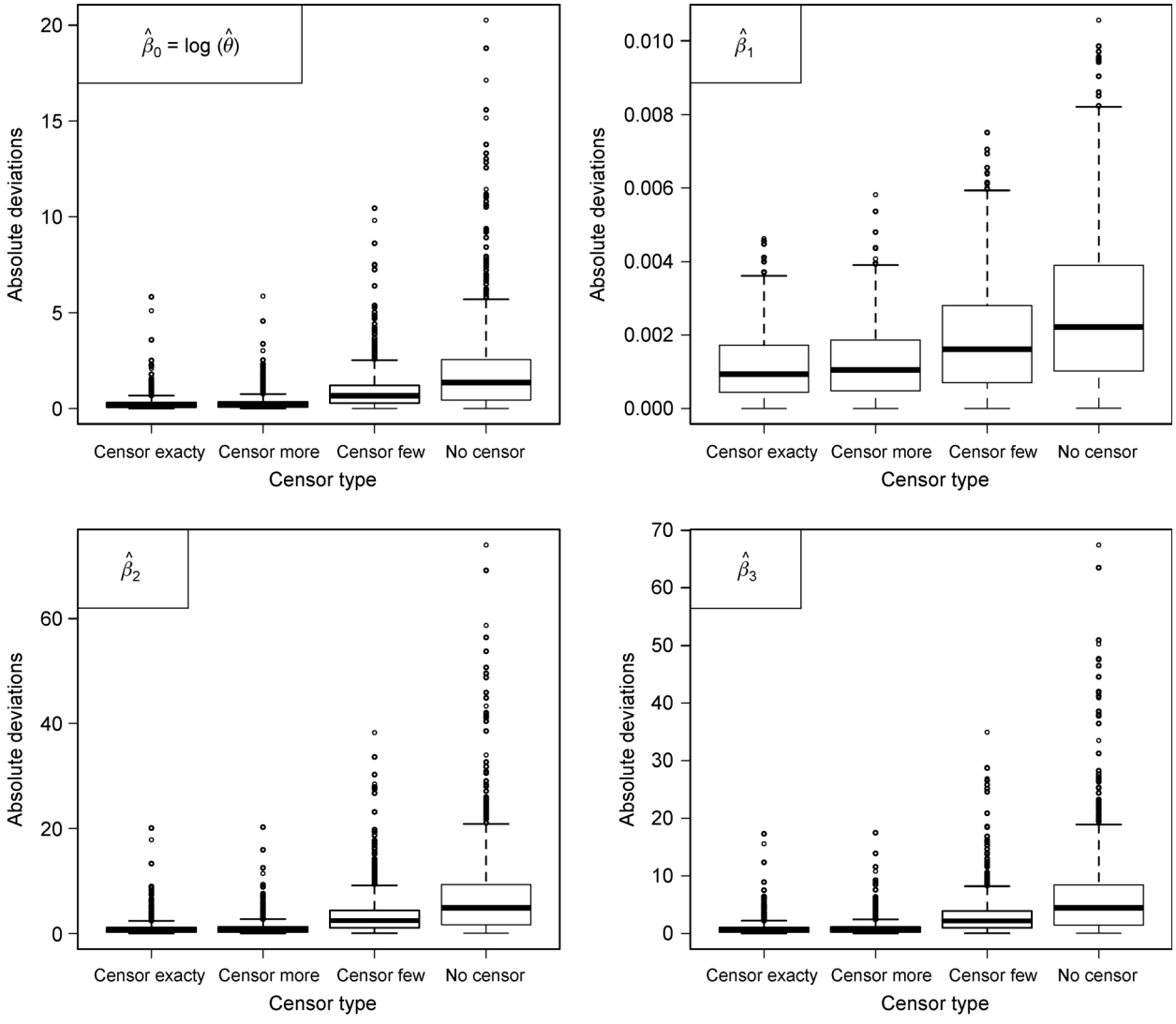

Figure 6. Boxplots of absolute deviations for estimates ($\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$). Each graph shows the results based on the censored-Poisson model with three different lower/upper bound settings as well as the non-censored Poisson model.

simulations was to make our simulated data more comprehensive/complicated so that both low expression counts and high expression counts were available for our simulation analysis.)

$$ELM = \begin{bmatrix} 5 & 0 & 86 & 3 & 322 & 219 & 581 \\ 5 & 394 & 0 & 3 & 0 & 219 & 581 \\ 0 & 394 & 86 & 3 & 322 & 219 & 0 \end{bmatrix}.$$

Then, we set $\theta_1 = 4.0\times10^{-33}$, $\theta_2 = 3.2\times10^{-33}$, $\theta_3 = 2.2\times10^{-33}$. We still set $\{\beta_1, \beta_2, \beta_3\} = \{\beta_G, \beta_{GC}, \beta_{GC^2}\} = (0.1, 199.29, -183.30)$. After the data simulations, we still added some contaminations as described above. We still repeated simulating data for 100 normal subjects and 100 tumor subjects for 1,000 times. Again, the above four different censoring strategies were considered and the absolute deviation of estimates was used to compare different results. Figures 7 and 8 shows the results. The absolute deviations based on "no censor" were clearly overall larger among different censoring strategies. It was not surprising that "censor exactly" was the best choice, but "censor more" was also a comparable choice. The absolute deviations based on "censor few" were overall between these based on "no censor" and "censor more" (consistently observed for different parameter estimates).

DISCUSSION AND CONCLUSIONS

In this study, we proposed an artificial censoring approach to the analysis of RNA-seq data. Due to the complicated experimental procedure for data collection, it was difficult to consider simple statistical models/distributions in the related data analysis. Particularly, it was difficult to fit the data of low expression and high expression. With an artificial censoring approach, we achieved desirable robust analysis results. Furthermore, similar as traditional semiparametric statistical methods, our approach could be more powerful when it was difficult to specify an

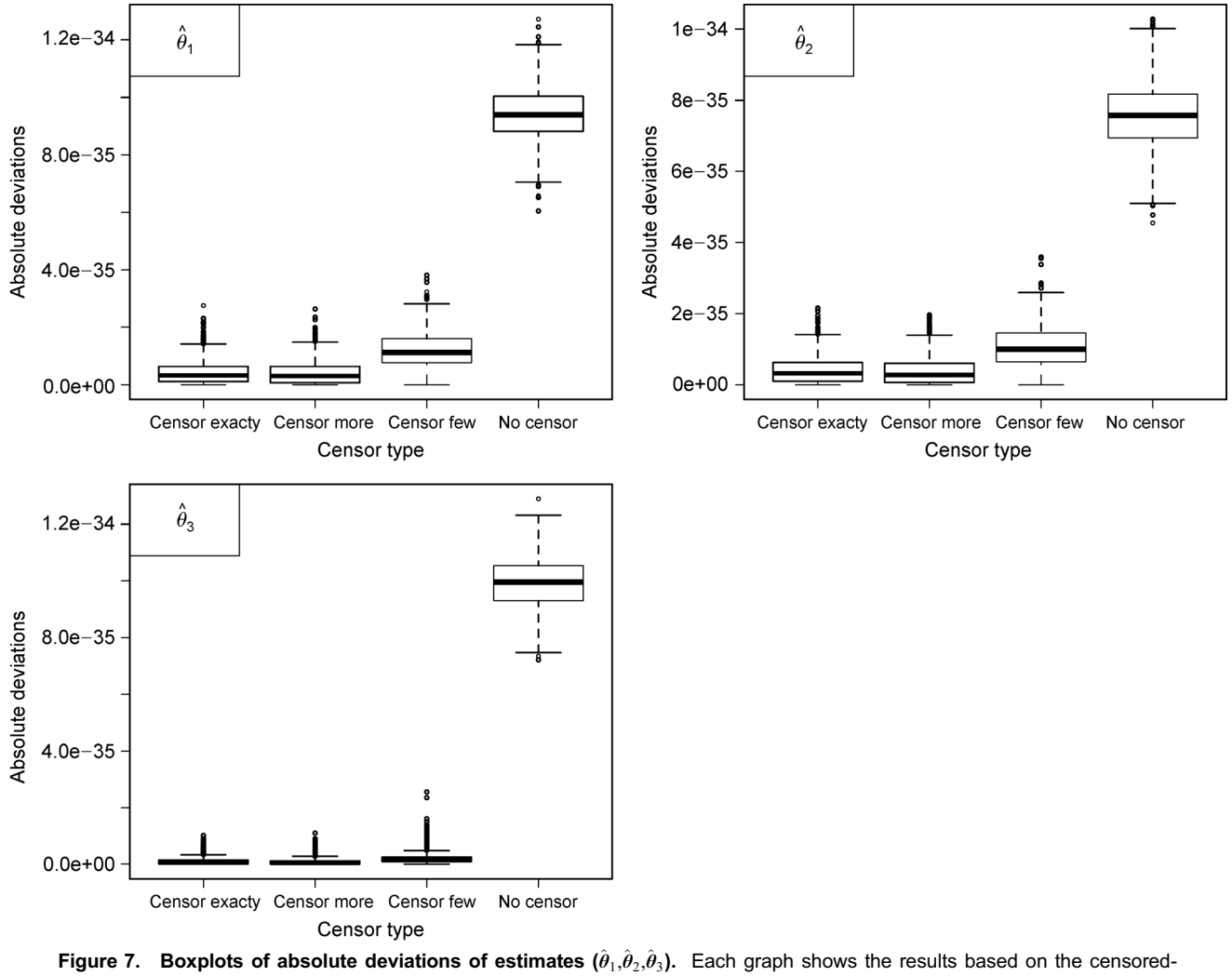

Figure 7. Boxplots of absolute deviations of estimates $(\hat{\theta}_1,\hat{\theta}_2,\hat{\theta}_3)$. Each graph shows the results based on the censored-Poisson model with three different lower/upper bound settings as well as the non-censored Poisson model.

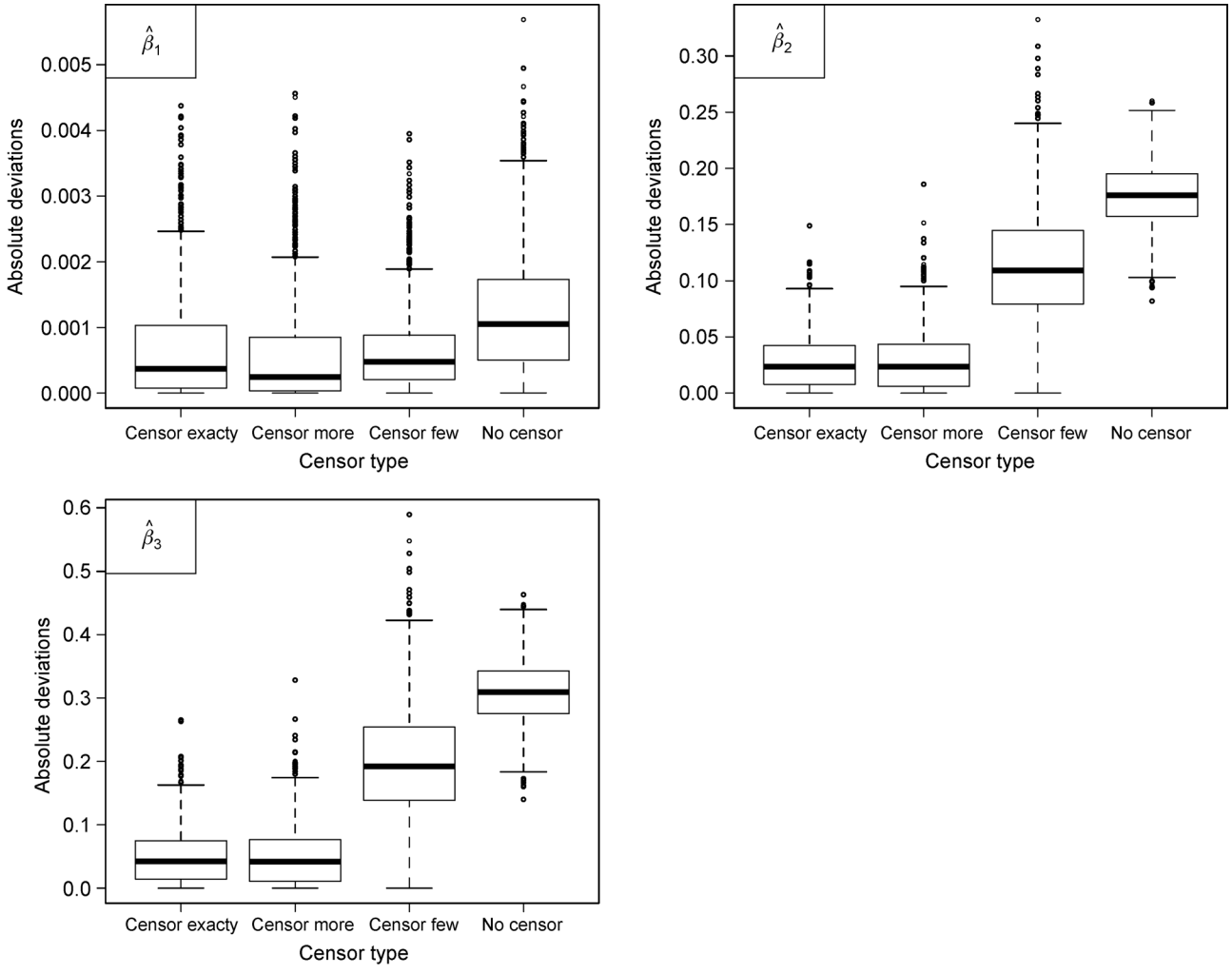

Figure 8. Boxplots of absolute deviations of estimates ($\hat{\beta}_1$,$\hat{\beta}_2$,$\hat{\beta}_3$). Each graph shows the results based on the censored-Poisson model with three different lower/upper bound settings as well as the non-censored Poisson model.

appropriate distribution for the overall range of data. The simulation analysis results and application results presented in this study confirmed our artificial censoring approach.

We demonstrated the improved analysis results after applying an artificial censoring to a traditional Poisson regression model for RNA-seq data analysis. Our proposed artificial censoring approach can certainly be considered for other well-developed models or methods for RNA-seq data analysis, such as Poisson mixture models and negative binomial models. When the artificial censoring is considered, a selected model/method can be more generally useful and efficient, especially in the situation that a large number of features (*e.g.*, genes) are analyzed simultaneously with the same form of models. Notice that, for a selected model/method for analyzing RNA-seq data, our approach is actually a modification that introduces more flexibility in fitting the data. Without any artificial censoring, it is still the originally selected model/method. With artificial censoring, it can be considered as a degenerated form of the originally selected model/method. We have demonstrated such a modification (artificial censoring) to the traditional Poisson regression model. For the modification of artificial censoring to other models/methods, it is necessary to devote research efforts for the related methodological developments and analysis evaluations, which will be pursued as our future research topics.

It was difficult for us to identify an optimization approach for setting the lower and upper bounds for artificial censoring. Therefore, in this study, we would simply suggest setting these two values as approximately 15-percentile and 85-percentile of data, respectively. Other percentile-based values could certainly be con-

sidered. Our simulation study results were also useful for this purpose in practice. We would leave this flexibility to users who are interested in considering artificial censoring in their RNA-seq data analysis.

Numerical computing is essential to our approach, and there are some related common practical difficulties. These have been well addressed in the literature of numerical computing. To avoid the decrease of likelihood during iterative computing, we would suggest the well-established backtracking procedure. To avoid numerical singularities in the calculations of inverse of Hessian matrices, we would suggest the well-established block-computing approach. To set appropriate initial values, we would suggest these from a non-censored model (*e.g.*, a traditional Poisson regression model).

In RNA-seq data analysis, the non-uniformity of short reads has been a challenging concern. Li *et al*. [5] introduced two models for fitting the non-uniformity in short read rates based on local sequences. Our approach is based on the traditional Poisson regression models, and similar considerations can also be flexibly incorporated into our models for the concern of non-uniformity of short reads. The artificial censoring approach can also be considered in the mixture Poisson-model based statistical methods for analyzing RNA-seq data (to achieve more robust analysis results). Furthermore, this approach can be considered in the recently developed statistical methods for single-cell RNA-seq data analysis. Additionally, it is interesting to extend our artificial censoring approach to the negative binomial distribution based methods for RNA-seq data analysis.

MATERIALS AND METHODS

Censored-Poisson regression model

Our methodological development was motivated by the models proposed by Jiang and Wong [7], Salzman, Jiang and Wong [8] and Shi and Jiang [9]. Before the description of our model, we list the related mathematical notations in Table 4.

Our model is still based on the traditional Poisson distribution/regression. For a gene $g \in G$, a subject $m \in M$, we assume that the expected value of the number of read counts Y_{mj} from exon j is given by the following equation.

$$\lambda_{mj} = E(Y_{mj}) = n_m * \sum_{i=1}^{I} l_{ij}\theta_i * \exp(X^{\mathrm{T}}\boldsymbol{\beta}), \qquad (1)$$

where X is the covariates matrix (*e.g.*, group assignment, GC content, etc.) for the coefficient vector $\boldsymbol{\beta}$. The list of covariates could be different for different RNA-seq data sets and/or analysis purpose. (In a practical RNA-seq data analysis, the patient's demographic/clinical features can certainly be considered when available. Feature/variable selection is also an important concern related to this. These topics are out of the scope of this study.)

Table 4 Mathematical notations for our censored-Poisson regression model

Symbol	Meaning
G	Set of genes
g	A gene in the set of genes
M	Number of subjects
m	Subject index
I	Number of isoforms of a gene
i	Isoform index
J	Number of exons of an isoform
j	Exon index
Y	Poisson random variable
λ	Mean of Poisson random variable
n_m	Number of read counts from subject m
X	Covariates matrix
$\boldsymbol{\beta}$	Coefficient vector
$\boldsymbol{\theta}$	Isoform-specific coefficient vector
l_{ij}	Length of j-th exon from i-th isoform

In the above equation, each θ_i can be included into the exponential function as an isoform-specific intercept $\beta_{0i} = \log(\theta_i)$. It is essentially a Poisson regression model with a specified mean structure. This model may be flexibly used in practice for evaluating differential expression (group effect), GC content effect, etc. However, in practice, a simple Poisson regression model usually lacks of robustness (*e.g.*, due to a simple distribution assumption). In this study, we consider that it is difficult to model low count values (less than a given value a as lower bound) and high count values (greater than a given value b as upper bound) with a simple distribution, but the count values between a and b can be efficiently described by a Poisson distribution ($0 < a < b < \infty$). [This is based on our data analysis experience. Rigorously speaking, we would like to consider this as an assumption, especially when a large number of features (*e.g.*, genes) were analyzed with the same form of models.] Therefore, we propose to artificially censor count values less than a as one interval category, and to artificially censor count values greater than b as another interval category. (Notice that no data were discarded in our analysis.)

For each Y_{mj}, let δ_{mj} be a related indicator: $\delta_{mj} = 1$ when $Y_{mj} < a$ or zero otherwise; let δ'_{mj} also be a related indicator: $\delta'_{mj} = 1$ when $Y_{mj} > b$ or zero otherwise.

We propose the following likelihood function.

$$L = \prod_{m=1}^{M}\prod_{j=1}^{J} [Pr(Y_{mj} < a)]^{\delta_{mj}} [Pr(Y_{mj} > b)]^{\delta'_{mj}}$$

$$[Pr(Y_{mj} = y_{mj})]^{1-\delta_{mj}-\delta'_{mj}}, \qquad (2)$$

which can be calculated as:

$$L=\prod_{m=1}^{M}\prod_{j=1}^{J}\left[\sum_{k<a}\frac{e^{-\lambda_{mj}}(\lambda_{mj})^{k}}{k!}\right]^{\delta_{mj}}\left[\sum_{k>b}\frac{e^{-\lambda_{mj}}(\lambda_{mj})^{k}}{k!}\right]^{\delta'_{mj}}\left[\frac{e^{-\lambda_{mj}}(\lambda_{mj})^{y_{mj}}}{y_{mj}!}\right]^{1-\delta_{mj}-\delta'_{mj}}. \quad (3)$$

We use the well-established Newton-Raphson method to obtain the maximum likelihood estimates for $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. The related mathematical details are provided in an Appendix, which includes several non-trivial formula simplifications. These simplifications are essential to improve the necessary numerical computing (by utilizing their existing R-functions).

TCGA RNA-seq data

The Cancer Genome Atlas (TCGA) is a comprehensive cancer research project [23]. Pre-processed RNA-seq data sets for different types of cancer have been made publically available. We downloaded the TCGA RNA-seq data for breast cancer study. During the progress of our research development, the database had been constantly updated. At the time of our application analysis, we downloaded the data for 101 normal subjects and 96 tumor subjects, and these data were still appropriate as illustrative examples for our method.

UCSC Genome Browser

TCGA data used the UCSC Genome Browser hg19 (2009) as the reference genome. To obtain isoform information for a given gene, we searched the corresponding exon locations and isoform structure from the UCSC genome browser [24]. In summary, we obtained the exon information (*e.g.*, location, length) based on the data from TCGA and UCSC Genome Browser. In this study, we focused on the exon based RNA-seq data analysis. Therefore, we have adequate isoform information and RNA-seq data for our analysis.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Xing Chen and Yinglei Lai declare that they have no conflict of interests.

This article does not contain any study materials with human or animal subjects performed by any of the authors.

APPENDIX

Mathematical derivations

Multiple isoforms are mixed and the isoform-specific parameters are θ_i, $i=1,...,I$. We choose to estimate each θ_i and β_w separately. For applying Newton-Raphson method, we need the first derivatives and the second derivatives with respect to (w.r.t.) all the parameters. In the following, we first provide the first derivatives and the second derivative w.r.t β's. Then, we provide these w.r.t. θ's, etc.

Based on the model (Eq. (1)) and the likelihood function (Eq. (2)), the first derivative w.r.t β_w is

$$\frac{\partial}{\partial\beta_w}\log L=\sum_{m=1}^{M}\sum_{j=1}^{J}\left\{\delta_{mj}\frac{\partial}{\partial\beta_w}\log Pr(Y_{mj}<a)+\delta'_{mj}\frac{\partial}{\partial\beta_w}\log Pr(Y_{mj}>b)+(1-\delta_{mj}-\delta'_{mj})\frac{\partial}{\partial\beta_w}\log Pr(Y_{mj}=y_{mj})\right\}. \quad (A1)$$

The first part of lower censoring is

$$\begin{aligned}
&\frac{\partial}{\partial\beta_w}\log Pr(Y_{mj}<a)\\
&=\frac{1}{Pr(Y_{mj}<a)}\frac{\partial}{\partial\beta_w}\left[\sum_{k=0}^{a-1}\frac{e^{-n_m\sum_{i=1}^{I}l_{ij}\theta_i\exp(X^{\mathrm{T}}\boldsymbol{\beta})}(n_m\sum_{i=1}^{I}l_{ij}\theta_i\exp(X^{\mathrm{T}}\boldsymbol{\beta}))^{k}}{k!}\right]\\
&=\frac{1}{Pr(Y_{mj}<a)}\frac{\partial}{\partial\beta_w}\left[e^{-n_m\sum_{i=1}^{I}l_{ij}\theta_i\exp(X^{\mathrm{T}}\boldsymbol{\beta})}+\sum_{k=1}^{a-1}\frac{e^{-n_m\sum_{i=1}^{I}l_{ij}\theta_i\exp(X^{\mathrm{T}}\boldsymbol{\beta})}(n_m\sum_{i=1}^{I}l_{ij}\theta_i\exp(X^{\mathrm{T}}\boldsymbol{\beta}))^{k}}{k!}\right]\\
&=\frac{e^{-\lambda_{mj}}}{Pr(Y_{mj}<a)}\left[-\lambda_{mj}-\sum_{k=1}^{a-1}\frac{\lambda_{mj}^{k+1}}{k!}+\sum_{k=1}^{a-1}\frac{\lambda_{mj}^{k}}{(k-1)!}\right]x_{mjw}\\
&=-\frac{aPr(Y_{mj}=a)}{Pr(Y_{mj}<a)}x_{mjw}.
\end{aligned}$$

Similarly, the second part of upper censoring is

$$\frac{\partial}{\partial\beta_w}\log Pr(Y_{mj}>b)=\frac{\frac{\partial}{\partial\beta_w}Pr(Y_{mj}>b)}{Pr(Y_{mj}>b)}=\frac{\frac{\partial}{\partial\beta_w}[1-Pr(Y_{mj}\leqslant b)]}{Pr(Y_{mj}>b)}=\frac{(b+1)Pr(Y_{mj}=b+1)}{Pr(Y_{mj}>b)}x_{mjw}.$$

Lastly, the third part of no-censoring is

$$\frac{\partial}{\partial\beta_w}\log Pr(Y_{mj}=b)=\frac{\partial}{\partial\beta_w}\left[-n_m\sum_{i=1}^{I}l_{ij}\theta_i\exp(X^{\mathrm{T}}\boldsymbol{\beta})+y_{mj}\log n_m\sum_{i=1}^{I}l_{ij}\theta_i\exp(X^{\mathrm{T}}\boldsymbol{\beta})\right]=(y_{mj}-\lambda_{mj})x_{mjw}.$$

Finally, combining the above three terms, we get the first derivative of the log-likelihood function w.r.t β_w.

$$\frac{\partial}{\partial\beta_w}\log L=\sum_{m=1}^{M}\sum_{j=1}^{J}\left\{\delta_{mj}\frac{-aPr(Y_{mj}=a)}{Pr(Y_{mj}<a)}x_{mjw}+\delta'_{mj}\frac{(b+1)Pr(Y_{mj}=b+1)}{Pr(Y_{mj}>b)}x_{mjw}+(1-\delta_{mj}-\delta'_{mj})(y_{mj}-\lambda_{mj})x_{mjw}\right\}. \tag{A2}$$

Furthermore, we continue to work on the second derivative w.r.t β_{w_1}, β_{w_2}

$$\begin{aligned}\frac{\partial^2}{\partial\beta_{w_2}\partial\beta_{w_1}}\log L=\sum_{m=1}^{M}\sum_{j=1}^{J}\Bigg\{&\delta_{mj}\frac{\partial^2}{\partial\beta_{w_2}\partial\beta_{w_1}}\log Pr(Y_{mj}<a)+\delta'_{mj}\frac{\partial^2}{\partial\beta_{w_2}\partial\beta_{w_1}}\log Pr(Y_{mj}>b)\\&+(1-\delta_{mj}-\delta'_{mj})\frac{\partial^2}{\partial\beta_{w_2}\partial\beta_{w_1}}\log Pr(Y_{mj}=y_{mj})\Bigg\}.\end{aligned} \tag{A3}$$

Then, we work on three parts separately.

$$\begin{aligned}&\frac{\partial^2}{\partial\beta_{w_2}\partial\beta_{w_1}}\log Pr(Y_{mj}<a)\\&=\frac{\partial}{\partial\beta_{w_2}}\left[\frac{\frac{\partial}{\partial\beta_{w_1}}Pr(Y_{mj}<a)}{Pr(Y_{mj}<a)}\right]\\&=\frac{\frac{\partial^2}{\partial\beta_{w_2}\partial\beta_{w_1}}Pr(Y_{mj}<a)Pr(Y_{mj}<a)-\frac{\partial}{\partial\beta_{w_1}}Pr(Y_{mj}<a)\frac{\partial}{\partial\beta_{w_2}}Pr(Y_{mj}<a)}{[Pr(Y_{mj}<a)]^2}\\&=\frac{ax_{mjw_1}x_{mjw_2}[(a+1)Pr(Y_{mj}=a+1)-aPr(Y_{mj}=a)]}{Pr(Y_{mj}<a)}-\frac{ax_{mjw_1}x_{mjw_2}Pr(Y_{mj}=a)Pr(Y_{mj}=a)}{[Pr(Y_{mj}<a)]^2}\\&=\frac{ax_{mjw_1}x_{mjw_2}}{Pr(Y_{mj}<a)}\left\{(a+1)Pr(Y_{mj}=a+1)-aPr(Y_{mj}=a)-\frac{a[Pr(Y_{mj}<a)]^2}{Pr(Y_{mj}<a)}\right\}.\end{aligned}$$

For the second part, it has the similar derivation procedure.

$$\begin{aligned}&\frac{\partial^2}{\partial\beta_{w_2}\partial\beta_{w_1}}\log Pr(Y_{mj}>b)\\&=\frac{\frac{\partial^2}{\partial\beta_{w_2}\partial\beta_{w_1}}Pr(Y_{mj}<b)Pr(Y_{mj}<b)-\frac{\partial}{\partial\beta_{w_1}}Pr(Y_{mj}>b)\frac{\partial}{\partial\beta_{w_2}}Pr(Y_{mj}>b)}{[Pr(Y_{mj}>b)]^2}\\&=\frac{(b+1)x_{mjw_1}x_{mjw_2}}{Pr(Y_{mj}>b)}\left\{-(b+2)Pr(Y_{mj}=b+2)+(b+1)Pr(Y_{mj}=b+1)-\frac{(b+1)[Pr(Y_{mj}=b+1)]^2}{Pr(Y_{mj}>b)}\right\}.\end{aligned}$$

Finally, the third part looks like

$$\begin{aligned}\frac{\partial^2}{\partial\beta_{w_2}\partial\beta_{w_1}}\log Pr(Y_{mj}=y_{mj}) &= \frac{\partial}{\partial\beta_{w_2}}\left[\frac{\partial}{\partial\beta_{w_1}}\log Pr(Y_{mj}=y_{mj})\right]\\ &= \frac{\partial}{\partial\beta_{w_2}}[y_{mj}-\lambda_{mj}x_{mjw_1}]\\ &= \frac{\partial}{\partial\beta_{w_2}}(-\exp(X^{\mathrm{T}}\boldsymbol{\beta}))x_{mjw_1}\\ &= -x_{mjw_1}x_{mjw_2}\lambda_{mj}.\end{aligned}$$

In summary, combining all three terms to get the full second derivative function.

$$\frac{\partial^2}{\partial\beta_{w_2}\partial\beta_{w_1}}\log L$$

$$=\sum_{m=1}^{M}\sum_{j=1}^{J}\left\{\delta_{mj}\frac{ax_{mjw_1}x_{mjw_2}}{Pr(Y_{mj}<a)}\left\{(a+1)Pr(Y_{mj}=a+1)-aPr(Y_{mj}=a)-\frac{a[Pr(Y_{mj}<a)]^2}{Pr(Y_{mj}<a)}\right\}\right.$$

$$+\delta'_{mj}\frac{(b+1)x_{mjw_1}x_{mjw_2}}{Pr(Y_{mj}>b)}\left\{-(b+2)Pr(Y_{mj}=b+2)+(b+1)Pr(Y_{mj}=b+1)-\frac{(b+1)[Pr(Y_{mj}=b+1)]^2}{Pr(Y_{mj}>b)}\right\}$$

$$\left.-(1-\delta_{mj}-\delta'_{mj})-x_{mjw_1}x_{mjw_2}\lambda_{mj}\right\}. \tag{A4}$$

Based on the model (Eq. (1)) and the likelihood function (Eq. (2)), the first derivative w.r.t θ_i is

$$\frac{\partial}{\partial\theta_i}\log L=\sum_{m=1}^{M}\sum_{j=1}^{J}\left\{\delta_{mj}\frac{\partial}{\partial\theta_i}\log Pr(Y_{mj}<a)+\delta'_{mj}\frac{\partial}{\partial\theta_i}\log Pr(Y_{mj}>b)+(1-\delta_{mj}-\delta'_{mj})\frac{\partial}{\partial\theta_i}\log Pr(Y_{mj}=y_{mj})\right\}. \tag{A5}$$

The first part of lower censoring is

$$\frac{\partial}{\partial\theta_i}\log Pr(Y_{mj}<a)=\frac{1}{Pr(Y_{mj}<a)}\frac{\partial}{\partial\theta_i}\left\{e^{-n_m\sum_{i=1}^{I}l_{ij}\theta_i\exp(X^{\mathrm{T}}\boldsymbol{\beta})}+\sum_{k=1}^{a-1}\frac{e^{-n_m\sum_{i=1}^{I}l_{ij}\theta_i\exp(X^{\mathrm{T}}\boldsymbol{\beta})}(n_m\sum_{i=1}^{I}l_{ij}\theta_i\exp(X^{\mathrm{T}}\boldsymbol{\beta}))^k}{k!}\right\}$$

$$=\frac{n_m l_{ij}\exp(X^{\mathrm{T}}\boldsymbol{\beta})}{Pr(Y_{mj}<a)}\left\{-e^{-\lambda_{mj}}+\sum_{k=1}^{a-1}\left[-\frac{e^{-\lambda_{mj}}\lambda_{mj}^k}{k!}+\frac{e^{-\lambda_{mj}}\lambda_{mj}^{k-1}}{(k-1)!}\right]\right\}$$

$$=-\frac{n_m l_{ij}\exp(X^{\mathrm{T}}\boldsymbol{\beta})}{Pr(Y_{mj}<a)}\frac{e^{-\lambda_{mj}}\lambda_{mj}^{a-1}}{(a-1)!}$$

$$= -\frac{n_m l_{ij} \exp(X^{\mathrm{T}}\boldsymbol{\beta}) Pr(Y_{mj} = a-1)}{Pr(Y_{mj} < a)}.$$

Similarly, the second part for upper censoring is

$$\frac{\partial}{\partial \theta_i} \log Pr(Y_{mj} > b) = \frac{\frac{\partial}{\partial \theta_i} Pr(Y_{mj} > b)}{Pr(Y_{mj} > b)} = \frac{\frac{\partial}{\partial \theta_i}[1 - Pr(Y_{mj} \leqslant b)]}{Pr(Y_{mj} > b)} = \frac{n_m l_{ij} \exp(X^{\mathrm{T}}\boldsymbol{\beta}) Pr(Y_{mj} = b)}{Pr(Y_{mj} > b)}.$$

Finally, the third part of no-censoring is

$$\begin{aligned}
\frac{\partial}{\partial \theta_i} \log Pr(Y_{mj} = y_{mj}) &= \frac{\partial}{\partial \theta_i}\left[-n_m \sum_{i=1}^{I} l_{ij}\theta_i \exp(X^{\mathrm{T}}\boldsymbol{\beta}) + y_{mj} \log\left(n_m \sum_{i=1}^{I} l_{ij}\theta_i \exp(X^{\mathrm{T}}\boldsymbol{\beta})\right)\right] \\
&= -n_m l_{ij} \exp(X^{\mathrm{T}}\boldsymbol{\beta}) + \frac{y_{mj} n_m l_{ij}\theta_i \exp(X^{\mathrm{T}}\boldsymbol{\beta})}{\lambda_{mj}} \\
&= n_m l_{ij} \exp(X^{\mathrm{T}}\boldsymbol{\beta})\left(\frac{y_{mj}}{\lambda_{mj}} - 1\right).
\end{aligned}$$

In summary, the final function of first derivative w.r.t θ_i is

$$\frac{\partial}{\partial \theta_i} \log L = \sum_{m=1}^{M}\sum_{j=1}^{J}\left\{\delta_{mj}\frac{-n_m l_{ij} \exp(X^{\mathrm{T}}\boldsymbol{\beta}) Pr(Y_{mj} = a-1)}{Pr(Y_{mj} < a)} + \delta'_{mj}\frac{n_m l_{ij}\theta_i \exp(X^{\mathrm{T}}\boldsymbol{\beta}) Pr(Y_{mj} = b)}{Pr(Y_{mj} > b)} + (1 - \delta_{mj} - \delta'_{mj}) n_m l_{ij} \exp(X^{\mathrm{T}}\boldsymbol{\beta})\left(\frac{y_{mj}}{\lambda_{mj}} - 1\right)\right\}. \tag{A6}$$

Furthermore, we continue to work on the second derivative w.r.t θ_{i_1} θ_{i_2}.

$$\frac{\partial^2}{\partial \theta_{i_2} \partial \theta_{i_1}} \log L = \sum_{m=1}^{M}\sum_{j=1}^{J}\left\{\delta_{mj}\frac{\partial^2}{\partial \theta_{i_2} \partial \theta_{i_1}} \log Pr(Y_{mj} < a) + \delta'_{mj}\frac{\partial^2}{\partial \theta_{i_2} \partial \theta_{i_1}} \log Pr(Y_{mj} > b) + (1 - \delta_{mj} - \delta'_{mj})\frac{\partial^2}{\partial \theta_{i_2} \partial \theta_{i_1}} \log Pr(Y_{mj} = y_{mj})\right\}. \tag{A7}$$

By following the above Eq. (A7), we derive the three parts separately as before.

The first part for lower censoring is

$$\begin{aligned}
&\frac{\partial^2}{\partial \theta_{i_2} \partial \theta_{i_1}} \log Pr(Y_{mj} < a) \\
&= \frac{\partial}{\partial \theta_{i_2}}\left[\frac{\frac{\partial}{\partial \theta_{i_1}} Pr(Y_{mj} < a)}{Pr(Y_{mj} < a)}\right] \\
&= \frac{\frac{\partial^2}{\partial \theta_{i_2} \partial \theta_{i_1}} Pr(Y_{mj} < a) Pr(Y_{mj} < a) - \frac{\partial}{\partial \theta_{i_1}} Pr(Y_{mj} < a)\frac{\partial}{\partial \theta_{i_2}} Pr(Y_{mj} < a)}{[Pr(Y_{mj} < a)]^2} \\
&= -\frac{n_m^2 l_{i_1 j} l_{i_2 j} \exp(2X^{\mathrm{T}}\boldsymbol{\beta})[-Pr(Y_{mj} = a-1) + Pr(Y_{mj} = a-2)]}{Pr(Y_{mj} < a)} - \frac{n_m^2 l_{i_1 j} l_{i_2 j} \exp(2X^{\mathrm{T}}\boldsymbol{\beta})[Pr(Y_{mj} = a-1)]^2}{[Pr(Y_{mj} < a)]^2} \\
&= \frac{n_m^2 l_{i_1 j} l_{i_2 j} \exp(2X^{\mathrm{T}}\boldsymbol{\beta})}{Pr(Y_{mj} < a)}\left\{Pr(Y_{mj} = a-1) - Pr(Y_{mj} = a-2) - \frac{[Pr(Y_{mj} = a-1)]^2}{Pr(Y_{mj} < a)}\right\}.
\end{aligned}$$

Then the second part for upper censoring is

$$\frac{\partial^2}{\partial\theta_{i_2}\partial\theta_{i_1}}\log Pr(Y_{mj}>b)=\frac{\partial}{\partial\theta_{i_2}}\left[\frac{\frac{\partial}{\partial\theta_{i_1}}Pr(Y_{mj}>b)}{Pr(Y_{mj}>b)}\right]$$

$$=\frac{\frac{\partial^2}{\partial\theta_{i_2}\partial\theta_{i_1}}Pr(Y_{mj}>b)Pr(Y_{mj}>b)-\frac{\partial}{\partial\theta_{i_1}}Pr(Y_{mj}>b)\frac{\partial}{\partial\theta_{i_2}}Pr(Y_{mj}>b)}{[Pr(Y_{mj}>b)]^2}$$

$$=\frac{n_m^2 l_{i_1 j}l_{i_2 j}\exp(2X^{\mathrm{T}}\boldsymbol{\beta})[-Pr(Y_{mj}=b)+Pr(Y_{mj}=b-1)]}{Pr(Y_{mj}>b)}-\frac{n_m^2 l_{i_1 j}l_{i_2 j}\exp(2X^{\mathrm{T}}\boldsymbol{\beta})[Pr(Y_{mj}=b)]^2}{[Pr(Y_{mj}>b)]^2}$$

$$=\frac{n_m^2 l_{i_1 j}l_{i_2 j}\exp(2X^{\mathrm{T}}\boldsymbol{\beta})}{Pr(Y_{mj}>b)}\left\{-Pr(Y_{mj}=b)+Pr(Y_{mj}=b-1)-\frac{[Pr(Y_{mj}=b)]^2}{Pr(Y_{mj}>b)}\right\}.$$

For the third part of no-censoring, we have

$$\frac{\partial^2}{\partial\theta_{i_2}\partial\theta_{i_1}}\log Pr(Y_{mj}=y_{mj})=\frac{\partial}{\partial\theta_{i_2}}\left[-l_{i_1 j}\exp(X^{\mathrm{T}}\boldsymbol{\beta})+\frac{y_{mj}l_{i_1 j}}{\sum_{i=1}^{I}l_{ij}\theta_i}\right]=-\frac{y_{mj}l_{i_1 j}l_{i_2 j}}{(\sum_{i=1}^{I}l_{ij}\theta_i)^2}.$$

Combining all three parts, we get the general function of second derivative w.r.t θ_{i_1}, θ_{i_2}.

$$\begin{aligned}\frac{\partial^2}{\partial\theta_{i_2}\partial\theta_{i_1}}\log L=\sum_{m=1}^{M}\sum_{j=1}^{J}\Bigg\{&\delta_{mj}\frac{n_m^2 l_{i_1 j}l_{i_2 j}\exp(2X^{\mathrm{T}}\boldsymbol{\beta})}{Pr(Y_{mj}<a)}\left\{Pr(Y_{mj}=a-1)-Pr(Y_{mj}=a-2)-\frac{[Pr(Y_{mj}=a-1)]^2}{Pr(Y_{mj}<a)}\right\}\\&+\delta'_{mj}\frac{n_m^2 l_{i_1 j}l_{i_2 j}\exp(2X^{\mathrm{T}}\boldsymbol{\beta})}{Pr(Y_{mj}>b)}\left\{-Pr(Y_{mj}=b)+Pr(Y_{mj}=b-1)-\frac{[Pr(Y_{mj}=b)]^2}{Pr(Y_{mj}>b)}\right\}\\&+(1-\delta_{mj}-\delta'_{mj})\left[-\frac{y_{mj}l_{i_1 j}l_{i_2 j}}{(\sum_{i=1}^{I}l_{ij}\theta_i)^2}\right]\Bigg\}.\end{aligned}\tag{A8}$$

Next, we need to derive the second derivative w.r.t θ_i, β_w for the Hessian matrix. The general function looks like

$$\frac{\partial^2}{\partial\theta_{i_2}\partial\theta_{i_1}}\log L=\sum_{m=1}^{M}\sum_{j=1}^{J}\left\{\delta_{mj}\frac{\partial^2}{\partial\beta_w\partial\theta_i}\log Pr(Y_{mj}<a)+\delta'_{mj}\frac{\partial^2}{\partial\beta_w\partial\theta_i}\log Pr(Y_{mj}>b)+(1-\delta_{mj}-\delta'_{mj})\frac{\partial^2}{\partial\beta_w\partial\theta_i}\log Pr(Y_{mj}=y_{mj})\right\}.\tag{A9}$$

Deriving the first part of Eq. (A9) for lower censoring.

$$\frac{\partial^2}{\partial\beta_w\partial\theta_i}\log Pr(Y_{mj}<a)$$

$$=\frac{\partial}{\partial\beta_w}\left[\frac{\frac{\partial}{\partial\theta_i}Pr(Y_{mj}<a)}{Pr(Y_{mj}<a)}\right]$$

$$=\frac{\frac{\partial^2}{\partial\beta_w\partial\theta_i}Pr(Y_{mj}<a)Pr(Y_{mj}<a)-\frac{\partial}{\partial\theta_i}Pr(Y_{mj}<a)\frac{\partial}{\partial\beta_w}Pr(Y_{mj}<a)}{[Pr(Y_{mj}<a)]^2}$$

$$=\frac{an_m l_{ij}\exp(X^{\mathrm{T}}\boldsymbol{\beta})x_{mjw}[Pr(Y_{mj}=a)-Pr(Y_{mj}=a-1)]}{Pr(Y_{mj}<a)}-\frac{n_m l_{ij}\exp(X^{\mathrm{T}}\boldsymbol{\beta})Pr(Y_{mj}=a-1)ax_{mjw}Pr(Y_{mj}=a)}{[Pr(Y_{mj}<a)]^2}$$
$$=\frac{an_m l_{ij}\exp(X^{\mathrm{T}}\boldsymbol{\beta})x_{mjw}}{Pr(Y_{mj}<a)}\left\{Pr(Y_{mj}=a)-Pr(Y_{mj}=a-1)-\frac{Pr(Y_{mj}=a)Pr(Y_{mj}=a-1)}{Pr(Y_{mj}<a)}\right\}.$$

Then, the second part for upper censoring is

$$\frac{\partial^2}{\partial\beta_w\partial\theta_i}\log Pr(Y_{mj}>b)$$
$$=\frac{\frac{\partial^2}{\partial\beta_w\partial\theta_i}Pr(Y_{mj}>b)Pr(Y_{mj}>b)-\frac{\partial}{\partial\theta_i}Pr(Y_{mj}>b)\frac{\partial}{\partial\beta_w}Pr(Y_{mj}>b)}{[Pr(Y_{mj}>b)]^2}$$
$$=\frac{(b+1)n_m l_{ij}\exp(X^{\mathrm{T}}\boldsymbol{\beta})x_{mjw}[-Pr(Y_{mj}=b+1)+Pr(Y_{mj}=b)]}{Pr(Y_{mj}>b)}-\frac{n_m l_{ij}\exp(X^{\mathrm{T}}\boldsymbol{\beta})Pr(Y_{mj}=b)(b+1)x_{mjw}Pr(Y_{mj}=b+1)}{[Pr(Y_{mj}>b)]^2}$$
$$=\frac{(b+1)n_m l_{ij}\exp(X^{\mathrm{T}}\boldsymbol{\beta})x_{mjw}}{Pr(Y_{mj}>b)}\left\{-Pr(Y_{mj}=b+1)+Pr(Y_{mj}=b)-\frac{Pr(Y_{mj}=b)Pr(Y_{mj}=b+1)}{Pr(Y_{mj}>b)}\right\}.$$

For the third part of non-censor term, it should be

$$\frac{\partial^2}{\partial\beta_w\partial\theta_i}\log Pr(Y_{mj}=y_{mj})=\frac{\partial}{\partial\beta_w}\left\{\frac{\partial}{\partial\theta_i}\log Pr(Y_{mj}=y_{mj})\right\}=\frac{\partial}{\partial\beta_w}\left\{n_m l_{ij}\exp(X^{\mathrm{T}}\boldsymbol{\beta})\left(\frac{y_{mj}}{\lambda_{mj}}-1\right)\right\}=-n_m l_{ij}\exp(X^{\mathrm{T}}\boldsymbol{\beta})x_{mjw}.$$

Thus, we can combine all three terms to obtain the second derivative w.r.t θ_i, β_w,

$$\frac{\partial^2}{\partial\beta_w\partial\theta_i}\log L=\sum_{m=1}^{M}\sum_{j=1}^{J}\left\{\delta_{mj}\frac{an_m l_{ij}\exp(X^{\mathrm{T}}\boldsymbol{\beta})x_{mjw}}{Pr(Y_{mj}<a)}\left\{Pr(Y_{mj}=a)-Pr(Y_{mj}=a-1)-\frac{Pr(Y_{mj}=a)Pr(Y_{mj}=a-1)}{Pr(Y_{mj}<a)}\right\}\right.$$
$$+\delta'_{mj}\frac{(b+1)n_m l_{ij}\exp(X^{\mathrm{T}}\boldsymbol{\beta})x_{mjw}}{Pr(Y_{mj}>b)}\left\{-Pr(Y_{mj}=b+1)+Pr(Y_{mj}=b)-\frac{Pr(Y_{mj}=b)Pr(Y_{mj}=b+1)}{Pr(Y_{mj}>b)}\right\}$$
$$\left.-(1-\delta_{mj}-\delta'_{mj})n_m l_{ij}\exp(X^{\mathrm{T}}\boldsymbol{\beta})x_{mjw}\right\}.$$

After we derived all first derivatives and second derivatives, we can build the vector of first derivative and Hessian matrix to apply Newton-Raphson method. The first derivative vector is $D^{\mathrm{T}}=\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log L,\frac{\partial}{\partial\boldsymbol{\beta}}\log L\right\}$ and the Hessian matrix is

$$H=\begin{bmatrix}\frac{\partial^2}{\partial\boldsymbol{\theta}^2}\log L & \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\theta}}\log L\\ \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\beta}}\log L & \frac{\partial^2}{\partial\boldsymbol{\beta}^2}\log L\end{bmatrix},$$

Then, we can implement Newton-Raphson method to obtain the maximum likelihood estimate (MLE) for each θ_i and β_w by the iterative numerical computing: $(\boldsymbol{\theta},\boldsymbol{\beta})^{\mathrm{T}}_{new}=(\boldsymbol{\theta},\boldsymbol{\beta})^{\mathrm{T}}_{old}-H^{-1}D$.

REFERENCES

1. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc., 7, 562–578
2. Alkhateeb, A., and Rueda, L. (2017) Zseq: An approach for preprocessing next-generation sequencing data. J. Comput. Biol., 24, 746–755
3. Pérez-Rubio, P., Lottaz, C., and Engelmann, J. C. (2019) FastqPuri: high-performance preprocessing of RNA-seq data. BMC Bioinformatics, 20, 226
4. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat. Methods, 5, 621–628
5. Li, J., Jiang, H., and Wong, W., H. (2010) Modeling non-uniformity in short-read rates in RNA-seq data. Genome Biol., 11, R50
6. Li, B. and Dewey, C. N. (2011) RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. BMC Bioinformatics, 12, 323
7. Jiang, H. and Wong, W. H. (2009) Statistical inferences for isoform expression in RNA-seq. Bioinformatics, 25, 1026–1032
8. Salzman, J., Jiang, H. and Wong, W. H. (2011) Statistical modeling of RNA-seq data. Stat. Sci., 26, 62–83
9. Shi, Y. and Jiang, H. (2013) rSeqDiff: detecting differential isoform expression from RNA-seq data using hierarchical likelihood ratio test. PLoS One, 8, e79448
10. Dohm, J. C., Lottaz, C., Borodina, T. and Himmelbauer, H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res., 36, e105
11. Aird, D., Ross, M. G., Chen, W. S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C. and Gnirke, A. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol., 12, R18
12. Benjamini, Y. and Speed, T. P. (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res., 40, e72
13. Hansen, K. D., Irizarry, R. A. and Wu, Z. (2012) Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics, 13, 204–216
14. Robinson, M. D. and Smyth, G. K. (2007) Moderated statistical tests for assessing differences in tag abundance. Bioinformatics, 23, 2881–2887
15. Robinson, M. D. and Smyth, G. K. (2008) Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics, 9, 321–332
16. Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. Genome Biol., 11, R106
17. Anders, S., McCarthy, D. J., Chen, Y., Okoniewski, M., SmythG. K., Huber, W. and Robinson, M. D. (2013) Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat. Protoc., 8, 1765–1786
18. Rau, A., Maugis-Rabusseau, C., Martin-Magniette, M.-L. and CeleuxG. (2015) Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models. Bioinformatics, 31, 1420–1427
19. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. and Salzberg, S. L. (2016) Transcript-level expression analysis of rna-seq experiments with hisat, stringtie and ballgown. Nat. Protoc., 11, 1650–1667
20. Kazakiewicz, D., Claesen, J., Górczak, K., Plewczynski, D. and Burzykowski, T. (2019) A multivariate negative-binomial model with random effects for differential gene-expression analysis of correlated mrna sequencing data. J. Comput. Biol., 26, 1339–1348
21. Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A. and Dewey, C. N. (2010) RNA-seq gene expression estimation with read mapping uncertainty. Bioinformatics, 26, 493–500
22. Khoury, M. P. and Bourdon, J.-C. (2011) p53 isoforms: An intracellular microprocessor? Genes Cancer, 2, 453–465
23. Cancer Genome Atlas Network. (2012) Comprehensive molecular portraits of human breast tumours. Nature, 490, 61–70
24. Rosenbloom, K. R.Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., *et al*. (2015) The UCSC Genome Browser database: 2015 update. Nucleic Acids Res., 43, D670–D681

