The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1195–1216
DOI: 10.1214/16-AOAS906

DECONVOLUTION OF BASE PAIR LEVEL RNA-SEQ READ COUNTS FOR QUANTIFICATION OF TRANSCRIPT EXPRESSION LEVELS[1]

BY HAN WU* AND YU ZHU*,†

*Purdue University** *and Tsinghua University*†

RNA-Seq has emerged as the method of choice for profiling the transcriptomes of organisms. In particular, it aims to quantify the expression levels of transcripts using short nucleotide sequences or short reads generated from RNA-Seq experiments. In real experiments, the label of the transcript, from which each short read is generated, is missing, and short reads are mapped to the genome rather than the transcriptome. Therefore, the quantification of transcript expression levels is an indirect statistical inference problem.

In this article, we propose to use individual exonic base pairs as observation units and, further, to model nonzero as well as zero counts at all base pairs at both the transcript and gene levels. At the transcript level, two-component Poisson mixture distributions are postulated, which gives rise to the Convolution of Poisson mixture (CPM) distribution model at the gene level. The maximum likelihood estimation method equipped with the EM algorithm is used to estimate model parameters and quantify transcript expression levels. We refer to the proposed method as CPM-Seq. Both simulation studies and real data demonstrate the effectiveness of CPM-Seq, showing that CPM-Seq produces more accurate and consistent quantification results than Cufflinks.

1. Introduction. Transcription is the first step of gene expression, and all transcription products, which are called RNA molecules or transcripts, form the transcriptome of a cell under a given developmental stage or physiological condition. The largest family of transcripts are mRNAs. The number of transcripts is much larger than the number of genes due to gene alternative splicing.

Transcriptome profiling, which is to comprehensively detect, catalog and quantify all transcripts in the transcriptome, is a grand challenge in molecular biology and functional genomics. In the past two decades, microarray has been the primary method for interrogating the transcriptome. Recently, the development of next generation sequencing (NGS) technology has revolutionized the way genomic research is conducted. In particular, NGS technology provides a new avenue for mapping and quantifying the transcriptome. As one of such new technologies,

Received October 2013; revised October 2015.
[1]Supported in part by NSF Grant DMS-10-00443.
Key words and phrases. RNA-Seq, transcriptome profiling, finite Poisson mixture model, convolution.

RNA-Seq directly measures the abundance of transcripts and has become an attractive alternative for profiling the transcriptome [Hu et al. (2012), Mortazavi et al. (2008)].

In a typical RNA-Seq experiment, RNA molecules are fragmented into small pieces and converted to a library of cDNA fragments with adapters attached to one end or both ends. Each fragment, after amplification, is then sequenced using one of the NGS technologies, generating hundreds of millions of short nucleotide sequences or short reads. After sequencing, the resulting reads are either assembled *de novo* or aligned to the reference genome to produce a genome-scale transcriptional profile. Using the mapped short reads (single-end reads or paired-end reads), the number of reads that each base pair of the reference genome receives can be calculated, and the resulting counts are collectively known as the *base level read counts data*.

The read counts data can be used to quantify either the expression level of a region of interest (ROI) on the reference genome, such as an exon or a gene, or the expression levels of transcripts. The expression level of an ROI is relatively easier to quantify because the total number of reads received by the ROI directly reflects its abundance, and thus can be used to measure its expression level after proper normalization. Various statistical model-based quantification methods for ROI's have been proposed in the literature, which include GPseq [Srivastava and Chen (2010)], PMSeq [Wu, Qin and Zhu (2012)] and POME [Hu et al. (2012)]. The expression levels of transcripts are, however, more difficult to quantify because one may not be able to allocate the short reads uniquely to transcripts. Due to alternative splicing, multiple transcripts can give rise to identical reads. In other words, the labels of the transcripts that identical reads are generated from are missing. Therefore, the quantification of transcript expression levels becomes an indirect inference problem.

A number of methods have been proposed for transcript expression level quantification in the literature. Jiang et al. proposed a Poisson model for the total number of reads in each exon or exon-exon junction [Salzman, Jiang and Wong (2011)]. The intensity of the Poisson distribution is further assumed to be a linear combination of the expression levels of the transcripts that contain the exon or exon-exon junction. Trapnell et al. proposed a method called Cufflinks [Trapnell et al. (2010)]. Cufflinks uses a probabilistic model to represent the generating scheme for each read. The probabilistic model involves the expression levels of the transcripts that can produce this read. Li et al. proposed a Lasso regression approach (called IsoLasso) to quantifying the expression levels of transcripts [Li, Feng and Jiang (2011)]. IsoLasso first divides a gene into a set of segments based on exon-intron boundaries and then applies Lasso to regress read counts in the segments against the transcript expression indexes. Li and Dewey proposed a method called RNA-Seq by Expectation Maximization (RSEM) for quantifying the expression level of transcripts [Li and Dewey (2011)]. In RSEM, a noise transcript is introduced to account for reads without alignments. As a result, Li and Dewey claimed

that RSEM achieves more accurate quantification results than other existing methods. Salzman et al. later used the simple Poisson distribution to model the total number of counts that fall into each exon and exon-exon junction [Salzman, Jiang and Wong (2011)]. Li et al. proposed to use a sparse linear model for isoform discovery and abundance estimation (SLIDE) [Li et al. (2011)]. SLIDE is similar to IsoLasso except that SLIDE uses read counts in bins instead of segments, and the mapped reads are considered in the same bin if their starting and ending positions belong to the same exons, respectively. All six methods discussed above use the typical approach for solving indirect problems, which is to use a probabilistic or statistical model to relate observations (i.e., observed reads) to unobserved quantities (i.e., transcript abundances). They differ from each other in terms of the units of observations and the types of models they used. In particular, the Poisson model proposed by Jiang et al. used exon or exon-exon junction as the unit, both Cufflinks and RSEM treat each observed read as the unit, IsoLasso used each segment as the unit, and SLIDE uses each bin as the unit [Li, Feng and Jiang (2011)].

The types of units discussed in the previous paragraph may suffer from some drawbacks. One consequence of using bins, exons or segments as observation units is the loss of information caused by aggregation. A well-known advantage of NGS technologies is that they provide single base resolution. In a typical RNA-Seq experiment, due to fragmentation and sampling, some base pairs of a gene receive reads, while others do not receive reads. It is clear that base pairs with positive read counts can reflect the abundance level of transcripts. We argue that base pairs without reads also contain information about the abundance of transcripts and reflect various uncertainties in an RNA-Seq experiment. Therefore, using only base pairs with read counts but not those without read counts may again result in information loss. The method proposed by Jiang et al. and IsoLasso model the aggregated read counts in exons, segments or junctions rather than base pair level counts. Cufflinks models only observed reads, but does not consider base pairs that do not receive reads. SLIDE models only aggregated bin counts, whereas base pairs with zero read counts and bins without observed reads are ignored. Therefore, these methods may fail to utilize all information contained in RNA-Seq read counts data, and, as a consequence, may fail to accurately quantify the expression levels of some transcripts. In some cases, they may suffer from the nonidentifiability problem, that is, they may not be able to distinguish transcripts that are distinguishable.

To improve upon these existing methods, in this article, we treat each base pair as the observation unit and propose a novel approach that models both zero and nonzero read counts for quantifying transcript expression levels. The generating scheme for the read count at each base pair can be considered involving two steps. In the first step, short reads are generated from each transcript that contains this base pair according to a certain distribution; and in the second step, all these short reads are mapped to the same base pair on the reference genome, which gives rise to the observed read count. In the first step, the label of the transcript from which each short read is generated is conceptually available, whereas, after the second

step, this label becomes missing. Therefore, the second step can be regarded as a convolution step, from which short reads of different transcripts are mixed and the information about their origins is lost.

Instead of directly proposing models for the observed read counts on the genome, we propose to first model the short read counts on the transcripts, and refer to the resulting models as the *transcript level models*. After the transcript level models are available, the models for the base level read counts on the reference genome can then be derived as the convolution of the transcript level models, which are referred to the *genome level models*. In particular, we propose to use the mixture of Poisson models at the transcript level, which leads to the convolution of the mixture of Poisson models at the genome level. In this article, we do not consider the transcript assembly problem. Instead we focus on the quantification of a given set of candidate transcripts. The candidate transcripts set can include annotated transcripts, novel transcripts of current research interest or those assembled by other methods.

In the next section, we propose the Convolution of Poisson Mixture (CPM) distribution to model RNA-Seq base level read counts data, develop algorithms for computing the estimates of parameters, and further propose the quantification method for transcript expression levels. We refer to our proposed method as the CPM-Seq method. We report in Section 3 the results from simulation studies and real data applications regarding the performance of CPM-Seq and its comparison with Cufflinks and RSEM. We conclude the article with further discussion and future research directions.

2. Methods.

2.1. *Isoforms and read types*. The exons of a gene form a partition of the gene's exonic region. In general, these exons represent the smallest units that can be entirely transcribed or skipped during the transcription of the gene. However, it can also happen that only a part of an exon is transcribed. When this happens, the involved exon needs to be further divided into sub-exons so that each exon or sub-exon is either completely retained or excluded in a transcript. Suppose gene g contains k_g exons or sub-exons, which are labeled as $e_1, \ldots, e_{k_g}$ from the left end to the right end of the gene, respectively. Suppose the number of base pairs in e_i is n_i for $1 \le i \le k_g$. Let $n_0 = 0$ and $n = n_1 + \cdots + n_{k_g}$. We index the exonic base pairs of gene g from left to right as $1, \ldots, n$. It is clear that exon e_i consists of the base pairs $\{\sum_{j=0}^{i-1} n_j + 1, \ldots, \sum_{j=0}^{i} n_j\}$ for $1 \le i \le k_g$.

As discussed in the Introduction, the transcription of gene g can produce different transcripts, which are called, alternatively, spliced isoforms. For example, if only e_1 and e_5 are kept during transcription while all the other exons are skipped, the resulting transcript consists of e_1 and e_5, which can be denoted as $T_1 = e_1e_5$; and if only e_1, e_2 and e_6 are kept and the others are skipped, the resulting transcript

is $T_2 = e_1e_2e_6$. Let $\mathcal{T}_g = \{T_1, \ldots, T_N\}$, where $N = |\mathcal{T}_g|$, be a set of candidate transcripts. For example, if $k_g = 5$, the total number of all possible transcripts will be $2^5 - 1 = 31$, and N will be the number of transcripts of interest, which is less than or equal to 31.

For every short read mapped to the annotated region of gene g, its starting and ending positions can be obtained. The starting position and ending position of a single-end read are denoted as *start* and *end*, respectively. Each paired-end read contains two mate pairs, and we denote the starting positions and ending positions of the two mate pairs as $start_1$, end_1, $start_2$ and end_2, respectively. For any two single-end reads, if their starting positions and ending positions belong to exons e_{l_1} and e_{l_2}, respectively, they are said to be of the same type, which is denoted as $r_{l_1l_2}$. Similarly for a paired-end read, if its starting and ending positions $start_1$, end_1, $start_2$ and end_2 are in exons e_{l_1}, e_{l_2}, e_{l_3} and e_{l_4}, it is said to be of type $r_{l_1l_2l_3l_4}$. Let $\mathcal{R}$ denote the collection of all possible read types. The mapped reads can also be classified into nonjunction reads and junction reads. Nonjunction reads are those reads whose starting and ending positions are in the same exon, whereas junction reads are those whose starting and ending positions are not in the same exon. Correspondingly, all possible read types can be classified into nonjunction read types and junction read types. For example, r_{1111} is a nonjunction read type because reads of type r_{1111} have their starting and ending positions in the same exon e_1, whereas r_{1122} is a junction read type because reads of type r_{1122} have the starting and ending positions of the first mate pair in exon e_1 but the starting and ending positions of the second mate pair in exon e_2. Let $\mathcal{N}$ denote the collection of nonjunction read types and $\mathcal{J}$ the collection of junction read types. Then we have $\mathcal{R} = \mathcal{N} \cup \mathcal{J}$. At base pair m of gene g, the number of reads or the total read count starting at this base pair can be obtained, which is denoted as S_m. Furthermore, this total read count can be partitioned into read counts of different types as $S_m = \sum_{r \in \mathcal{R}} Y_m^r$, where Y_m^r denotes the total count of type r ($r \in \mathcal{R}$) reads starting at base pair m.

Consider a paired-end read of type $r_{l_1l_2l_3l_4} \in \mathcal{R}$. If the read is a nonjunction read, then $l_1 = l_2 = l_3 = l_4 = l$, and $r_{l_1l_2l_3l_4}$ can be simplified to be r_{ll}. If the read is a junction read, there are seven possible scenarios, which are $l_1 = l_2 = l_3 < l_4$, $l_1 = l_2 < l_3 = l_4$, $l_1 < l_2 = l_3 = l_4$, $l_1 = l_2 < l_3 < l_4$, $l_1 < l_2 = l_3 < l_4$, $l_1 < l_2 < l_3 = l_4$, and $l_1 < l_2 < l_3 < l_4$. In this article, we aggregate these seven types into one type denoted by $r_{l_1l_4}$. In other words, we consider only the starting position of the first mate pair and the ending position of the second mate pair, and leave the other positions arbitrary. Therefore, only two indexes (i.e., l_1, l_4) are needed to characterize reads involving more than one exon. This simplification is mainly to facilitate clear presentation of the proposed convolution model framework in this short article. The proposed convolution model can be applied to the original read types involving four indexes, although the presentation and computation will become much more involved. Note that such simplification is not needed when dealing with single-end reads.

As discussed in the Introduction, mapped reads and their corresponding counts can be directly used to quantify the overall expression level of a gene, but they may not be directly used to quantify the expression levels of transcripts. The counts of different types of reads Y_m^r's contain more information than the total read counts S_m, but they still cannot be directly used for quantifying transcript expression levels due to the same reason. To properly characterize the distribution of Y_m^r, a convolution model is needed. As discussed in the Introduction, transcript expression level quantification is an indirect statistical inference problem, which is to infer transcript expression levels from the genome level read counts data. To facilitate the inference, we first propose statistical models to represent the reads-generating mechanism for each individual transcript in the RNA-Seq experiment, which are referred to as the transcript level models. Second, we use a convolution model to characterize the read counts of different types at the genome level.

2.2. *Convolution of Poisson mixture models.* For gene g with k_g exons, there are a maximum of $k_g(k_g+1)/2$ all possible types of reads, among which k_g types are nonjunction types and $k_g(k_g-1)/2$ types are junction types. We still use $\mathcal{N}$ and $\mathcal{J}$ to denote the collection of all nonjunction types and the collection of all junction types under consideration, respectively. It is clear that $\mathcal{R}=\mathcal{N}\cup\mathcal{J}$, which is the collection of all possible types. We use E_g to denote the exonic region of gene g, that is, $E_g=e_1\cup e_2\cup\cdots\cup e_{k_g}$. For any type $r\in\mathcal{R}$ and base pair $m\in E_g$, recall Y_m^r denotes the count of type r reads starting at base pair m. We use $Y^r=\{Y_m^r; m\in E^r\}$ to represent the collection of type r read counts, where E^r is the collection of all possible base pairs that can become the starting positions of type r reads. We use $\mathbf{Y}=\{Y^r; r\in\mathcal{R}\}=\{Y_m^r; r\in\mathcal{R}, m\in E^r\}$ to represent the counts for all types of reads defined on all possible base pairs.

Recall that $\mathcal{T}_g=\{T_1,T_2,\ldots,T_N\}$ is the collection of N candidate transcripts for gene g under consideration. Consider transcript $T_t\in\mathcal{T}_g$ for $1\le t\le N$. For base pair m of transcript T_t, that is, $m\in T_t$, and read type $r\in\mathcal{R}$, we define X_{tm}^r to be the count of type r reads generated from T_t in an RNA-Seq experiment, and assume X_{tm}^r follows a two-component mixture of Poisson distribution with the probability mass function

$$f\big(X_{tm}^r=x|\lambda_t,p_t\big)=\sum_{i=1}^{2}p_{ti}\,\mathrm{Poi}(x;\lambda_{ti}), \tag{2.1}$$

where $p_t=(p_{t1},p_{t2})'$ is the vector of mixing proportions satisfying $\sum_{i=1}^2 p_{ti}=1$, $\lambda_t=(\lambda_{t1},\lambda_{t2})'$ is the vector of intensity rates of the two Poisson components, $\mathrm{Poi}(x;\lambda_{ti})=(\lambda_{ti})^x\exp(-\lambda_{ti})/x!$, and x is any non-negative integer. The first Poisson component with intensity rate λ_{t1} is used to model the base pairs that either are not covered in the RNA-Seq experiment or covered with an abnormally smaller number of reads due to various sequencing uncertainties, whereas the second Poisson component with intensity rate λ_{t2} is used to model the base pairs that

are normally covered in the RNA-Seq experiment and λ_{t2} represents the abundance of the transcript.

The two-component Poisson mixture distribution can be considered a generalized zero-inflated Poisson model. The difference between this model and the standard zero-inflated model is that its first component accounts for not only excessive zero counts but also small counts. The reason we choose this model over the standard zero-inflated model is threefold. First, there are three sources that can give rise to zero and small counts, which include background noise, skipped base pairs due to random fragmentation and base pairs improperly covered in an RNA-Seq experiment; and the first component is used to simultaneously characterize those sources. Second, using single-isoform genes, Wu et al. have found that the two-component Poisson mixture distribution fits real RNA-Seq data well [Wu, Qin and Zhu (2012)]. Third, as discussed in the Introduction, zero and small counts also contain information about the abundance of the transcript; and, together with the second component, the intensity rate of the first component will be used to produce a more accurate shrinkage estimate of the transcript expression level (see Section 7 of the supplementary material [Wu and Zhu (2016)]). The standard zero-inflated Poisson model cannot be used to serve the same purpose.

Note that X_{tm}^r for $m \in T_t$, $T_t \in \mathcal{T}_g$ and $r \in \mathcal{R}$ may not be always directly observable. After the reads are mapped to the annotated region of gene g, the transcript label t is missing as discussed previously. Instead of observing X_{tm}^r, we may only observe Y_m^r, which is the total count of type r reads for $m \in E^r$. There, however, exists a relationship between Y_m^r and X_{tm}^r, which can be obtained explicitly when the collection of transcripts $\mathcal{T}_g$ is given. Consider a base pair $m \in E^r$ for $r \in \mathcal{R}$. Suppose a total of N_r transcripts $\{T_{i_1}, \ldots, T_{i_{N_r}}\} \subset \mathcal{T}_g$ can give rise to type r reads. Let $X_{i_1 m}^r, \ldots, X_{i_{N_r} m}^r$ be the counts of type r reads at base pair m from the candidate transcripts $T_{i_1}, \ldots, T_{i_{N_r}}$, respectively. Then Y_m^r is the sum of X_{km}^r for $k = i_1, \ldots, i_{N_r}$, that is, $Y_m^r = X_{i_1 m}^r + \cdots + X_{i_{N_r} m}^r$. Therefore, the distribution Y_m^r is the convolution of the distributions of $X_{i_1 m}^r, \ldots, X_{i_{N_r} m}^r$. Because X_{km}^r follows the two-component mixture of Poisson distribution as defined previously in model (2.1) with $f(X_{km}^r = x) = \sum_{i=1}^2 p_{ki} \operatorname{Poi}(x; \lambda_{ki})$ for $k \in \{i_1, i_2, \ldots, i_{N_r}\}$, the distribution of Y_m^r can be derived explicitly, which is a 2^{N_r}-component mixture of Poisson distribution with the following probability mass function:

$$
\begin{aligned}
p(Y_m^r = y) &\equiv f\big(y | \{T_{i_1} \cdots T_{i_{N_r}}\}\big) \\
&= \prod_{k=i_1}^{i_{N_r}} \Bigg[\sum_{j_k=1}^{2} p_{k j_k} \operatorname{Poi}(y; \lambda_{k j_k}) \Bigg] \\
&= \sum_{j_{i_1}} \cdots \sum_{j_{i_{N_r}}} \big[p_{i_1 j_{i_1}} \cdots p_{i_{N_r} j_{i_{N_r}}} \operatorname{Poi}(y; \lambda_{i_1 j_{i_1}} + \cdots + \lambda_{i_{N_r} j_{i_{N_r}}}) \big],
\end{aligned}
\tag{2.2}
$$

where y is a non-negative integer. There are in total $4N_r$ unknown parameters in the model above: $2N_r$ proportion parameters p_{k1} and p_{k2} satisfying $p_{k1} + p_{k2} = 1$

for $k = i_1, i_2, \ldots, i_{N_r}$, and $2N_r$ intensity rates λ_{k1} and λ_{k2} for $k = i_1, i_2, \ldots, i_{N_r}$. The intensity rates λ_{t1} may vary from transcript to transcript, but they mainly depend on the coverage uncertainty in the RNA-Seq experiment as discussed in the paragraph following (2.1), so we further assume that they are equal, that is, $\lambda_{i_1 1} = \cdots = \lambda_{i_{N_r} 1}$.

In the discussion above, we do not distinguish junction reads from nonjunction reads, even though junction reads demonstrate different characteristics. First, junction reads contain more information about the transcripts, from which they are generated, than nonjunction reads. Consider junction reads of type $r_{l_1 l_2} \in \mathcal{J}$ with $l_1 < l_2$ and nonjunction reads of type $r_{l_1 l_1} \in \mathcal{N}$. Let $\mathcal{T}^{r_{l_1 l_1}}$ be the collection of candidate transcripts that can generate reads of nonjunction type $r_{l_1 l_1}$, and $\mathcal{T}^{r_{l_1 l_2}}$ the collection of candidate transcripts that can generate reads of junction type $r_{l_1 l_2}$. It is clear that $\mathcal{T}^{r_{l_1 l_2}}$ is a subset of $\mathcal{T}^{r_{l_1 l_1}}$. Therefore, reads of type $r_{l_1 l_2}$ are generated from a smaller set of transcripts, and contain more direct information about the transcripts than reads of type $r_{l_1 l_1}$. For the same reason, junction reads are often used for assembling novel transcripts.

In current real RNA-seq data, however, we found that the number of junction reads is relatively smaller than the number of nonjunction reads. In other words, given a junction read type $r \in \mathcal{J}$, for $m \in E^r$, the number of positive y_m^r's is small. We postulate that the excessive large number of base pairs with $y_m^r = 0$, for $r \in \mathcal{J}$, is caused by other unknown missing mechanisms, which cannot be properly modeled. Therefore, when estimating the model parameters, it may not be appropriate to use the original distribution of Y_m^r, for $r \in \mathcal{J}$. One approach to solving this difficulty is to consider only positive counts $y_m^r > 0$ and the conditional distribution of y_m^r given $y_m^r > 0$. Another advantage of using the positive counts and their conditional distributions is to avoid the ambiguity caused by difference in fragment length in the definition of E^r for $r \in \mathcal{J}$. We define $y_+^r = \{y_m^r : m \in E_+^r\}$, where $E_+^r = \{m : y_m^r > 0\}$. For $r \in \mathcal{J}$, the conditional distribution of Y_m^r given $Y_m^r > 0$ for $m \in E^r$ is given as follows:

$$p(Y_m^r = y | Y_m^r > 0) = \left\{ \prod_{k=i_1}^{i_{N_r}} \left[\sum_{j_k=1}^{2} p_{k j_k} \operatorname{Poi}(y; \lambda_{k j_k}) \right] \right\} \Big/ \{1 - p(Y_m^r = 0)\}.$$

2.3. *Illustrative example*. We use a concrete example to illustrate exons, read types and the models discussed in the previous subsections. Based on the reference sequence database (refseq) of the human genome (hg18), gene DARC has two biological exons, which we denote as exons e_1' and e_2', respectively, and depict in Figure 1 of the supplementary material [Wu and Zhu (2016)]. There are two annotated transcripts or isoforms associated with DARC, which are labeled as NM002036 and NM001122951, respectively. NM002036 consists of e_1' and a part of e_2', while NM001122951 consists of the entire e_2'. To make the transcripts

either contain or skip an exon entirely, we split exon e_2' into two sub-exons denoted as e_2 and e_3. We re-denote exon e_1' as e_1. Therefore, exons e_1, e_2 and e_3 form a partition of the exonic region of gene DARC, NM002036 consists of e_1 and e_3, and NM001122951 consists of e_2 and e_3. We relabel the two transcripts NM002036 and NM001122951 as T_1 and T_2, respectively, and assume that they form the collection of candidate transcripts, that is, $\mathcal{T} = \{T_1, T_2\}$. We are interested in quantifying the expression levels of T_1 and T_2 based on the reads that are mapped to the annotated region of gene DARC.

Let E_g be the exonic region of gene DARC, that is, $E_g = e_1 \cup e_2 \cup e_3$. There are in total 2212 base pairs in E_g with 969 base pairs in e_1, 202 base pairs in e_2, and 1041 base pairs in e_3. We index the base pairs in E_g from the left end (5′-end) to the right end (3′-end) by $1, 2, \ldots, 2212$. Hence, e_1 contains base pairs 1 through 969, e_2 contains base pairs 970 through 1171, and e_3 contains base pairs 1172 through 2212. Transcript T_1 is capable of generating reads of three different types, which are r_{11}, r_{13} and r_{33}, respectively. Similarly, transcript T_2 is capable of generating reads of three types, which are r_{22}, r_{23} and r_{33}. Therefore, the collection of all possible read types is $\mathcal{R} = \{r_{11}, r_{22}, r_{33}, r_{13}, r_{23}\}$, and the collections of nonjunction types and junction types are $\mathcal{N} = \{r_{11}, r_{22}, r_{33}\}$ and $\mathcal{J} = \{r_{13}, r_{23}\}$, respectively. Thus, $E^{r_{11}} = \{1, \ldots, 969\}$, $E^{r_{13}} = \{1, \ldots, 969\}$, $E^{r_{33}} = \{1172, \ldots, 2212\}$, $E^{r_{22}} = \{970, \ldots, 1171\}$, and $E^{r_{23}} = \{970, \ldots, 1171\}$. Let $X_{tm}^{r_{uv}}$ represent the count of type r_{uv} reads generated from transcript T_t at base pair m for $r_{uv} \in \mathcal{R}$, $T_t \in \mathcal{T}$ and $m \in E^{r_{uv}}$. Due to limited space, we list only the distributions of $X_{1m}^{r_{33}}$ and $X_{2m}^{r_{33}}$ below, and the distributions of other $X_{tm}^{r_{uv}}$'s can be found in Section 3 of the supplementary material [Wu and Zhu (2016)]:

$$\text{(2.3)} \qquad \begin{cases} X_{1m}^{r_{33}} \sim p_{11}\,\mathrm{Poi}(\lambda_{11}) + p_{12}\,\mathrm{Poi}(\lambda_{12}), & \text{for } m \in E^{r_{33}}; \\ X_{2m}^{r_{33}} \sim p_{21}\,\mathrm{Poi}(\lambda_{21}) + p_{22}\,\mathrm{Poi}(\lambda_{22}), & \text{for } m \in E^{r_{33}}. \end{cases}$$

Here the intensity rates λ_{11} and λ_{21} account for the no-coverage or abnormally low coverage at the base pairs of transcripts T_1 and T_2, respectively. Note that $\lambda_{11} = \lambda_{21}$. The intensity rates λ_{12} and λ_{22} account for the abundance or the expression levels of transcripts T_1 and T_2, respectively.

As discussed previously, $X_{tm}^{r_{uv}}$'s are not always directly observable. Instead, we only observe the aggregated read counts $Y_m^{r_{uv}}$'s. The relationship between $Y_m^{r_{uv}}$ and $X_{tm}^{r_{uv}}$ and the distributions of $Y_m^{r_{uv}}$ can be obtained. For example, $Y_m^{r_{33}} = \sum_{i=1}^{2} X_{im}^{r_{33}}$, and the distribution of $Y_m^{r_{33}}$ is $\sum_{i=1}^{2}\sum_{j=1}^{2} p_{1i}\,p_{2j} \cdot \mathrm{Poi}(\lambda_{1i} + \lambda_{2j})$, $m \in E^{r_{33}}$. Notice that, in this example, both transcripts T_1 and T_2 can produce reads of type r_{33}. Therefore, $Y_m^{r_{33}}$ is the convolution of $X_{1m}^{r_{33}}$ and $X_{2m}^{r_{33}}$. The distributions of the other $Y_m^{r_{uv}}$'s can be found in Section 3 of the supplementary material [Wu and Zhu (2016)]. Given $\mathbf{y} = \{y_m^r : m \in E^r, r \in \mathcal{R}\}$, the maximum likelihood method can be used to estimate the mixing proportions and the intensity rates of the models.

2.4. *Composite likelihood function.* For gene g with k_g exons $e_1, \ldots, e_{k_g}$, the collection of candidate transcripts $\mathcal{T}_g = \{T_1, \ldots, T_N\}$, and the collection of possible read types $\mathcal{R} = \mathcal{N} \cup \mathcal{J}$, given the count data of all possible types $\mathbf{y} = \{y_m^r, m \in E^r, r \in \mathcal{R}\}$, the likelihood function for the model parameters can be derived as follows. We consider nonjunction read types first. For $r \in \mathcal{N}$, $y^r = \{y_m^r : m \in E^r\}$. Recall that $\mathcal{T}_g^r = \{T_{i_1} \cdots T_{i_{N_r}}\}$, which is a sub-collection of $\mathcal{T}_g$ including the transcripts that can produce reads of type r. Let $\theta^r = \{(p_{k1}, p_{k2}, \lambda_{k1}, \lambda_{k2}) : k \in \{i_1, \ldots, i_{N_r}\}\}$, which is the collection of the model parameters associated with the transcripts in $\mathcal{T}_g^r$. The probability mass function of y_m^r is given in model (2.2). Given y^r, the likelihood function of θ^r is

$$
\begin{aligned}
&L^r(\theta^r | y^r) \\
&= \prod_{m \in E^r} \prod_{k=i_1}^{i_{N_r}} \Bigg[\sum_{j_k=1}^{2} p_{k j_k} \operatorname{Poi}(\lambda_{k j_k}) \Bigg] \qquad (2.4) \\
&= \prod_{m \in E^r} \Bigg\{ \sum_{j_{i_1}}^{2} \cdots \sum_{j_{i_{N_r}}}^{2} [(p_{i_1 j_{i_1}} \cdots p_{i_{N_r} j_{i_{N_r}}}) \operatorname{Poi}(y_m^r; \lambda_{i_1 j_{i_1}} + \cdots + \lambda_{i_{N_r} j_{i_{N_r}}})] \Bigg\}.
\end{aligned}
$$

Next, we consider junction read types $r \in \mathcal{J}$. For $r \in \mathcal{J}$, $y^r = \{y_m^r : m \in E^r\}$. Due to the reasons discussed in Section 2.2, we only use the nonzero counts $y_m^r > 0$ and the conditional distribution of Y_m^r given $Y_m^r > 0$. We define $\mathcal{T}^r$ and θ^r in the same way as those for nonjunction types. Given the positive counts $y_+^r = \{y_m^r : m \in E_+^r\}$, the conditional likelihood function of θ^r is

$$
L_c^r(\theta^r | y_+^r) = \prod_{m \in E_+^r, \text{ s.t. } y_m^r > 0} p(y_m^r | \theta^r, y_m^r > 0), \qquad (2.5)
$$

where c in L_c^r indicates conditional likelihood. The detailed calculation of the conditional likelihood function can be found in Section 4 of the supplementary material [Wu and Zhu (2016)]. Assume that there are $|\mathcal{T}_g| = N$ candidate transcripts in $\mathcal{T}_g$, which are indexed by $1, 2, \ldots, N$. Let $\theta = \{(p_{i1}, p_{i2}, \lambda_{i1}, \lambda_{i2}) : 1 \le i \le N\}$, which is the collection of all model parameters. We assume read counts of different types are independent with each other. Let $\tilde{\mathbf{y}} = \{y^r : r \in \mathcal{N}\} \cup \{y_+^r : r \in \mathcal{J}\}$. Given $\tilde{\mathbf{y}}$, the likelihood function of θ can be obtained by combining the likelihood functions (2.4) and (2.5) as follows:

$$
L(\theta | \tilde{\mathbf{y}}) = \prod_{r \in \mathcal{N}} L^r(\theta^r | y^r) \cdot \prod_{r \in \mathcal{J}} L_c^r(\theta^r | y^r, y^r > 0).
$$

We refer to $L(\theta | \tilde{\mathbf{y}})$ as the composite likelihood function of θ, since it involves both the ordinary likelihood and conditional likelihood functions. The maximum composite likelihood estimate (MCLE) of θ is defined as the solution to the following maximization problem:

$$
\max_{\theta} L(\theta | \tilde{\mathbf{y}}),
$$

subject to

$$\begin{cases} \lambda_{ij} > 0, & \text{for } 1 \le i \le N \text{ and } 1 \le j \le 2; \\ \lambda_{11} = \cdots = \lambda_{N1}; & \\ p_{i1} + p_{i2} = 1, & \text{for } 1 \le i \le N. \end{cases}$$

The MCLE of θ is denoted as $\hat{\theta}$. The behavior and properties of general MCLEs have been studied by Varin et al. [Varin, Reid and Firth (2011)]. Under some regularity conditions, MCLEs are consistent, asymptotically normal and may, however, trade some efficiency for gain in convenience in modeling and computation. In the next subsection, we use the EM algorithm to compute $\hat{\theta}$.

2.5. *EM algorithm.* Directly optimizing the original composite likelihood function $L(\theta|\tilde{\mathbf{y}})$ is difficult and time consuming. Instead we apply the EM algorithm to calculate the MCLE $\hat{\theta}$. The EM algorithm uses a two-step data-generating scheme as follows. For type $r \in \mathcal{R}$ and $m \in E^r$, recall that Y^r_m follows a 2^{N_r}-component mixture of Poisson distribution, and we index the components by $i_1 j_{i_1} \cdots i_{N_r} j_{i_{N_r}}$, for $j_{i_k} \in \{1, 2\}$ and $k \in \{1, \ldots, N_r\}$. We define membership indicator variables $Z^r_{m,(i_1 j_{i_1} \cdots i_{N_r} j_{i_{N_r}})}$ such that $Z^r_{m,(i_1 j_{i_1} \cdots i_{N_r} j_{i_{N_r}})} = 1$ if Y^r_m is from the component of the convolution distribution with intensity $(\lambda_{i_1 j_{i_1}} + \cdots + \lambda_{i_{N_r} j_{i_{N_r}}})$; and $Z^r_{m,(i_1 j_{i_1} \cdots i_{N_r} j_{i_{N_r}})} = 0$, otherwise. Let $z^r = \{z^r_{m,(i_1 j_{i_1} \cdots i_{N_r} j_{i_{N_r}})}, m \in E^r\}$, for $r \in \mathcal{N}$ and $z^r_+ = \{z^r_{m,(i_1 j_{i_1} \cdots i_{N_r} j_{i_{N_r}})}, m \in E^r_+\}$ for $r \in \mathcal{J}$. Let $\tilde{\mathbf{z}} = \{z^r : r \in \mathcal{N}\} \cup \{z^r_+ : r \in \mathcal{J}\}$, which is the membership indicator of $\tilde{\mathbf{y}}$. With both $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{z}}$, the complete composite log-likelihood for θ can be written as

$$\begin{aligned} l(\theta|\tilde{\mathbf{y}}, \tilde{\mathbf{z}}) &= \log\bigl(L\bigl(\theta^r|\mathbf{y}, \mathbf{z}\bigr)\bigr) \\ &= \sum_{r \in \mathcal{N}} l^r\bigl(\theta^r|y^r, z^r\bigr) + \sum_{r \in \mathcal{J}} l^r_c\bigl(\theta^r|y^r, z^r, y^r > 0\bigr). \end{aligned}$$

For detailed information about the log-likelihood for nonjunction and junction reads, see Section 6 of the supplementary material [Wu and Zhu (2016)].

Suppose the current parameter estimate is $\hat{\theta}^{\text{cur}} = (\hat{\lambda}^{\text{cur}}, \hat{p}^{\text{cur}})'$. The E-step is to calculate the expected complete log-likelihood function $Q(\theta|\hat{\theta}^{\text{cur}}, \tilde{\mathbf{y}}) = E_{\tilde{\mathbf{z}}}[\log(l(\theta|\hat{\theta}^{\text{cur}}, \tilde{\mathbf{y}}, \tilde{\mathbf{z}}))]$, where the expectation is over the conditional distribution of $\tilde{\mathbf{z}}$ given $\hat{\lambda}^{\text{cur}}$, $\hat{p}^{\text{cur}}$ and $\tilde{\mathbf{y}}$. Notice that $Q(\theta|\hat{\theta}^{\text{cur}}, \tilde{\mathbf{y}})$ can also be written as $E[\sum_{r \in \mathcal{N}} l^r(\theta^r|y^r, \hat{\theta}^{\text{cur}}, z^r) + \sum_{r \in \mathcal{J}} l^r_c(\theta^r|y^r, \hat{\theta}^{\text{cur}}, z^r_+, y^r > 0)]$. The function Q consists of two parts, one of which involves the nonjunction types, and the other involves the junction types.

The M-step is to maximize Q with respect to λ and p, and the resulting maximizers can be used to update $\hat{\theta}^{\text{cur}} = (\hat{\lambda}^{\text{cur}}, \hat{p}^{\text{cur}})'$. We use a block coordinate descent algorithm to optimize Q. First, we fix the value of p at $\hat{p}^{\text{cur}}$ and maximize Q with respect to λ. The resulting maximizer is $(\hat{\lambda}_{11}, \hat{\lambda}_{12}, \ldots, \hat{\lambda}_{N2}) = \arg\max_{\lambda_{11}, \lambda_{12}, \ldots, \lambda_{N2}} Q$, and the current estimate of $\hat{\lambda}^{\text{cur}}$ is updated to be $(\hat{\lambda}_{11}, \hat{\lambda}_{12},$

$\ldots, \hat{\lambda}_{N2})$. Second, we fix the value of λ at $\hat{\lambda}^{\text{cur}}$, and optimize Q with respect to p_{t2} for $1 \le t \le N$. Let $\hat{p}_{t2} = \arg\max_{p_{t2}} Q$ and then $\hat{p}_{t1} = 1 - \hat{p}_{t2}$ for $1 \le t \le N$. Then current estimate $\hat{p}^{\text{cur}}$ is updated to be $(\hat{p}_{12}, \ldots, \hat{p}_{N2})$.

The EM algorithm iterates between the E-step and M-step until some convergence criterion is satisfied. It is worth pointing out that as the number of candidate transcripts increases, the computational complexity also increases. The EM algorithm for the CPM model suffers from the curse of dimensionality and the problem of local optima. To deal with the first problem, more sophisticated optimization algorithms or parallel computing techniques could be implemented. To deal with the second problem, we adopt the strategy of using multiple initializations. We repeat the EM algorithm with different initial values of the parameters and choose the estimates that achieve the largest likelihood value.

2.6. *Quantification method.* Suppose the MCLEs of the model parameters for transcript T_t are calculated to be $\hat{\lambda}_{t1}$, $\hat{\lambda}_{t2}$, $\hat{p}_{t1}$ and $\hat{p}_{t2}$. Following the quantification procedure proposed by Wu et al. [Wu, Qin and Zhu (2012)], the expression level of transcript T_t is quantified to be $g_t^s = (s\hat{\lambda}_{t1}\hat{p}_{t1} + \hat{\lambda}_{t2}\hat{p}_{t2})/(s\hat{p}_{t1} + \hat{p}_{t2})$, where s is a prespecified number between 0 and 1 and $\hat{\lambda}_{t1} < \hat{\lambda}_{t2}$. More discussions about the selection of the tuning parameter s can be found in Section 7 of the supplementary material [Wu and Zhu (2016)].

2.7. *Illustrative example* (*continued*). Wong's lab studied the transcriptome of a brain tissue in an RNA-Seq experiment [Au et al. (2010)] which generates 50 bp paired-end reads. A total of 313 reads are mapped to gene DARC, which contains two isoforms, as we discussed in Section 2.3. The annotated two transcripts of gene DARC are able to generate five types of reads. The frequency of each type of reads is summarized in Table 1, and the base level counts are plotted in Figure 2 of the supplementary material [Wu and Zhu (2016)].

Note that r_{23} and r_{13} are two possible junction reads. Since type r_{23} reads are not observed, they will not be included in the calculation. There are 41 type r_{13} reads, and the count of those reads will be included in the calculation. Therefore, in this example, $\mathcal{R} = \mathcal{N} \cup \mathcal{J} = \{r_{11}, r_{22}, r_{33}, r_{13}\}$ and the observed read counts are $\mathbf{y} = (y^{r_{11}}, y^{r_{22}}, y^{r_{33}}, y_+^{r_{13}})'$. The composite likelihood can be written as $L(\theta|\mathbf{y}) = \prod_{r_{uu}\in\mathcal{N}} L^{r_{uu}}(\theta^{r_{uu}}|y^{r_{uu}}) \cdot \prod_{r_{uv}\in\mathcal{J}} L_c^{r_{uv}}(\theta^{r_{uv}}|y^{r_{uv}})$, where $\lambda_{tj} > 0$ for $1 \le$

TABLE 1
Frequency table for all read types for DARC in illustrative example

Type	r_{11}	r_{13}	r_{22}	r_{23}	r_{33}
Counts	103	41	0	0	169

TABLE 2
Quantification results for gene DARC in illustrative example

Transcript	$\hat{\boldsymbol{\lambda}}$	$\hat{\boldsymbol{p}}$	CPM-Seq	Cufflinks
T_1	$\hat{\lambda}_{11} = 0.066$	$\hat{p}_{11} = 0.953$	0.319	2.577
	$\hat{\lambda}_{12} = 1.352$	$\hat{p}_{12} = 0.047$		
T_2	$\hat{\lambda}_{21} = 0.066$	$\hat{p}_{21} = 0.830$	0.033	≈ 0
	$\hat{\lambda}_{22} \approx 0.000$	$\hat{p}_{22} = 0.170$		

$t \le 2, 1 \le j \le 2$. Additionally, we have $\lambda_{11} = \lambda_{21}$, and $p_{t1} + p_{t2} = 1$ for $1 \le t \le 2$. Applying the EM algorithm, the MCLEs of the parameters are obtained and reported in Table 2.

The expression levels of T_1 and T_2 are quantified to be 0.319 and 0.033 by CPM-Seq, respectively. We also applied Cufflinks to quantify the expression levels of these two transcripts, and they are 2.577 and 0, respectively. From the counts of the observed types of reads in Table 1, we can see that there do not exist type r_{22} and r_{23} reads, suggesting the absence or the low expression level of transcript T_2. The large counts of type r_{11} and type r_{33} reads indicate the presence of transcript T_1. Therefore, in this example, the quantification results of CPM-Seq and Cufflinks are consistent with each other and with the observed read counts.

3. Results. To further compare CPM-Seq with Cufflinks and RSEM, two popularly used methods in practice, we need to have a gold standard as the benchmark. In our simulation study, we can simulate expression levels and treat them as the gold standard. In real data applications, however, the true expression levels of transcripts are not available, hence, we instead use the qRT-PCR measurements as the gold standard. We present the results of three simulation studies, which are Examples 1, 2 and 3, conducted at different scales. We further use two real datasets to compare our proposed method with Cufflinks and RSEM. The first dataset contains the single-end sequencing data and qRT-PCR measurements of eight transcripts. The second dataset contains the paired-end sequencing data of a brain sample. The comparison results based on these two datasets are presented in Examples 4 and 5, respectively.

3.1. *Simulation study.* We can generate RNA-Seq read counts in two possible ways. We can simulate the data from a prespecified parametric model or we can use an RNA-Seq simulator. To make our simulation study more convincing, we choose the Flux simulator to generate RNA-Seq short reads. The Flux simulator was developed by Gabriel et al. [Griebel et al. (2012)] to simulate RNA-Seq experiments *in silico* and is among the most sophisticated simulators. Given a set of transcripts and their expression levels, the Flux simulator simulates the protocols of an RNA-Seq experiment step-by-step to generate the short reads.

EXAMPLE 1. We conducted a small-scale simulation study to compare the performances of CPM-Seq and Cufflinks. Five genes were selected from chromosome one of the human genome. Each gene contained three annotated isoforms. Using the Flux simulator, 75 bp paired-end reads were generated for these 15 isoforms as follows. First, the simulator randomly assigned expression levels to all 15 isoforms in the annotation. Second, the simulator randomly fragmented these isoform molecules into small pieces, which were then amplified *in silico*. Third, the simulator sequenced these fragments and generated three thousand 75 bp paired-end reads. Once the reads were obtained, we mapped them back to the reference genome using Tophat [Trapnell, Pachter and Salzberg (2009)]. We converted the mapped reads to read count data of different types. Based on the count data of different types, we applied CPM-Seq and Cufflinks separately to quantify the expression levels of the 15 transcripts. We refer to the resulting measurements as the CPM-Seq measurements and Cufflinks measurements, respectively. The expression levels of the transcripts assigned by the simulator in the first step were treated as the gold standard.

The Pearson correlation coefficient between the CPM-Seq measurements and the gold standard is 0.715, and the Pearson correlation coefficient between the Cufflinks measurements and the gold standard is 0.665. The scatter plots of the CPM-Seq and Cufflinks measurements against the gold standard are given in Figure 3 in the supplementary material [Wu and Zhu (2016)]. We also calculated the Spearman rank correlation coefficient between CPM-Seq and the gold standard (0.871), and the Spearman rank correlation coefficient between Cufflinks and the gold standard (0.275). The scatter plots of the ranks of the CPM-Seq and Cufflinks measurements against those of the gold standard are given in Figure 1.

We can see that in terms of the Pearson correlation coefficient, CPM-Seq slightly outperforms Cufflinks. However, in terms of the Spearman rank corre-

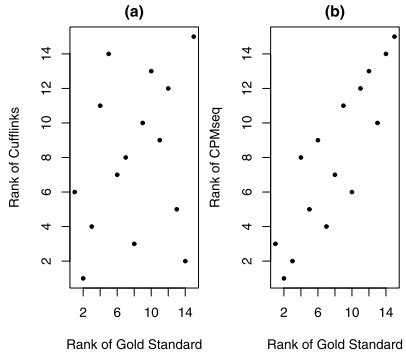

FIG. 1. *Scatter plots of ranks of quantification results by Cufflinks and CPM-Seq for* 15 *transcripts in Example* 1. *Plot* (a) *is for the ranks of Cufflinks measurements versus those of the gold standard* (*the Spearman rank correlation coefficient* = 0.275). *Plot* (b) *is for the ranks of the CPM-Seq measurements versus those of the gold standard* (*the Spearman rank correlation coefficient* = 0.871).

TABLE 3
Spearman rank correlation coefficients for RSEM, CPM-Seq and Cufflinks versus the gold standard

CPM-SEQ	Cufflinks	RSEM
0.604	0.387	0.589

lation coefficient, CPM-Seq outperforms Cufflinks dramatically. We believe that the Spearman rank correlation coefficient characterizes the performances of the two different methods much better than the Pearson correlation coefficient. The overall superior performance of CPM-Seq over Cufflinks is demonstrated by the strong linear pattern in Plot (b) of Figure 1 for CPM-Seq, and the lack of linear pattern in Plot (a) for Cufflinks. When comparing the 15 transcripts pairwise (105 pairs in total), CPM-Seq ranked 90 pairs correctly, whereas Cufflinks only ranked 63 pairs correctly. This example suggests that the Spearman rank correlation coefficient provides a more reliable measure of the performance of a quantification method, and thus we will use it in the other examples in the rest of the paper.

EXAMPLE 2. We conducted a medium scale simulation study to compare CPM-Seq with both Cufflinks and RSEM. We used the Flux Simulator and generated reads for 17 genes, each of which contains 3 transcripts. As in Example 1, generated reads were mapped back to the reference genome using Tophat. CPM-Seq, Cufflinks and RSEM were applied to quantify the transcripts' expression levels. CPM-Seq and RSEM achieved comparable Spearman rank correlation with the gold standard, which were 0.604 and 0.589, respectively. Cufflinks was only able to achieve a correlation of 0.387 with the gold standard. The performances of these three methods are summarized in Table 3 and Figure 2.

EXAMPLE 3. We also conducted a relatively large-scale simulation study with ten replicated runs. In each run, the Flux simulator randomly assigned expression levels to all isoforms of human chromosome 1 in refseq hg 18, and generated three million 75 bp paired-end reads. Once the reads were obtained, they were mapped back to the reference genome using Tophat. We eliminated genes that had less then 20 reads, and genes that received more than 60 reads at least at one base pair. In the first run, there were 987 genes left after filtering. Among these genes, 710 genes had single isoform, 156 genes had two isoforms, 74 genes had three isoforms, 27 genes had four isoforms, and 27 genes had five isoforms. We applied CPM-Seq and Cufflinks separately to quantify the expression levels of these isoforms and calculated their Spearman rank correlation coefficients with the gold standard. The Spearman rank correlation coefficient for CPM-Seq was 0.616 and the Spearman rank correlation coefficient for Cufflinks was 0.538. Therefore,

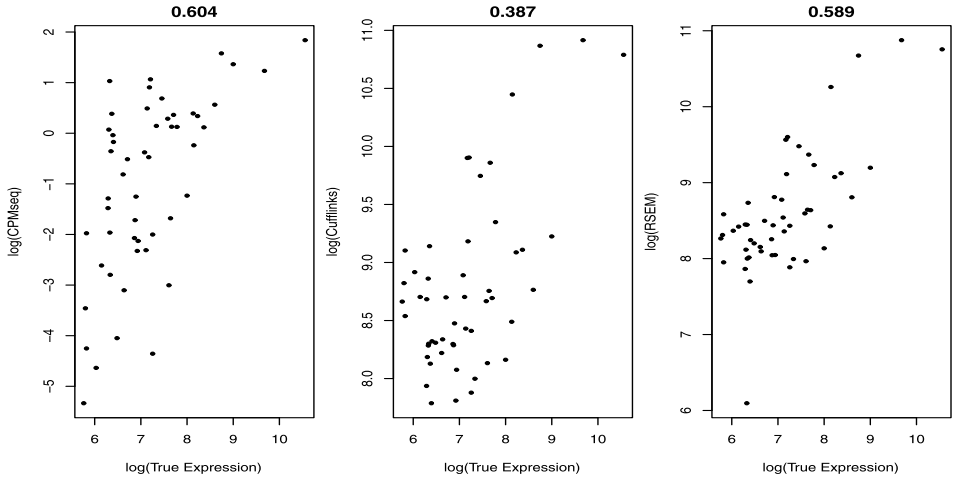

FIG. 2. *Scatter plots of quantification results by CPM-Seq, Cufflinks and RSEM with the gold standard in Example 2.*

CPM-Seq outperformed Cufflinks in this run. The simulation results of the other 9 runs are reported in Table 4. To compare the performances of CPM-Seq and Cufflinks in all ten runs, we applied the paired t-test, and the resulting p-value was ≤ 0.0001, suggesting a significant improvement of CPM-Seq over Cufflinks.

3.2. *Real data application.* Two real datasets are further used to compare CPM-Seq, Cufflinks and RSEM, and the corresponding results are presented in Examples 4 and 5 below. Example 4 is based on a small-scale study with qRT-PCR measurements which are used as the gold standard. Example 5 is based on

TABLE 4
Spearman rank correlation coefficients for CPM-Seq and Cufflinks of ten simulation runs

	CPM-Seq with gold	**Cufflinks with gold**	**CPM-Seq with Cufflinks**
Run 1	0.616	0.538	0.498
Run 2	0.551	0.540	0.436
Run 3	0.606	0.550	0.471
Run 4	0.612	0.555	0.464
Run 5	0.589	0.548	0.511
Run 6	0.547	0.527	0.464
Run 7	0.603	0.562	0.490
Run 8	0.610	0.532	0.449
Run 9	0.602	0.569	0.498
Run 10	0.594	0.585	0.536

the large-scale study that does not have qRT-PCR measurements. Therefore, we do not have a gold standard in Example 5. Instead, we use the characteristics of the read counts data themselves to facilitate the comparison of the two methods.

EXAMPLE 4. Two human cell lines named MCF7 and HME were studied by Wang et al. using RNA-Seq [Wang et al. (2008)]. The resulting data can be downloaded from the NCBI Short Read Archive at http://www.ncbi.nlm.nih.gov/sra under accession number GSE12946. There are 21.6 million 32 bp reads for the MCF7 cell line and 17.8 million 32 bp sequenced reads for the HME cell line. Using bowtie [Langmead et al. (2009)], we mapped the reads to the ucsc hg18 reference genome and obtained the base level read counts data for both cell lines. We applied CPM-Seq, Cufflinks and RSEM to quantify the transcript expression levels.

The original study of Wang et al. did not provide the qRT-PCR measurements of the transcripts. Fortunately, Kim et al. [Kim et al. (2011)] used the qRT-PCR techonology to measure eight transcripts of four genes of these two cell lines in a separate study. We used these qRT-PCR measurements as the gold standard to compare the performances of CPM-Seq, Cufflinks and RSEM. In Table 5, we report those eight transcripts and their qRT-PCR, CPM-Seq, Cufflinks and RSEM measurements for the HME cell line. The same transcripts and their qRT-PCR, CPM-Seq, Cufflinks and RSEM measurements for the MCF7 cell line are reported in Section 9 of the supplementary material [Wu and Zhu (2016)]. The Spearman rank correlation coefficients between the gold standard and the three quantification methods for both of the two cell lines are calculated and reported in Table 6. We can see that CPM-Seq achieves higher correlation with the gold standard than Cufflinks and RSEM in both cell lines.

TABLE 5
Quantification results of qRT-PCR, CPM-Seq and Cufflinks for eight transcripts in Example 4

	HME			
Transcript ID	**qRT-PCR**	**CPM-Seq**	**Cufflinks**	**RSEM**
uc002cvs.1	423.5	0.9771	1.3396	1.52
uc002cvt.2	234.7	1.4227	33.6959	40.05
uc002qlp.1	277.9	1.1251	7.8938	6.97
uc002qlq.1	621.8	1.3131	18.9976	19.51
uc002xmo.1	8.1	0.0019	0.1141	0.44
uc002xmn.1	10.7	0.0039	0.2860	0.29
uc003ngr.1	12.4	0.8350	14.3832	9.93
uc003ngs.1	538.2	0.0472	1.6722	2.01

TABLE 6
Spearman rank correlation coefficients of CPM-Seq and Cufflinks with the gold standard in Example 4

	CPM-Seq	Cufflinks	RSEM
HME	0.571	0.476	0.452
MCF7	0.619	−0.024	−0.024

EXAMPLE 5. In this example, we will see that even when CPM-Seq is concordant with Cufflinks and RSEM, their quantification results for some genes can be quite different from each other. We analyzed RNA-Seq data of the Human Brain Reference RNA sample, which was originally generated by Wong's lab using the Illumina Genome Analyzer platform [Au et al. (2010)]. We processed one lane of eight million 50 bp paired-end reads. The data set can be downloaded from the NCBI Short Read Archive (SRA) at http://www.ncbi.nlm.nih.gov/sra under the accession numbers GSM475204 and GSM475205 [Au et al. (2010)]. Tophat was used to map the reads to refseq hg18. We filtered out genes that had received in total less than 20 reads, genes that had received more than 60 reads at least at one base pair, and genes with more than five exons. After filtering, 433 genes on chromosome one remained and included 743 isoforms. Among the 433 genes, there were 277 single-isoform genes, 87 two-isoform genes, 51 three-isoform genes and 18 four-isoform genes. Because these multi-isoform genes contain many sub-exons, it is not possible to observe every type of junction read even if all of the junctions are expressed. Therefore, we used the composite likelihood, which includes all non-junction reads and the positive junction reads. We applied CPM-Seq, Cufflinks and RSEM to quantify the expression level of each transcript. The Spearman rank correlation coefficient between CPM-Seq and Cufflinks was 0.589, and the Spearman rank correlation coefficient between CPM-Seq and RSEM was 0.590. Therefore, CPM-Seq, Cufflinks and RSEM show a good overall concordance in this example.

Despite their general concordance, the quantification results of CPM-Seq differ from those of Cufflinks and RSEM for a large number of genes. A more careful comparison between the CPM-Seq, Cufflinks and RSEM measurements of these genes indicates that the CPM-Seq measurements are more reasonable. We give such an example below.

According to human refseq hg18, gene ZNF238 contains two exons, which we denote as e_1' and e_2', and it has two annotated transcripts labeled as NM205768 and NM006352. Transcript NM205768 consists of e_1' and a part of e_2', and transcript NM006352 consists of the entire e_2'. To make the transcripts either contain or skip an exon entirely, we split exon e_2' into two sub-exons denoted as e_2 and e_3. We re-denote exon e_1' as e_1. Therefore, exons e_1, e_2 and e_3 form a partition of the exonic region of gene ZNF238. The total length of gene ZNF238's exonic region is 4387, and the exonic base pairs are indexed as $1, \ldots, 4387$. Exons e_1, e_2 and e_3 contain

TABLE 7
Frequency table for each type of read for gene ZNF238 in Example 5

Type	r_{11}	r_{13}	r_{22}	r_{23}	r_{33}
Counts	2	8	4	2	1313

base pairs $\{1, \ldots, 187\}$, $\{188, \ldots, 695\}$, $\{696, \ldots, 4387\}$, respectively. NM205768 consists of e_1 and e_3, and NM006352 consists of e_2 and e_3. We relabel the two transcripts NM205768 and NM006352 as T_1 and T_2, respectively, and assume that they form the collection of candidate transcripts, that is, $\mathcal{T} = \{T_1, T_2\}$.

As discussed previously, not all junction reads will be observed due to insufficient coverage or technological limitations of the RNA-Seq experiment. In the 50 bp paired-end reads data generated by Wong's lab, we observed only five types of reads for gene ZNF238. The frequency of each type of read is summarized in Table 7, and the base level counts are plotted in Figure 3.

We applied CPM-Seq to quantify the expression levels of T_1 and T_2, and the MCLEs of the model parameters and quantification results are reported in Table 8.

The parameter estimates indicate that both T_1 and T_2 are expressed and the expression level of T_1 (0.741) is about 1.5-fold higher than that of T_2 (0.296). We also applied Cufflinks and RSEM to quantify these two transcripts, and the quantification results are also presented in Table 8. Cufflinks quantified the expression level of T_1 to be 11.653, and quantified the expression level of T_2 to be 4.128e-05, which is almost zero. It appears that Cufflinks suggests that T_1 was expressed but T_2 was not. RSEM quantified the expression level of T_1 to be 233.61, and quantified the expression level of T_2 to be 23.36. According to the RSEM measurements,

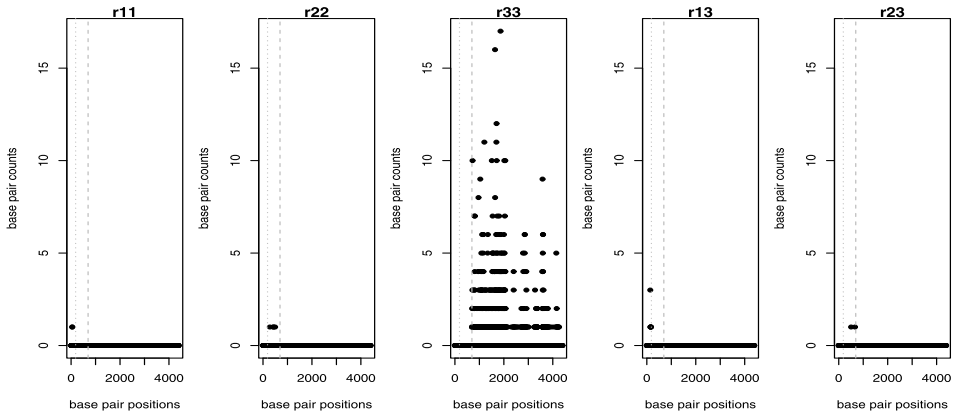

FIG. 3. *Base pair level counts for each type of read of ZNF238 in Example 5. The x-axis indicates base pair positions, and the y-axis represents the counts.*

TABLE 8
Quantification results of CPM-Seq and Cufflinks for T_1 and T_2 in Example 5

Transcript	$\hat{\boldsymbol{\lambda}}$	$\hat{\boldsymbol{p}}$	CPM-Seq	Cufflinks	RSEM
T_1	$\hat{\lambda}_{11} = 0.040$	$\hat{p}_{21} = 0.898$	0.741	11.653	233.61
	$\hat{\lambda}_{12} = 1.978$	$\hat{p}_{12} = 0.102$			
T_2	$\hat{\lambda}_{21} = 0.040$	$\hat{p}_{21} = 0.993$	0.296	≈ 0 (4.128e–05)	23.36
	$\hat{\lambda}_{22} = 7.998$	$\hat{p}_{22} = 0.007$			

T_2 was also expressed, which is consistent with CPM-Seq; however, the expression level of T_1 is 9-fold higher than that of T_1, which is not completely consistent with CPM-Seq.

Recall that $T_1 = e_1e_3$, $T_2 = e_2e_3$, and e_1, e_2 and e_3 are 187 bp, 507 bp and 3691 bp long in length, respectively. It is clear e_3 is the longest among the three exons, followed by e_2 and then e_1, and e_3 is actually much longer than e_2 and e_1. From Table 7, e_3 received the majority of the reads mapped to gene ZNF238 (1313 out of 1329 reads). These reads are of type r_{33}. Because both T_1 and T_2 contain e_3 and both transcripts can give rise to reads of type r_{33}, they cannot be directly allocated to the transcripts. Reads of types r_{11} and r_{13} (10 in total) suggest the expression of T_1, whereas reads of type r_{22} and r_{23} (6 in total) suggest the expression of T_2. The counts of these types of reads are relatively small compared to the count of reads of type r_{33}, due to the short lengths of e_1 and e_2. We believe that this was the reason that Cufflinks was not able to separate the two transcripts and instead allocated all reads of type r_{33} to T_1. RSEM was able to separate the two transcripts, however, its ratio of separation is not consistent with the ratio of counts of reads that are not of the r_{33} type. On the other hand, CPM-Seq was able to infer the expression levels of T_1 and T_2 using the convolution model that models the read count at each base pair. We believe that CPM-Seq successfully identified and properly quantified the two transcripts in this example.

4. Discussion. As discussed in the Introduction, it is an indirect inference problem to identify transcripts and quantify their expression levels using RNA-Seq data, and various types of observation units proposed in the literature such as exons, segments and bins can make transcripts and their expression levels nonidentifiable in many cases. In this article, we propose a novel approach called CPM-Seq for quantifying transcript expression levels, which treats individual base pairs as observation units and uses the convolution of a mixture Poisson distribution to model the RNA-Seq data. Both a simulation study and real data application have demonstrated the effectiveness of CPM-Seq, and further shown that CPM-Seq is capable of producing more accurate and consistent quantification results than Cufflinks and RSEM.

There are several immediate directions to further improve CPM-Seq. First, more efficient computational algorithms may substantially accelerate CPM-Seq. Our current R implementation ran for 1 minute, 10 hours and 30 minutes for Examples 1, 2 and 3, respectively, on an Intel i5 2.5 GHz Processor with four cores. We observed that CPM-Seq takes a relatively longer time to handle genes with a large number of transcripts. The computation is more challenging with more candidate transcripts in the annotation. Because CPM-Seq quantifies the transcripts on a gene by gene basis, an immediate idea to accelerate CPM-Seq is to parallelize the current algorithm using modern distributed and parallel computing systems such as Spark [Zaharia et al. (2010)].

The second direction to improve CPM-Seq is to incorporate the fragment length distribution into the CPM model. In the literature, the fragment length distribution is typically modeled by $N(\mu, \sigma^2)$, where μ and σ^2 are either known or can be estimated from reads mapped to genes with a single exon. When considering reads of different types, we also need to include the possible lengths of the reads. For example, for read type r, instead of simply counting the reads of type r, we need to count the reads of type r and length l.

The third direction to improve CPM-Seq is to incorporate the lasso type of penalty into CPM-Seq. Currently, CPM-Seq can handle collections of candidate transcripts of moderate size. When the collection of candidate transcripts becomes large, both the computational and estimation efficiencies of CPM-Seq will be compromised. Under the sparsity assumption, incorporating the lasso-type penalty into CPM-Seq is expected to improve both the estimation and computational efficiencies of CPM-Seq. We have been working on all those three directions and the results are expected to be reported in a future publication.

Acknowledgments. We thank the Associate Editor and the reviewers for their constructive comments.

SUPPLEMENTARY MATERIAL

Supplementary document for deconvolution of base pair level RNA-Seq read counts for quantification of transcript expression levels (DOI: 10.1214/16-AOAS906SUPP; .pdf). We provide a supplementary document to show the details of the Poisson mixture distribution, the conditional distribution of y_m^r, the distribution of the illustrative example, the composite likelihood function, the details of the EM algorithm, the quantification method, supporting figures for the illustrative example, quantification results for MCF7, and the supporting figure for Example 1.

REFERENCES

AU, K. F., JIANG, H., LIN, L., XING, Y. and WONG, W. H. (2010). Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res.* **38** 4570–4578.

GRIEBEL, T., ZACHER, B., RIBECA, P., RAINERI, E., LACROIX, V., GUIGÓ, R. and SAMMETH, M. (2012). Modelling and simulating generic RNA-seq experiments with the flux simulator. *Nucleic Acids Res*. **40** 10073–10083.

HU, M., ZHU, Y., TAYLOR, J. M. G., LIU, J. S. and QIN, Z. S. (2012). Using Poisson mixed-effects model to quantify transcript-level gene expression in RNA-seq. *Bioinformatics* **28** 63–68.

KIM, H., BI, Y., PAL, S., GUPTA, R. and DAVULURI, R. V. (2011). IsoformEx: Isoform level gene expression estimation using weighted non-negative least squares from mRNA-seq data. *BMC Bioinformatics* **12** 305.

LANGMEAD, B., TRAPNELL, C., POP, M. and SALZBERG, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. **10** R25.

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.

LI, W., FENG, J. and JIANG, T. (2011). IsoLasso: A LASSO regression approach to RNA-Seq based transcriptome assembly. *J. Comput. Biol*. **18** 1693–1707. MR2860071

LI, J. J., JIANG, C.-R., BROWN, J. B., HUANG, H. and BICKEL, P. J. (2011). Sparse linear modeling of next-generation mRNA sequencing (RNA-seq) data for isoform discovery and abundance estimation. *Proc. Natl. Acad. Sci. USA* **108** 19867–19872.

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.

SALZMAN, J., JIANG, H. and WONG, W. H. (2011). Statistical modeling of RNA-Seq data. *Statist. Sci*. **26** 62–83. MR2849910

SRIVASTAVA, S. and CHEN, L. (2010). A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Res*. **38** e170.

TRAPNELL, C., PACHTER, L. and SALZBERG, S. L. (2009). TopHat: Discovering splice junctions with RNA-seq. *Bioinformatics* **25** 1105–1111.

TRAPNELL, C., WILLIAMS, B. A., PERTEA, G., MORTAZAVI, A., KWAN, G., VAN BAREN, M. J., SALZBERG, S. L., WOLD, B. J. and PACHTER, L. (2010). Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol*. **28** 511–515.

VARIN, C., REID, N. and FIRTH, D. (2011). An overview of composite likelihood methods. *Statist. Sinica* **21** 5–42. MR2796852

WANG, E. T., SANDBERG, R., LUO, S., KHREBTUKOVA, I., ZHANG, L., MAYR, C., KINGSMORE, S. F., SCHROTH, G. P. and BURGE, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* **456** 470–476.

WU, H., QIN, Z. S. and ZHU, Y. (2012). PM-Seq: Using finite Poisson mixture models for RNA-seq data analysis and transcript expression level quantification. *Statistics in Biosciences* **5** 71–87.

WU, H. and ZHU, Y. (2016). Supplement to "Deconvolution of base pair level RNA-seq read counts for quantification of transcript expression levels." DOI:10.1214/16-AOAS906SUPP.

ZAHARIA, M., CHOWDHURY, M., FRANKLIN, M. J., SHENKER, S. and STOICA, I. (2010). Spark: Cluster computing with working sets. In *HotCloud'*10 *Proceedings of the* 2*Nd USENIX Conference on Hot Topics in Cloud Computing* 10–10. USENIX Association, Berkeley, CA.

DEPARTMENT OF STATISTICS
PURDUE UNIVERSITY
WEST LAFAYETTE, INDIANA 47906
USA
E-MAIL: wu79@purdue.edu

DEPARTMENT OF STATISTICS
PURDUE UNIVERSITY
WEST LAFAYETTE, INDIANA 47906
USA
AND
CENTER FOR STATISTICAL SCIENCE
DEPARTMENT OF INDUSTRIAL ENGINEERING
TSINGHUA UNIVERSITY
BEIJING
CHINA
E-MAIL: yuzhu@purdue.edu