A Robust Method for Transcript Quantification with RNA-seq Data

Yan Huang[1], Yin Hu[1], Corbin D. Jones[2], James N. MacLeod[3], Derek Y. Chiang[4], Yufeng Liu[5], Jan F. Prins[6], and Jinze Liu[1] ⋆

[1]Department of Computer Science, [3]Department of Veterinary Science, University of Kentucky.
[2]Department of Biology, [4]Department of Genetics, [5]Department of Statistics and Operations Research, [6]Department of Computer Science, University of North Carolina at Chapel Hill.
{yan,yin}@netlab.uky.edu,{cdjones}@email.unc.edu,{jnmacleod}@uky.edu,{chiang}@med.unc.edu
{yfliu}@email.unc.edu,{prins}@cs.unc.edu,{liuj}@netlab.uky.edu

Abstract. The advent of high throughput RNA-seq technology allows deep sampling of the transcriptome, making it possible to characterize both the diversity and the abundance of transcript isoforms. Accurate abundance estimation or *transcript quantification* of isoforms is critical for downstream differential analysis (e.g. healthy vs. diseased cells), but remains a challenging problem for several reasons. First, while various types of algorithms have been developed for abundance estimation, short reads often do not uniquely identify the transcript isoforms from which they were sampled. As a result, the quantification problem may not be identifiable, i.e. lacks a unique transcript solution even if the read maps uniquely to the reference genome. In this paper, we develop a general linear model for transcript quantification that leverages reads spanning multiple splice junctions to ameliorate identifiability. Second, RNA-seq reads sampled from the transcriptome exhibit unknown position-specific and sequence-specific biases. We extend our method to simultaneously learn bias parameters during transcript quantification to improve accuracy. Third, transcript quantification is often provided with a candidate set of isoforms, not all of which are likely to be significantly expressed in a given tissue type or condition. By resolving the linear system with LASSO our approach can infer an accurate set of dominantly expressed transcripts while existing methods tend to assign positive expression to every candidate isoform. Using simulated RNA-seq datasets, our method demonstrated better quantification accuracy than existing methods. The application of our method on real data experimentally demonstrated that transcript quantification is effective for differential analysis of transcriptomes.

Keywords: Transcript quantification, Transcriptome, RNA-seq

1 Introduction

Recent studies have estimated that as many as 95% of all multi-exon genes are alternatively spliced, resulting in more than one transcript per gene [23, 33]. *Transcript quantification* determines the steady state levels of alternative transcripts within a sample, enabling the detection of differences in the expression of alternative transcripts under different conditions. Its application in detecting biomarkers between diseased and normal tissues can greatly impact biomedical research.

High-throughput sequencing technology (e.g. RNA-seq with Illumina, ABI Solid, etc.) provides deep sampling of the mRNA transcriptome. It allows the parallel sequencing of large number of mRNA molecules, generating tens of millions of short reads with lengthes up to 100bp at one end or both ends of mRNA fragments. Recent studies using RNA-seq have significantly expanded our knowledge on both the variety and the abundance of alternative splicing events [7, 36].

However, transcript quantification remains a challenging problem. First, it is commonly observed that "the more the isoforms, the harder to predict" [19]. Intuitively, transcript isoforms from the same gene often overlap significantly and a short read may be mapped to more than one transcript isoform. Determining the expression of individual transcripts from short read alignment, therefore, can lead to an *unidentifiable* model, where no unique solution exists. Secondly, transcript quantification often takes the candidate set of transcript isoforms, either from annotation databases such as Ensembl [2] and Refseq [3], or inferred from the splice graph using programs like Scripture [10], IsoInfer [8], IsoLasso [19], or Cufflinks [31]. It

⋆ To whom correspondence should be addressed.

is biologically unlikely to expect all candidate transcripts for a given gene to be significantly expressed concurrently in a cell. However, existing analytical approaches tend to assign positive expression values to every candidate transcript provided, thereby creating a situation in which large errors in abundance estimation can be computationally introduced for transcript isoforms that may, in reality, barely be expressed. An improved transcript quantification method, therefore, would determine or logically infer the subset of expressed transcript isoforms. Finally, various sampling biases have been observed regularly in RNA-seq datasets as a result of library preparation protocols. These biases typically include position-specific bias [6, 17, 25, 37] such as 3' bias and transcription start and end biases, and sequence-specific bias [18, 25, 32], where the read sampling in the transcriptome favors certain subsequences. How to compensate for these biases during transcript quantification is an open problem.

Transcript isoforms can differ not only in exons alternatively included or excluded but also in where two or more exons are connected together. In RNA-seq data, this information typically is implied by the spliced reads, i.e., the reads that cross one or more splice junctions. We have developed a general linear model for transcript quantification that leverages discriminative features in spliced reads to ameliorate the issue of identifiability and to simultaneously correct the sampling bias. Our contribution in this paper is three-fold: (1) We explicitly identify *MultiSplice*, a novel structural feature consisting of a contiguous set of exons that are expected to be spanned by the RNA-seq reads or transcript fragments of a given length. The MultiSplice, which includes single splice junctions as a special case, is used in two ways: its presence in the sample will infer the host transcript while its absence may reject it. MultiSplices are more powerful than single exons in disambiguating transcript isoforms, making more transcript quantification problems identifiable with long or paired-end reads; (2) We set up a linear system which minimizes the summed mean squared errors between the expected expression and the observed expression across all structure features along a gene while taking into account various bias effects; (3) We develop an iterative minimization algorithm in combination with LASSO [30] to resolve the aforementioned linear system in order to achieve the most accurate set of dominantly expressed transcripts while simultaneously correcting biases.

We have demonstrated the efficacy of our methods on both simulated RNA-seq datasets and real RNA-seq data: (1) We conducted the first study to investigate the question: what is the maximum read length needed in order to disambiguate all possible transcript isoforms in transcriptomes from different species; (2) We compared the proposed method with several state-of-the-art methods including Cufflinks, the Poisson model, and the ExonOnly model. Our results using simulated data from the human mRNA transcriptome demonstrated superior performance of the proposed method in most cases. When applied to 8 RNA-seq datasets from two breast cancer cell lines (MCF-7 and SUM-102), the quantification obtained from MultiSplice demonstrated good consistency within technical replicates from each transcriptome-wide assessment and substantial differences between the two biological groups (cell lines) in a small percentage of genes.

2 Related work

Various transcript quantification algorithms have been published recently. These methods can be divided mainly into two categories: read-centric and exon-centric. The representative methods using read-centric approaches include but are not limited to Cufflinks [31], IsoEM [22], and RSEM [17]. The central idea with read-centric approaches is to assign probability for each fragment to one transcript by maximizing the joint likelihood of read alignments based on the distribution of transcript fragments, and thereby estimating the transcript expression. When it is impossible to precisely allocate a fragment to a unique transcript, Cufflinks, for example, simply disregards or randomly assigns the read, causing information loss or inaccurate quantification. The second strategy, called exon-centric, considers the read abundance on an exonic segment as the cumulative abundance of all transcript isoforms. Methods in this category represent the transcript as a combination of exons and aim at estimating individual transcript abundance from the observed read counts or read coverage at each exon. The representative models in this category include the Poisson model [13, 24, 29] and linear regression approaches, such as rQuant [6], IsoLasso [19] and SLIDE [20].

Transcript abundance estimations can be unidentifiable, where no unique quantification exists. Both exon-centric and read-centric models may suffer from this problem. The paper by Lacroix *et al.* is one of the theoretical studies that have considered the identifiability problem of transcript quantification [16].

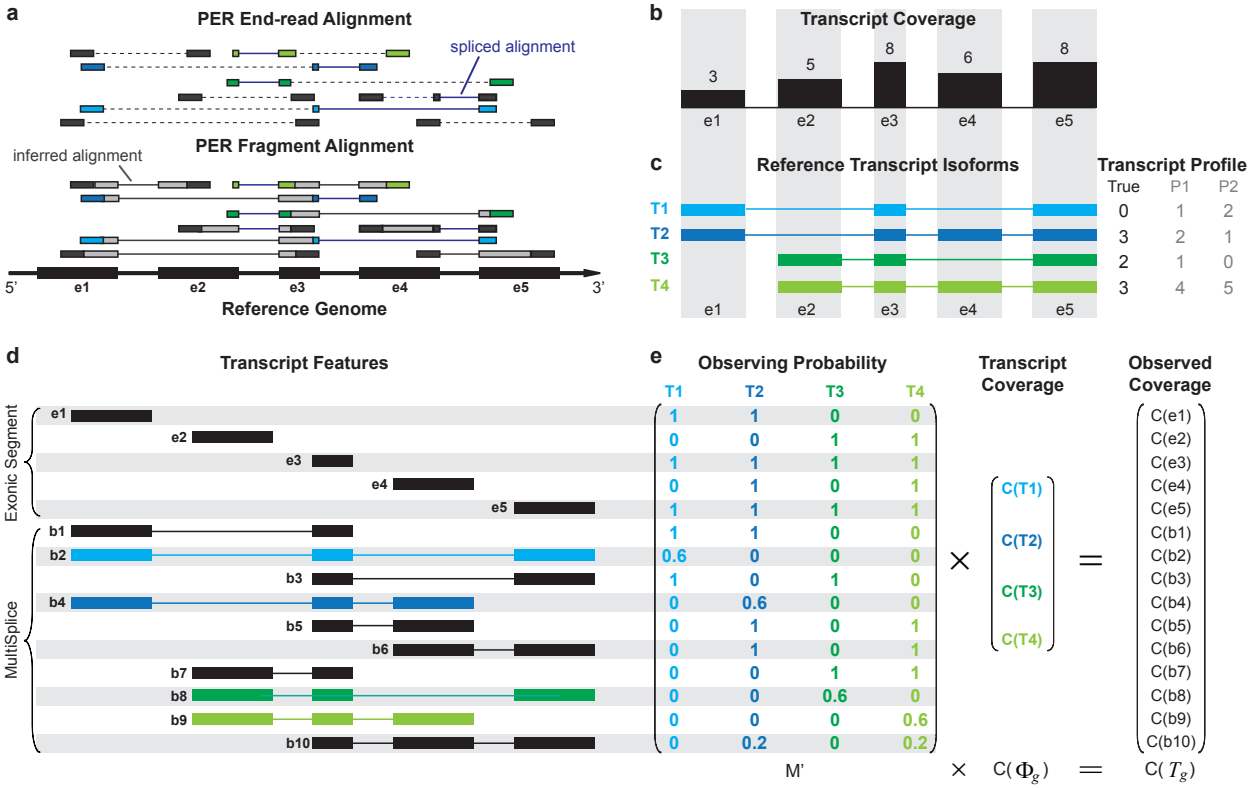


Fig. 1: Overview of the MultiSplice model. **a.** Sequenced RNA-seq short-reads are first mapped to the reference genome using an RNA-seq read aligner such as MapSplice [34]. In the presence of paired-end reads, MapPER [12] can be applied to find *PER fragment alignments* for the entire transcript fragment based on the distribution of insert size. **b.** Observed coverage on each exonic segment. **c.** Four transcripts originate from the alternative start and exon skipping events. Provided with these transcripts, abundance estimates would be unidentifiable for methods that only use coverage on exonic segments. Both transcript profiles P_1 and P_2, for instance, can explain the observed read coverage on each exon, but deviate from the true transcript expression profile. **d.** MultiSplices that connect multiple exonic segments in a transcript. **e.** A linear model can be set up where the expected coverage on every exonic or MultiSplice feature approximates its observed coverage. The transcript expression is solved as the one that minimizes the sum of squared relative error.

3 Method

In this section, we propose a method designated *MultiSplice*, for mRNA isoform quantification. We first define the observed features used in the MultiSplice model and the statistics collected. Then, we derive a general linear model to relate transcript level estimates to the observed expression on every feature.

Preliminaries. For a gene g, we use $\mathcal{E}_g$ to denote the set of exonic segments [13, 19] in g, which are disjoint genomic intervals on the genome that can be included in a transcript in its entirety. We use $\mathcal{T}_g$ to denote the set of mRNA isoforms transcribed from g. These mRNAs can be a set of annotated transcripts retrieved from a database such as Ensembl [2] or Refseq [3]. A transcript $t \in \mathcal{T}_g$ is defined by a sequence of exon segments, $t = e_1^t e_2^t \cdots e_{n_t}^t$, where $e \in \mathcal{E}_g$ and n_t denotes the number of exonic segments in the transcript t. The length of each exonic segment e is defined as the number of nucleotides in the exonic segment, denoted as $l(e)$. Hence, the length for every transcript is $l(t) = \sum_{i=1}^{n_t} l(e_i^t)$.

3.1 MultiSplice

In a typical RNA-seq dataset, a significant percentage of the read alignments are spliced alignments that connect more than one exon. With paired-end reads, the transcript fragment where its two ends are sampled can be inferred based on the distribution of the insert size [25]. Transcript fragments are typically between 200bp and 300bp, making them more likely to cross multiple exons, indicating these exons are present

together in one transcript. This information can be crucial in distinguishing alternative transcript isoforms. However, they are often ignored in current computational approaches.

In this subsection, we consider a sequence of adjacent exons in an mRNA transcript covered by transcript fragments. These structural features are the basis of *MultiSplice*. For generality, we assume that the RNA-seq reads are sampled from transcript fragments whose lengths follow a given distribution F_{fr} with probability density function f_{fr}. For example, the fragment length distribution F_{fr} is often modeled as a normal distribution with mean and variance learned from the genomic alignment of the RNA-seq reads. We also assume the maximum fragment length is l_{fr}.

Definition 1. *Let $b = e_i^t e_{i+1}^t \cdots e_{i+n_b}^t$ be a substring of a transcript sequence $t = e_1^t e_2^t \cdots e_{n_t}^t$, $n_b \geq 1$ and $i + n_b \leq n_t$. Then b is a MultiSplice in t if and only if*

$$\sum_{q=1}^{n_b-1} l(e_{i+q}) \leq l_{fr} - 2. \tag{1}$$

The condition in Equation 1 guarantees that a MultiSplice b connects $n_b + 1$ adjacent exons with at least 1 base landed on the left most exon e_i^t and the right most exon $e_{i+n_b}^t$. We use $\mathcal{B}_G$ to denote the set of all MultiSplices in gene G. From the definition, the set of MultiSplices vary according to the fragment length l_{fr}. The longer the fragments, the more MultiSplices are expected to function as structural features, and the higher power in disentangling highly similar alternative isoforms.

In Figure 1, for example, assume the maximum fragment length is $l_{fr} = 300bp$ with the expected fragment length of $250bp$ and the exonic segments of this gene have lengths of $l(e_1) = 200bp, l(e_2) = 200bp, l(e_3) = 100bp, l(e_4) = 200bp, l(e_5) = 200bp$. In reference transcript $T_1 = e_1e_3e_5$, $b_2 = e_1e_3e_5$ is a substring of T_1, and we have $l(e_3) = 100bp < 300bp = l_{fr}$ which allows a fragment to cover b_2. Therefore, b_2 is a MultiSplice feature of the gene. Combining MultiSplices from all the reference transcripts, b_1, b_3, b_5, b_6, and b_7 are MultiSplices consisting of a single splice junction, b_2, b_4, b_8, b_9, and b_{10} are MultiSplices consisting of two splice junctions.

3.2 Expected coverage and observed coverage

Given the gene g and a transcript $t \in \mathcal{T}_g$, let c_i be the number of transcript fragments covering the ith nucleotide of t. We define the coverage on t as the averaged number of transcripts covering each base in the transcript, $C(t) = \frac{1}{l(t)} \sum_{i=1}^{l(t)} c_i$. Then $C(t)$ is an estimator for the quantity of t in the sample, which provides a direct measure for the expression level of t. In our model, $C(t)$ is the unknown variable. The feature space that can be observed from the given RNA-seq sample is the union of all exonic segments and MultiSplices of the gene, $\Phi_g = \mathcal{E}_g \cup \mathcal{B}_g$. We aim at resolving the transcript expressions that minimize the difference between the observed expression and the expected expression of every feature.

For every $\phi \in \Phi_g$ and every transcript $t \in \mathcal{T}_g$, the expected coverage of feature ϕ from t can be expressed as a function of the transcript quantity $C(t)$, i.e., $E[C(\phi|t)] = m(\phi, t)C(t)$, where $m(\phi, t)$ contains the probability of observing ϕ in t assuming uniform sampling. Next, we define the *expected* coverage on exonic segments and MultiSplice respectively.

For an exonic segment e in t, assuming N_t fragments were sampled from t, the number of fragments falling in e then follows a binomial distribution with parameters N_t and $p_{e|t}$, $N_{e|t} \sim Bin(N_t, p_{e|t})$. When N_t is sufficiently large, the binomial distribution is well approximated using a normal distribution with mean $N_t p_{e|t}$ and variance $N_t p_{e|t}(1-p_{e|t})$, $N_{e|t} \dot{\sim} N(N_t p_{e|t}, N_t p_{e|t}(1-p_{e|t}))$. In expectation, $\frac{N_{e|t} l_{fr}}{l(e)}$ calculates the fragment coverage on e contributed by t, $C_{e|t}$, and $\frac{N_t l_{fr}}{l(t)}$ calculates the transcript coverage of t, C_t. Therefore, we have $\frac{N_{e|t} l_{fr}}{l_e} \dot{\sim} N(\frac{N_t p_{e|t} l_{fr}}{l_e}, \frac{r^2}{l_e^2} N_t p_{e|t}(1 - p_{e|t}))$ or $C_{e|t} \dot{\sim} N(C_t, \frac{r(l_t - l_e)C_t}{l_t l_e})$, and hence the expectation of observed coverage on e contributed by t equals the coverage of t, with $m(e, t) = 1$.

For a MultiSplice $b = e_i^t e_{i+1}^t \cdots e_{i+n_b}^t$, we are interested in the number of fragments containing it. Should a transcript fragment f_t cover b, f_t must start no later than the 3' end boundary of the leftmost exonic segment e_i^t and have at least 1 base landed on the rightmost exonic segment $e_{i+n_b}^t$. Therefore, there exists a window $w(b)$ before the 3' end of e_i^t with length $l(w(b)) = l(f_t) - \sum_{q=1}^{n_b-1} l(e_{i+q}) - 1$, where b can be covered by the

transcript fragment f_t. The probability that f_t covers b is hence $p_{b|t,l(f_t)} = \frac{l(f_t)-\sum_{q=1}^{n_b-1} l(e_{i+q})-1}{l(t)}$. Because $l(f_t)$ follows the fragment length distribution F, the expectation of $p_{b|t,l(f_t)}$ is then $p_{b|t} = E[p_{b|t,l(f_t)}] = \int p_{b|t,x} \cdot f(x)\,dx$, for x is the domain where the density function f is defined, resulting in $p_{b|t} = \frac{E[l(f_t)]-\sum_{q=1}^{n_b-1} l(e_{i+q})-1}{l(t)}$. Accordingly, the expected coverage of MultiSplice b gained from t is $E[C(b|t)] = m(b,t)C(t)$ if $m(b,t) = \frac{p_{b|t}l(t)}{E[l(f_t)]}$. In Figure 1, $m(b_2,T_1) = \frac{E[l(f_t)]-l(e_3)-1}{E[l(f_t)]} = \frac{250-100-1}{250} = 0.6$.

In summary, the probability that a MultiSplice feature ϕ contained in a transcript fragment f_t uniformly sampled from transcript t is:

$$m(\phi,t) = \begin{cases} 1 & \text{if } \phi \subset t \text{ and } \phi \in \mathcal{E}_G \\ \frac{E[l(f_t)]-\sum_{q=1}^{n_b-1} l(e_{i+q})-1}{E[l(f_t)]} & \text{if } \phi \subset t \text{ and } \phi \in \mathcal{B}_G \\ 0 & \text{if } \phi \not\subset t. \end{cases} \tag{2}$$

with $\phi \subset t$ standing for that ϕ is in t.

The *observed* coverage on an exonic segment $e \in \mathcal{E}_G$ as $C(e) = \frac{1}{l(e)}\sum_{i=1}^{l(e)} c_i$, where c_i is the number of reads covering the ith nucleotide in e. The read coverage $C(e)$ provides an estimator for the number of transcript copies that flow through the exonic segment e assuming uniform sampling. For a MultiSplice $b \in \mathcal{B}_G$, we use $C(b)$ to denote the read coverage on b defined as the number of transcript fragments that include b.

3.3 A generalized linear model for transcript quantification

We construct a matrix $\mathbf{M}' \in \Re^{|\Phi_G|\times|\mathcal{T}_G|}$ to represent the structure of the transcripts, whose entry on the row of ϕ and the column of t corresponds to the probability of observing feature ϕ from transcript t, $\mathbf{M}'(\phi,t) = m(\phi,t)$. The linear model is set up for every feature $\phi \in \Phi_G$ by equating the observed coverage on ϕ to the expected coverage from all transcripts:

$$C(\phi) = \sum_{t\in\mathcal{T}_G} \mathbf{M}'(\phi,t)C(t) + \epsilon_\phi, \text{for any } \phi \in \Phi_G. \tag{3}$$

Here $C(t) \geq 0$ for every $t \in \mathcal{T}_G$, ϵ_ϕ is the error term for feature ϕ in transcript t.

Lemma 1. *The MultiSplice model for transcript quantification is identifiable if the rank of M' is no less than the number of transcripts $|\mathcal{T}_G|$.*

Lemma 1 directly follows the the Rouché-Capelli theorem [11].

4 Bias correction

Under uniform sampling, the sampling probability is the same at every nucleotide of a transcript. The observed coverage on ϕ is unbiased for the expected coverage on t. In this case, $\sigma(\phi,t)$ is set to 1 for all transcripts and features. However, sampling bias is often introduced in RNA-seq sample preparation protocols and has been demonstrated to have significant effects in RNA-seq analysis [14, 35]. Therefore, we discuss in the following subsections how MultiSplice corrects various sampling bias via learning of the bias coefficients and simultaneously solves the linear model for transcript coverage $C(t)$ of every transcript t.

Figure 4(a-e) shows how various types of sampling bias alter the sampling probability and hence the coverage. Two types of sampling bias are commonly observed in RNA-seq data, namely, the position-specific bias and the sequence-specific bias [6, 4, 21, 27]. In our model, sampling bias may affect the sampling probability of exonic segments and MultiSplices. Therefore, we calculate the bias coefficient $\sigma(\phi,t)$ for every feature $\phi \in \Phi_G$ and every transcript t so that $E[C(\phi|t)] = \sigma(\phi,t)m(\phi,t)C(t)$. Next, we introduce each independent bias individually.

Sequence-specific bias. The sequence-specific bias refers to the perturbation of sampling probability related to certain sequences at the beginning or end of transcript fragments [4, 18]. The characteristic of this type of bias in the given RNA-seq sample can be learned in advance by examining the relationship between

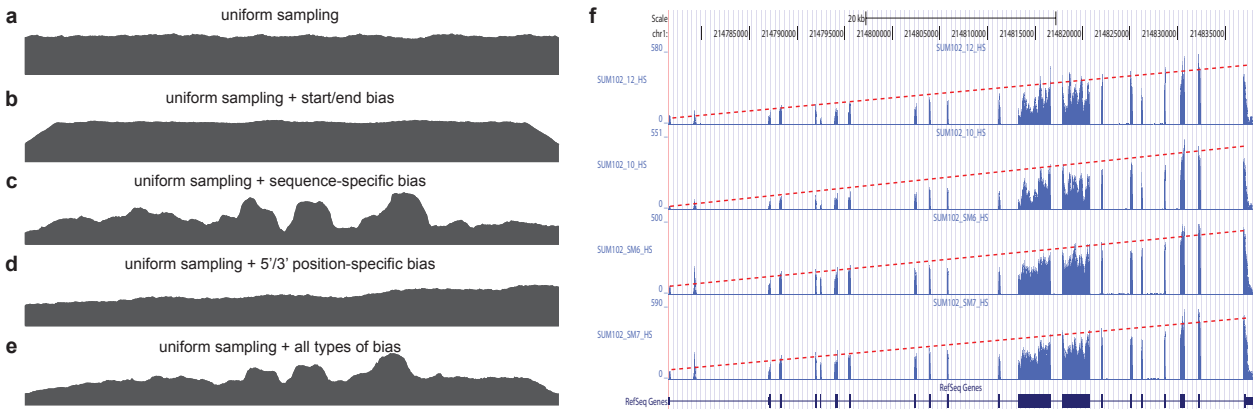


Fig. 2: Sampling bias present in the RNA-seq data. **a**. RNA-seq read coverage under uniform sampling. **b**. RNA-seq read coverage under uniform sampling with transcript start/end bias. **c**. RNA-seq read coverage under uniform sampling with sequence-specific bias. **d**. RNA-seq read coverage under uniform sampling with 5'/3' position-specific bias. **e**. RNA-seq read coverage under uniform sampling with all aforementioned types of bias. **f**. Sampling bias on gene CENPF in the breast cancer dataset used in Section 6. Please note that the second peak in the coverage plot is not an exon in CENPF. The observed coverage on each exon decreases almost linearly from the 3' end to the 5' end. The coverage also drops at the bases near the end of the gene. The non-uniformity in the two middle large exons is likely to be due to the sequence-specific sampling bias.

GC content and the observed coverage on single-isoform genes. To derive the sequence-specific bias at an arbitrary exonic position, we look into 8bp upstream to the 5' start to 11bp downstream according to [25]. A Markov chain is constructed to model the effect on the sampling probability at the position from the sequence of surrounding nucleotides. Then we use an approach based on the probabilistic suffix tree [5] to learn the sequence-specific bias coefficient $\alpha(t,i)$ for ith nucleotide in transcript t.

Transcript start/end bias. Sampling near transcript start site or transcript end site is often insufficient. The read coverage in these regions is typically lower than expected because the positions where a sampled read can cover are restricted by the transcript boundaries. The bias coefficient for start/end bias at the ith nucleotide in transcript t is written as:

$$\beta(t,i) = \begin{cases} i/E[l(fr)] & \text{if } i < E[l(fr)] \\ 1 & \text{if } E[l(fr)] \le i \le l(t) - E[l(fr)] \\ (l(t)-i)/E[l(fr)] & \text{if } i > l(t) - E[l(fr)]. \end{cases}$$

5'/3' position-specific bias. Position-specific bias refers to the alteration on sampling probability according to position in the transcript. For example, nucleotides to the 3' end of the transcript have higher probability to be sampled in Figure 4(d). Here we model the position-specific bias coefficient as a linear function, $\gamma(t,i) = \gamma_1^t \cdot i + \gamma_0^t$. The intercept γ_0^t gives the bias coefficient at the 5' transcript start site. The slope γ_1^t measures the extent of the bias: a positive γ_1^t indicates that 3' transcript end site has higher sampling probability than the start site; a zero γ_1^t indicates no positional bias in the transcript t.

Combined bias model. Assuming the above three types of bias have independent effect on read sampling, we derive the bias coefficient at ith nucleotide in transcript t as $\sigma(t,i) = \alpha(t,i) \cdot \beta(t,i) \cdot \gamma(t,i)$. The bias coefficient of an exonic segment $e \in \mathcal{E}_g$ is then the averaged bias coefficient on all positions in the exonic segment e, and the bias coefficient of a MultiSplice $b \in \mathcal{B}_g$ is the averaged bias coefficient on all positions in its sampling window $w(b)$. In summary, the bias coefficient for a MultiSplice feature $\phi \in \Phi_g$ in transcript t is

$$\sigma(\phi,t) = \begin{cases} \frac{\sum_{i\in\phi}\sigma(t,i)}{l(\phi)} & \text{if } \phi \subset t \text{ and } \phi \in \mathcal{E}_g \\ \frac{\sum_{i\in w\phi}\sigma(t,i)}{E[l(w(\phi))]} & \text{if } \phi \subset t \text{ and } \phi \in \mathcal{B}_g \\ 0 & \text{if } \phi \not\subset t. \end{cases} \quad (4)$$

5 Solving the generalized linear models with bias correction

Conventionally, we are interested in the set of transcript expressions that minimize the sum of squared errors, the absolute residuals between the expected coverage and the observed coverage. This solution is relatively sensitive to unexpected sampling noise which often occurs in real RNA-seq samples and may lead to a highly unstable extrapolation when the expression of the alternative splicing events discriminating the transcripts is notably lower than the average level of gene expression. Therefore, we define the sum of squared relative errors (SSRE), which measures the relative residual regarding the ratio of the expected coverage against the observed coverage.

$$SSRE = \sum_{f \in \mathcal{F}_G} \left(\frac{\sum_{t \in \mathcal{T}_G} \sigma(f,t)\mathbf{M}'(f,t)C(t)}{C(f)} - 1 \right)^2. \tag{5}$$

Bias parameter estimates. Among all the bias parameters, the sequence-specific bias is learned in advance while the start and end bias is a function of transcript fragment length. The only bias parameters unknown related to the 3' bias are defined by the intercept γ_0^t and slope γ_1^t for every transcript $t \in \mathcal{T}_g$. Therefore, we use an iterative-minimization strategy and search for a set of bias coefficients γ_0^t's and γ_1^t's that better fit the RNA-seq sample than the uniform sampling model. We start with the transcript coverage $C(t)$'s that are solved from the uniform sampling model (with $\gamma_0^t = 1$ and $\gamma_1^t = 0$ as initial condition). Analogous to the hill climbing algorithm [26], we then iteratively probe a locally optimal set of transcript coverage together with the bias coefficients around the uniform solution through minimizing the SSRE. In each iteration, a candidate solution is obtained through sequentially setting the partial derivatives to 0 with respect to every unknown parameter γ_0^t, γ_1^t, $C(t)$, and for every transcript $t \in \mathcal{T}_G$. If the candidate solution results in a smaller SSRE, the candidate solution is taken and the iteration continues. For details of the step to estimate the bias parameters, please refer to the Appendix section.

Solving the linear model with LASSO regularization. Lastly, we solve for the level of individual transcript expression with additional regularization, based on the bias coefficients from the previous step. One common problem in transcript quantification is that the set of expressed transcripts are not known a priori. Hence it becomes crucially important to identify the set of truly expressed transcripts provided in a candidate set. Therefore, we further apply the L1 regularization (known as LASSO) for its proven effectiveness in irrelevance-removal and solve for the set of transcript expression $C(\mathcal{T}_G)$ that minimizes the following loss function

$$L = \text{SSRE} \ + \ \text{L1 penalty} = \sum_{\phi \in \Phi_G} \left(\frac{\sum_{t \in \mathcal{T}_G} \sigma(\phi,t)\mathbf{M}'(\phi,t)C(t)}{C(\phi)} - 1 \right)^2 + \lambda ||C(\mathcal{T}_G)||_1, \tag{6}$$

where $\lambda \geq 0$ denotes the weight of the L1 shrinkage and $C(t) \geq 0$ for every $t \in \mathcal{T}_G$.

6 Experimental Results

To evaluate the performance of the MultiSplice model, we compared it with three other approaches. The *ExonOnly* model, where only exonic segments are used to represent transcript composition as proposed in SLIDE [20], was implemented using a linear regression approach with LASSO. The ExonOnly model provided the baseline comparison for MultiSplice. The *Poisson* model, which was originally proposed by [24], was implemented in C since it is not publicly available. Cufflinks [31] is a representative of read-centric model. Cufflinks 1.1.0 was downloaded from its website in September, 2011.

These algorithms were run on both simulated datasets and real datasets. Reads were first mapped by MapSplice 1.15.1 [34] to the reference genome. If the read was paired-end, MapPER [12] was applied to infer the alignment of the entire transcript fragment.

6.1 Transcriptome identifiability with increasing read length

We first study how the increase in read length may alleviate the lack of identifiability issues in transcript quantification using MultiSplice. We downloaded UCSC gene models in human (track UCSC genes:GRCh37/hg19), mouse (track UCSC Genes:NCBI37/mm9), worm (track WormBase Genes:WS190/ce6) and fly (track FlyBase Genes:BDGP R5/dm3). We computed the feature matrix used in MultiSplice given variable read length

and determined its rank. The transcript isoforms of a gene is identifiable if the rank of the feature matrix is no less than the number of transcripts. Figure 3 plots the additional number of genes that become identifiable as the read length increases from 50bp assuming single-end read RNA-seq data. For all four species, as the read length increases, MultiSplice is capable of resolving the transcript quantification issues of more genes. With 500bp reads, about 98% genes in both human and mouse become identifiable. Surprisingly, for worm and fly, 500bp reads do not gain significant improvement over 50bp reads. This is mostly due to the fact that the exon lengths of fly and worm are comparably much longer [9] than human and mouse, making it difficult for reads of moderate size to take effect. With current short read technology where read length is typically 100bp or less, paired-end reads with the size of transcript fragments around 500bp may be the most economical and effective for transcription quantification for genes with identifiability issues. This is under the assumption that it is possible to infer the transcript fragment from paired-end reads based on the tightly controlled distribution of insert-size.

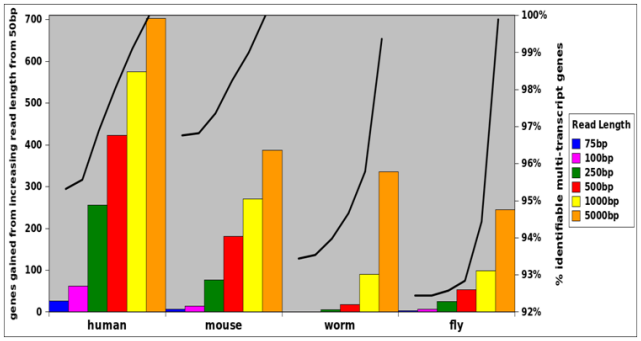


Fig. 3: Changes in mRNA identifiability as a function of transcript fragment/read length. Starting from levels achieved with 50bp single-end reads, the left side of the y-axis shows the additional number of genes that become identifiable using MultiSplice as the read length increases. The y-axis on the right side shows the total percentage of genes for which mRNA transcript structures are resolved. The UCSC annotated transcript sets of four species: human, mouse, fly and worm were used for this analysis.

6.2 Simulated human RNA-seq experiment

Data Simulation. Due to the lack of the ground truth within real datasets, simulated data has become an important resource for the evaluation of transcript quantification algorithms [6, 17, 22]. We developed an in-house simulator to generate RNA-seq datasets of a given sampling depth using human hg19 Refseq annotation. The simulation process consists of three steps: (1) randomly assign relative proportions to all the transcripts within a gene and set this as the true profile; (2) calculate the number of reads to be sampled from each transcript; (3) sample transcript fragments of a given length along the transcripts according to the per base coefficient $\sigma(t,i) = \frac{ki\alpha(t,i)\beta(t,i)}{l(t)} + 1$ for the ith base on transcript t, where $\alpha(t,i)$ and $\beta(t,i)$ are the sequence-specific bias and the transcript start/end bias as defined in Section 4 and k is the slope of the position-specific bias. Paired-end reads will be generating by taking the two ends of the transcript fragment. Please note the sequence bias per base has been learned from a real dataset, a technical replicate of MCF-7 data that will be introduced in the next section.

Accuracy measurement. Due to inconsistencies in the normalization scheme used by different software, the estimated abundance may not be comparable among different approaches. Hence, we computed relative proportions of transcript isoforms for each method. The similarity between the estimated result and the ground truth is measured by both Pearson correlation and Euclidean distance. Let X denote the vector of real isoform proportions of a gene and $\hat{X}$ denote the estimated proportions. The formula of the correlation is: $r(X,\hat{X}) = cov(X,\hat{X})/(\sigma_X \cdot \sigma_{\hat{X}})$. A value close to 1 means that our estimation is highly accurate and vice versa. Below, we adopt a boxplot to illustrate the performance of each method. The box is constructed by the 1st quartile, the median, and the 3rd quartile. The ends of the upper and lower whisker are given by the 3rd quartile $+1.5 \times IQR$*(inner quartile range)* and 1st quartile $-1.5 \times IQR$, respectively. Due to the space limit, we present the result of correlation measurement in the main manuscript. Results measured by Euclidian distance can be found in the Appendix section.

Sampling depth. Next we evaluate how the sequencing depth may affect the accuracy of transcript abundance estimation. Four groups of 2x50bp paired-end synthetic data (insert size 100bp) were generated on the whole human transcriptome with increasing number of reads: 6 million, 12 million, 18 million and 24 million. 14530 genes with multiple isoforms are selected for analysis. The genes were divided into three subsets: (1) 13576 genes to which identifiability holds for all methods. (2) 455 genes to which identifiability holds for MultiSplice. (3)499 genes to which identifiability does not hold for all methods.

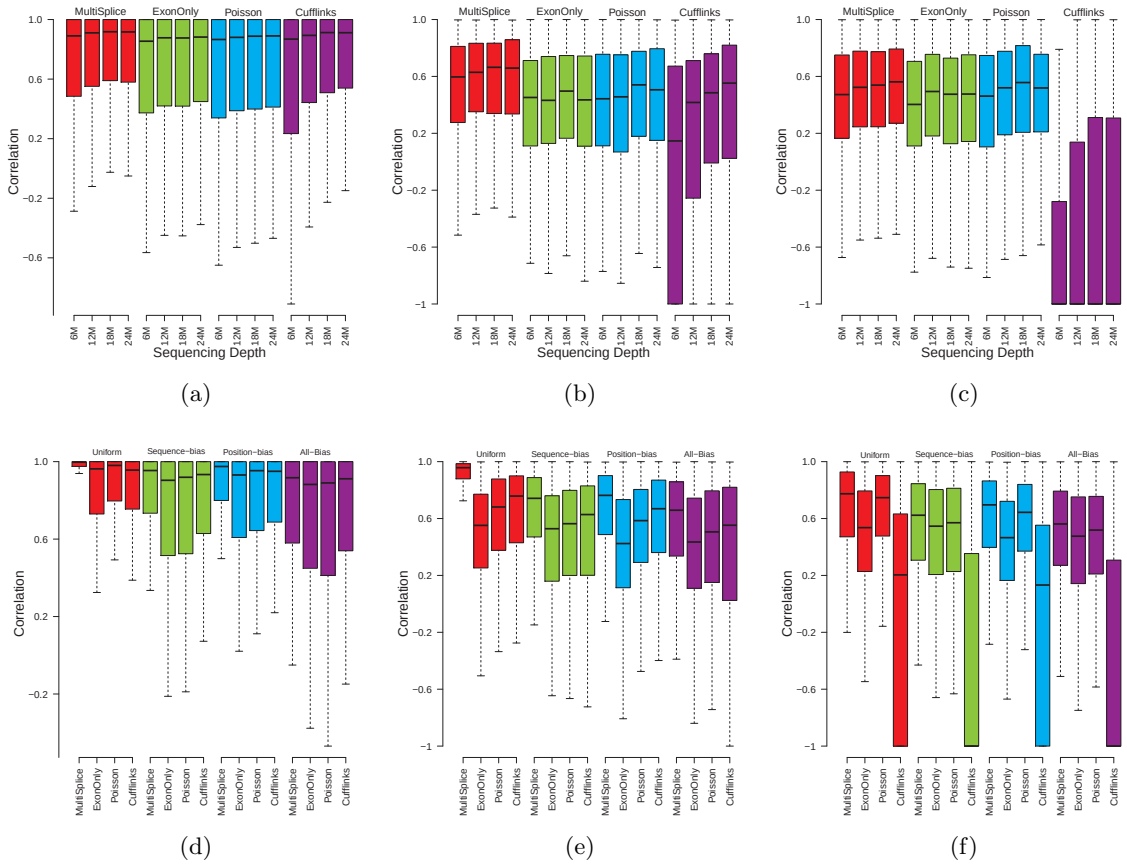


Fig. 4: a-c. Boxplots of the correlation between estimated transcript proportions and the ground truth under varying number of sampled reads: 6M, 12M, 18M and 24M over a total of 14530 human genes with more than one isoforms. (a),(b) and (c) correspond to the gene set that is identifiable with basic exon structure, identifiable with additional MultiSplice features, and unidentifiable, respectively. d-f: Boxplots of the correlation between estimated transcript proportions and the ground truth under four circumstances: uniform sampling, sampling with positional bias only, with sequence bias only and with all bias. (d),(e) and (f) correspond to the gene set that is identifiable with basic exon structure, identifiable with additional MultiSplice features, and unidentifiable, respectively.

For each subplot in Figure 4(a, b, c), the estimation accuracy for all methods generally improves as more reads are sampled. Cufflinks seems to be affected most by the sampling depth. For the genes whose identifiability conditions are satisfied for all methods, the correlation between the estimated transcript proportion is highly similar with the ground truth, with an average correlation close to 0.9 for all methods. In the second category, when the genes are still identifiable with MultiSplice, the estimation accuracy of MultiSplice remains high, with an average correlation above 0.6 while others slip below 0.5. For the category when identifiability is not satisfied for all methods, the estimation accuracy is degraded even more. However, MultiSplice still consistently gives better estimation results indicating that the inclusion of MultiSplice features make transcript quantification more stable than other methods. Cufflinks demonstrated the worst performance in this category, mainly because the unidentifiability conditions make it difficult to assign these reads to a transcript. Instead, it throws out most of multi-mapped reads. Apparently, increasing sampling depth cannot alleviate the issue of unidentifiability.

Bias correction. To study the effect of the bias correction, we have simulated data with uniform sampling, sampling with only positional bias, sampling with only sequence bias, and sampling with the combined positional and sequence bias. Here, we set the slope of the position-specific bias k to 2 with 24 million 2x50bp paired-end reads sampled from the whole transcriptome for each case. All the approaches achieve the best results when the sampling process is uniform. As positional or sequence bias is introduced, their performance tapers down. The presence of both positional and sequence biases has the largest impact in

all methods. Meanwhile, because MultiSplice and Cufflinks correct both sequence and positional bias, these two methods are more robust and outperform the ExonOnly and the Poisson methods in all categories.

Inference of expressed transcripts. Quantification of mRNAs usually rely on a set of candidate transcript structures as input. It is unknown in apriori whether each transcript is present in a sample or not. Therefore, accurate quantification methods should be able to infer the transcripts that are expressed as well as those that are not. To assess the capability of the various methods to infer expressed transcripts, we generated simulated 2x50bp paired-end reads from human genes with at least 3 transcripts. We randomly chose two transcripts from one gene and simulated reads only from these transcripts. The remaining transcripts were not sampled. We used the false positive rate to measure the accuracy of the inference. Non-expressed transcripts that were estimated with a positive abundance above a given threshold were counted as the false positives. As shown in Figure 5, MultiSplice demonstrated the lowest false positive rate in the identification of dominant transcripts. Poisson and Cufflinks tended to assign positive expression to every transcript including those that are not expressed. Even when the threshold was raised to 10%, the false positive rate remained high for some methods especially Cufflinks. MultiSplice, in general, outperformed the others in identifying of the correct set of expressed transcripts.

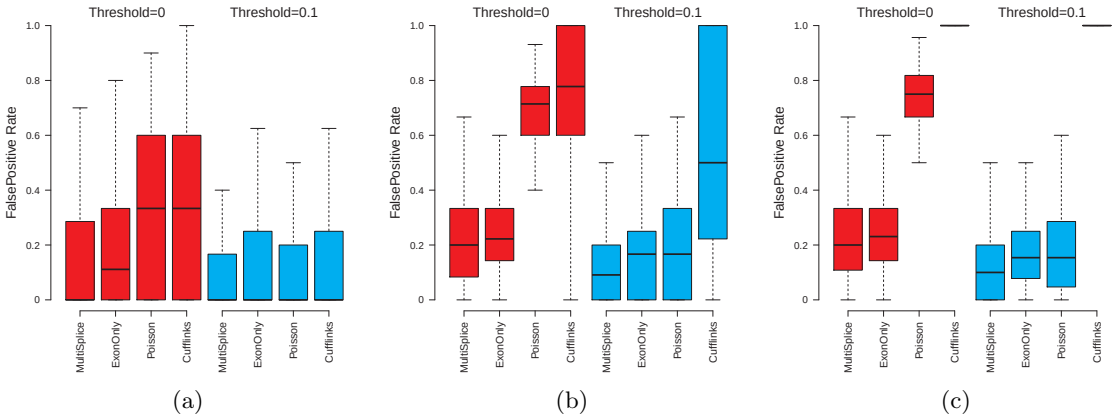


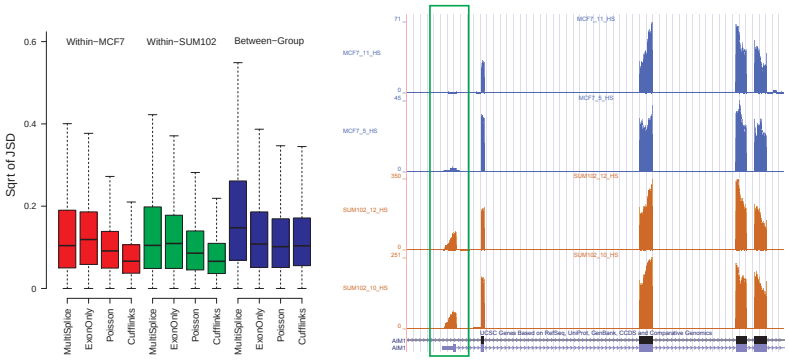
Fig. 5: Comparison of false positive rates in the inference of the expressed transcripts. Thresholds represent the minimum fraction of a transcript that is considered expressed. (a),(b) and (c) correspond to the gene set that is identifiable with the basic exon structure, identifiable with additional MultiSplice features, and unidentifiable, respectively.

6.3 Real human RNA-seq experiment

We attempted to use RNA-seq data generated from the samples in the Microarray Quality Control (MAQC) Project [15] with TaqMan qRT-PCR measurements of the abundance for approximate 1000 genes. Our primary interest is in disambiguating multiple isoforms using MultiSplice features. However, most of these genes express only a single isoform. Therefore, we applied the set of transcript quantification methods to a dataset that was originally used by Singh et al. to study differential transcription [28]. In this study, two groups of RNA-seq datasets were generated from SUM-102 and MCF-7, two breast cancer cell lines. Each group contains 4 samples as technical replicates. The RNA-seq data were generated from Illumina HISEQ2000. Each sample had 80 million 100bp single-end reads. About 60 million reads can be aligned to the reference genome by MapSplice. The Refseq human annotated transcripts were fed into each software for transcript quantification.

Since ground truth expression profiles do not exist for the real datasets, we investigated whether the different methods provided a consistent estimation within samples of technical replicates which only vary by random sampling. In contrast, a significant number of genes between MCF-7 and SUM-102 were expected to be differentially expressed [28]. To evaluate this, we computed Jensen−Shannon divergence (JSD), used in Cuffdiff [1] to measure the dissimilarity between two samples and calculated the *within-group* and *between-group* differences. As detailed in Figure 6(a), both MultiSplice and Cufflinks had smaller average within-group difference than the average between-group difference while the other two methods do not show clear difference. MultiSplice demonstrated higher between-group difference than Cufflinks, but also had relatively higher within-group differences as well. Most of these, however, were well below a JSD of 0.2 and

considered to be insignificant. A closer look at a number of cases showed that occasionally MultiSplice and Cufflinks may overestimate or underestimate the between-group difference respectively. Figure 6(b) (The complete figure with 8 samples can be found in the Appendix Figure 8(a)) shows a gene where Cufflinks underestimated the difference between the two groups. The second isoform of the gene AIM1 has a unique first exon (chr6:106989461-106989496). Clear difference in the read coverage on this exon can be observed between the two groups, indicating strong differential levels of expression, i.e., the second isoform is barely expressed in MCF-7 while almost comparable to the first isoform in SUM-102 cells. The between group square root of JSD is 0.21 by Cufflinks, much lower than 0.50 by MultiSplice.



(a) (b)

Fig. 6: **a.** Boxplots of the within-MCF-7, within-SUM-102, and between-group square root of JSD of all genes for all methods. **b.** A case where Cufflinks underestimated the difference between the two groups. The second isoform of Gene AIM1 has a unique first exon, whose read coverage differs significantly between the two groups. A detailed plot with all 8 samples can be found in the Appendix Figure 8(a).

The exon-skipping event found in gene CD46 is also differentially expressed (Figure 8(b), Appendix). The estimation of transcript quantification with MultiSplice was consistent with the observation in the qRT-PCR data showing that steady state levels of transcripts with the skipped exon were present in amounts more than two fold higher expression in SUM-102 than in MCF-7 cells.

7 Conclusion

In this paper, we propose a generalized linear system for the accurate quantification of alternative transcript isoforms with RNA-seq data. We introduce a set of new structural features, namely MultiSplice, to ameliorate the issue of *identifiability*. With MultiSplice features, 98% of Refseq transcript models in human and mouse become identifiable with 500bp reads (or paired-end reads with 500bp transcript fragments), an 8% increase from 50bp. Therefore, longer reads or paired-end reads with longer insert-sizes rather than further increases in sequencing depths can be crucial for the accurate quantification of mRNA isoforms with complex alternative transcription, even though a majority of the genes have relatively simple transcript variants. The results also demonstrate the robustness of the MultiSplice method under various sampling biases, consistently outperforming three other methods: Cufflinks, Poisson and ExonOnly. The application of our approach to real RNA-seq datasets for transcriptional profiling successfully identified a number of isoforms whose proportion changes differed significantly between two distinct breast cancer cell lines. In the near future, we will continue to experiment our algorithms with more complex gene models including those from Ensembl database and those transcripts that are directly assembled from RNA-seq.

8 Acknowledgements

We thank the referees for their insightful comments. We thank Christian F. Orellana for running Cufflinks on real datasets. We thank Dr. Charles Perou for the RNA-seq samples from the MCF-7 and SUM-102 cell lines. *Funding*: This work was supported by The National Science Foundation [CAREER award grant number 1054631 to J.L.]; the National Science Foundation [ABI/EF grant number 0850237 to J.L. and J.F.P.] and National Institutes of Health [grant number P20RR016481 to J.L.]. Additional support was provided by the National Institutes of Health: NCI TCGA [grant number CA143848 to Charles Perou] and an Alfred P. Sloan Foundation fellowship [D.Y.C.].

References

[1] *Cufflinks*. `http://cufflinks.cbcb.umd.edu`.

[2] *Ensembl Genome Browser*. `http://useast.ensembl.org/index.html`.

[3] *NCBI Reference Sequence (RefSeq)*. `http://www.ncbi.nlm.nih.gov/RefSeq`.

[4] Roberts A., Trapnell C., Donaghey J., Rinn J., and Pachter L. Improving rna-seq expression estimates by correcting for fragment bias. *Genome Biology*, 12(3):R22, 2011.

[5] Gill Bejerano. Algorithms for variable length markov chain modeling. *Bioinformatics*, 20:788789, 2004.

[6] Regina Bohnert and Gunnar R. rquant.web: a tool for rna-seq-based transcript quantitation. *Nucleic Acids Research*, 38(Suppl 2):W348–W351, 2010.

[7] Jean-Philippe Brosseau, Jean-Franois Lucier, Elvy Lapointe, Mathieu Durand, Daniel Gendron, Julien Gervais-Bird, Karine Tremblay, Jean-Pierre Perreault, and Sherif Abou Elela. High-throughput quantification of splicing isoforms. *RNA society*, 16:442–449, 2010.

[8] Jianxing Feng, Wei Li, and Tao Jiang. Inference of isoforms from short sequence reads. *RECOMB*, 2010.

[9] Kristi L. Fox-Walsh, Yimeng Dou, Bianca J. Lam, She pin Hung, Pierre F. Baldi, and Klemens J. Herte. The architecture of pre-mrnas affects mechanisms of splice-site pairing. *Proc. Natl. Acad. Sci.*, 102(45):16176–16181, 2005.

[10] Mitchell Guttman, Manuel Garber, Joshua Z Levin, Julie Donaghey, James Robinson, Xian Adiconis, Lin Fan, Magdalena J Koziol, Andreas Gnirke, Chad Nusbaum, John L Rinn, Eric S Lander, and Aviv Regev. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nature Biotechnology*, 28:503–510, 2010.

[11] Roger A. Horn and Charles R. Johnson. Matrix analysis. *Cambridge University Press*, 1990.

[12] Yin Hu, Kai Wang, Xiaping He, Derek Y. Chiang, Jan F. Prins, and Jinze Liu. A probabilistic framework for aligning paired-end rna-seq data. *Bioinformatics*, 26:1950–1957, 2010.

[13] Hui Jiang and Wing Hung Wong. Statistical inferences for isoform expression in rna-seq. *Bioinformatics*, 25:1026–1032, 2009.

[14] Iwanka Kozarewa, Zemin Ning, Michael A Quail, Mandy J Sanders, Matthew Berriman, and Daniel J Turner. Amplification-free illumina sequencing-library preparation facilitates improved mapping and assembly of (g+c)-biased genomes. *Nuc*, 6:291–295, 2009.

[15] Shi L, Reid LH, and et.al. Jones WD. The microarray quality control (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24(9):1151–61, 2006.

[16] Vincent Lacroix, Michael Sammeth, Roderic Guigo, and Anne Bergeron. Exact transcriptome reconstruction from short sequence reads. *Proceedings of the 8th international workshop on Algorithms in Bioinformatics*, 2008.

[17] Bo Li, Victor Ruotti, Ron M. Stewart, James A. Thomson, and Colin N. Dewey. Rna-seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26 (4):493–500, 2010.

[18] Jun Li, Hui Jiang, and Wing H Wong. Modeling non-uniformity in short-read rates in rna-seq data. *Genome Biology*, 11, 2010.

[19] Wei Li, Jianxing Feng, and Tao Jiang. Isolasso: A lasso regression approach to rna-seq based transcriptome assembly. *RECOMB*, 6577:167–188, 2011.

[20] Jingyi Jessica Lia, Ci-Ren Jiangb, James B. Browna, Haiyan Huanga, and Peter J. Bickela. Sparse linear modeling of next-generation mrna sequencing (rna-seq) data for isoform discovery and abundance estimation. *PNAS*, 2011.

[21] Olejniczak M., Galka P., and Krzyzosiak W.J. Sequence-non-specific effects of rna interference triggers and microrna regulators. *Nucl. Acids Res.*, 38(1):1–16, 2010.

[22] Marius Nicolae, Serghei Mangul, Ion I Mandoiu, and Alex Zelikovsky. Estimation of alternative splicing isoform frequencies from rna-seq data. *Algorithms for Molecular Biology*, 6:9, 2011.

[23] Qun Pan, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics*, 40:1413 – 1415, 2008.

[24] Hugues Richard, Marcel H. Schulz, Marc Sultan, Asja Nrnberger, Sabine Schrinner, Daniela Balzereit, Emilie Dagand, Axel Rasche, Hans Lehrach, Martin Vingron, Stefan A. Haas, and Marie-Laure Yaspo. Prediction of alternative isoforms from exon expression levels in rna-seq experiments. *Nucleic Acids Research*, 38:e112, 2010.

[25] Adam Roberts, Cole Trapnell, Julie Donaghey, John L Rinn, and Lior Pachter. Improving rna-seq expression estimates by correcting for fragment bias. *Genome Biology*, 12:R22, 2011.

[26] Stuart Russell and Peter Norvig. Artificial intelligence: A modern approach. page R22, 2003.

[27] Srivastava S. and Chen L. A two-parameter generalized poisson model to improve the analysis of rna-seq data. *Nucleic Acids Research*, pages 1–15, 2010.

[28] Darshan Singh, Christian F. Orellana, Yin Hu, Corbin D. Jones, Yufeng Liu, Derek Y. Chiang, Jinze Liu, and Jan F. Prins. Fdm: A graph-based statistical method to detect differential transcription using rna-seq data. *Bioinformatics*, 2011.

[29] Sudeep Srivastava and Liang Chen. A two-parameter generalized poisson model to improve the analysis of rna-seq data. *Nucleic Acids Research*, 38:e112, 2010.
[30] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society Series B.*, 58:267–288, 1996.
[31] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28:511–515, 2010.
[32] Ernest Turro, Shu-Yi Su, ngela Gonalves, Lachlan JM Coin, Sylvia Richardson, and Alex Lewin. Haplotype and isoform specific expression estimation using multi-mapping rna-seq reads. *Genome Biology*, 12:R13, 2011.
[33] Eric T. Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge1. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456:470–476, 2008.
[34] Kai Wang, Darshan Singh, Zheng Zeng, Yan Huang, Stephen Coleman, Gleb L. Savich, Xiaping He, Piotr Mieczkowski, Sara A. Grimm, Charles M. Perou, James N. MacLeod, Derek Y. Chiang, Jan F. Prins, and Jinze Liu. Mapsplice: Accurate mapping of rna-seq reads for splice junction discovery. *Nucleic Acid Research*, 38 (18):178, 2010.
[35] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10:57–63, 2009.
[36] Jie Wu, Martin Akerman, Shuying Sun, W. Richard McCombie, Adrian R. Krainer, and Michael Q. Zhang. Splicetrap: a method to quantify alternative splicing under single cellular conditions. *Bioinformatics*, 2011.
[37] Zhengpeng Wu, Xi Wang, and Xuegong Zhang. Using non-uniform read distribution models to improve isoform expression inference in rna-seq. *Bioinformatics*, 27:502–508, 2011.

Appendix

Iterative-minimization algorithm

In Section 5, we use an iterative-minimization strategy to search for a set of bias coefficients γ_0^t's and γ_1^t's for every transcript $t \in \mathcal{T}_g$ that better fit the RNA-seq sample than the uniform sampling model. We initiate the iterations with the transcript coverage $C(t)$'s solved from the uniform sampling model and the bias coefficients $\gamma_0^t = 1$ and $\gamma_1^t = 0$. In each iteration, for transcript t we set:

1. $\frac{\partial SSRE}{\partial C(t)} = 0$; 2. $\frac{\partial SSRE}{\partial \gamma_1^t} = 0$; 3. $\frac{\partial SSRE}{\partial \gamma_0^t} = 0$.

$$
\begin{aligned}
&\frac{\partial SSRE}{\partial C(t)} = 0 \\
\Rightarrow &\sum_{\phi \in \Phi_g} 2(C(\phi) - \sum_{s \in \mathcal{T}_g} \sigma(\phi, s)\mathbf{M}'(\phi, s)C(s)) \cdot \sigma(\phi, t)\mathbf{M}'(\phi, t) = 0 \\
\Rightarrow &\sum_{s \in \mathcal{T}_g} C(s)(\sum_{\phi \in \Phi_g} \sigma(\phi, s)\mathbf{M}'(\phi, s)\sigma(\phi, t)\mathbf{M}'(\phi, t)) = \sum_{\phi \in \Phi_g} C(\phi)\sigma(\phi, t)\mathbf{M}'(\phi, t) \\
\Rightarrow &C(t) = \frac{\sum_{\phi \in \Phi_g} C(\phi)\sigma(\phi, t)\mathbf{M}'(\phi, t) - \sum_{s \in \mathcal{T}_g, s \neq t} C(s)(\sum_{\phi \in \Phi_g} \sigma(\phi, s)\mathbf{M}'(\phi, s)\sigma(\phi, t)\mathbf{M}'(\phi, t))}{\sum_{\phi \in \Phi_g} \sigma(\phi, t)\mathbf{M}'(\phi, t)\sigma(\phi, t)\mathbf{M}'(\phi, t)}.
\end{aligned}
$$

$\sigma(\phi, t)$ is the only function related to γ_1^t and γ_0^t.

$$
\begin{aligned}
&\frac{\partial SSRE}{\partial \gamma_1^t} = 0 \\
\Rightarrow &\sum_{\phi \in \Phi_g} 2(C(\phi) - \sum_{s \in \mathcal{T}_g} \sigma(\phi, s)\mathbf{M}'(\phi, s)C(s)) \cdot \frac{\partial \sigma(\phi, t)}{\partial \gamma_1^t}\mathbf{M}'(\phi, t)C(t) = 0 \\
\Rightarrow &C(t) \sum_{\phi \in \Phi_g} \sigma(\phi, t)\mathbf{M}'(\phi, t)\frac{\partial \sigma(\phi, t)}{\partial \gamma_1^t}\mathbf{M}'(\phi, t) \\
&= \sum_{\phi \in \Phi_g} C(\phi)\frac{\partial \sigma(\phi, t)}{\partial \gamma_1^t}\mathbf{M}'(\phi, t) - \sum_{s \in \mathcal{T}_g, s \neq t} C(s)(\sum_{\phi \in \Phi_g} \sigma(\phi, s)\mathbf{M}'(\phi, s)\frac{\partial \sigma(\phi, t)}{\partial \gamma_1^t}\mathbf{M}'(\phi, t)).
\end{aligned}
$$

Similarly,

$$\frac{\partial SSRE}{\partial \gamma_0^t} = 0$$

$$\Rightarrow \sum_{\phi \in \Phi_g} 2(C(\phi) - \sum_{s \in \mathcal{T}_g} \sigma(\phi, s)\mathbf{M}'(\phi, s)C(s)) \cdot \frac{\partial \sigma(\phi, t)}{\partial \gamma_0^t}\mathbf{M}'(\phi, t)C(t) = 0$$

$$\Rightarrow C(t) \sum_{\phi \in \Phi_g} \sigma(\phi, t)\mathbf{M}'(\phi, t)\frac{\partial \sigma(\phi, t)}{\partial \gamma_0^t}\mathbf{M}'(\phi, t)$$

$$= \sum_{\phi \in \Phi_g} C(\phi)\frac{\partial \sigma(\phi, t)}{\partial \gamma_0^t}\mathbf{M}'(\phi, t) - \sum_{s \in \mathcal{T}_g, s \neq t} C(s)(\sum_{\phi \in \Phi_g} \sigma(\phi, s)\mathbf{M}'(\phi, s)\frac{\partial \sigma(\phi, t)}{\partial \gamma_0^t}\mathbf{M}'(\phi, t)).$$

Because $\sigma(\phi, t)$ is a linear combination of γ_1^t and γ_0^t, and hence $\sum_{\phi \in \Phi_g} \sigma(\phi, t)\mathbf{M}'\phi, t$ is also the linear combination of γ_1^t and γ_0^t. Then we can directly calculate $\frac{\partial \sigma(\phi,t)}{\partial \gamma_1^t}$ and $\frac{\partial \sigma(\phi,t)}{\partial \gamma_0^t}$.

Supplementary Figures

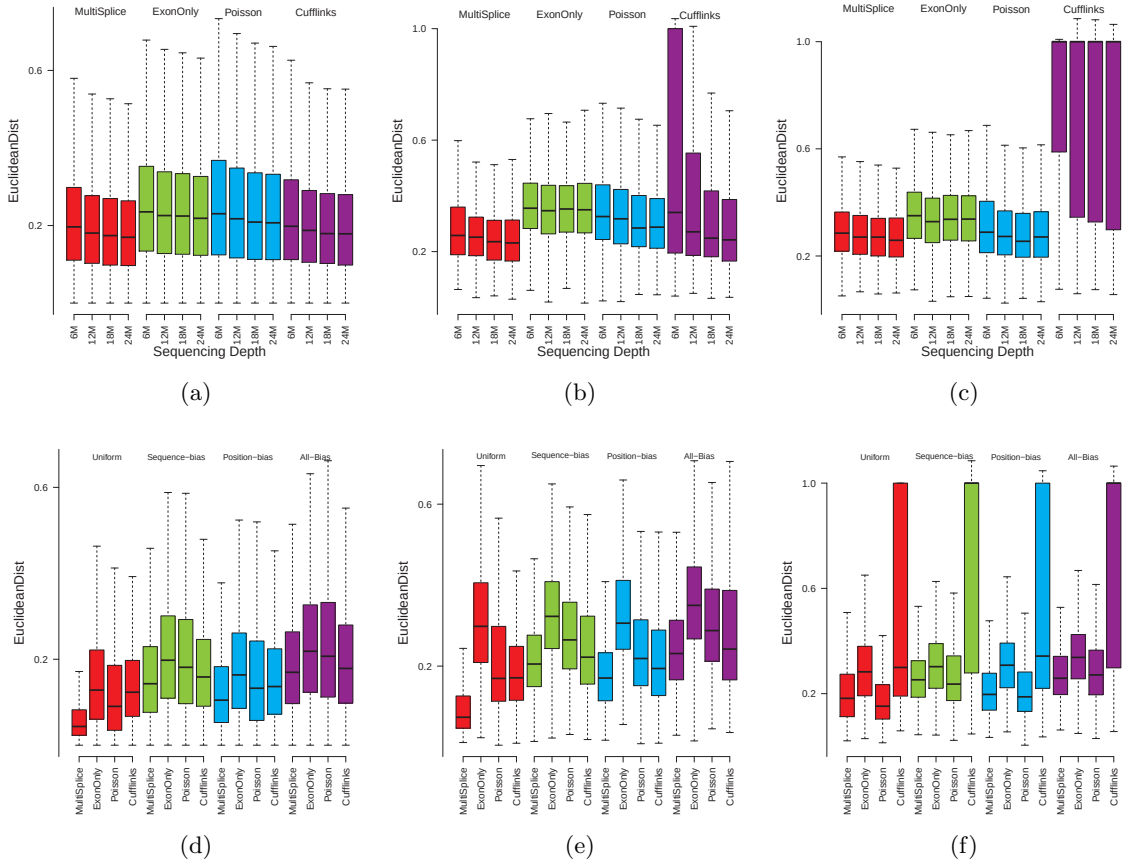
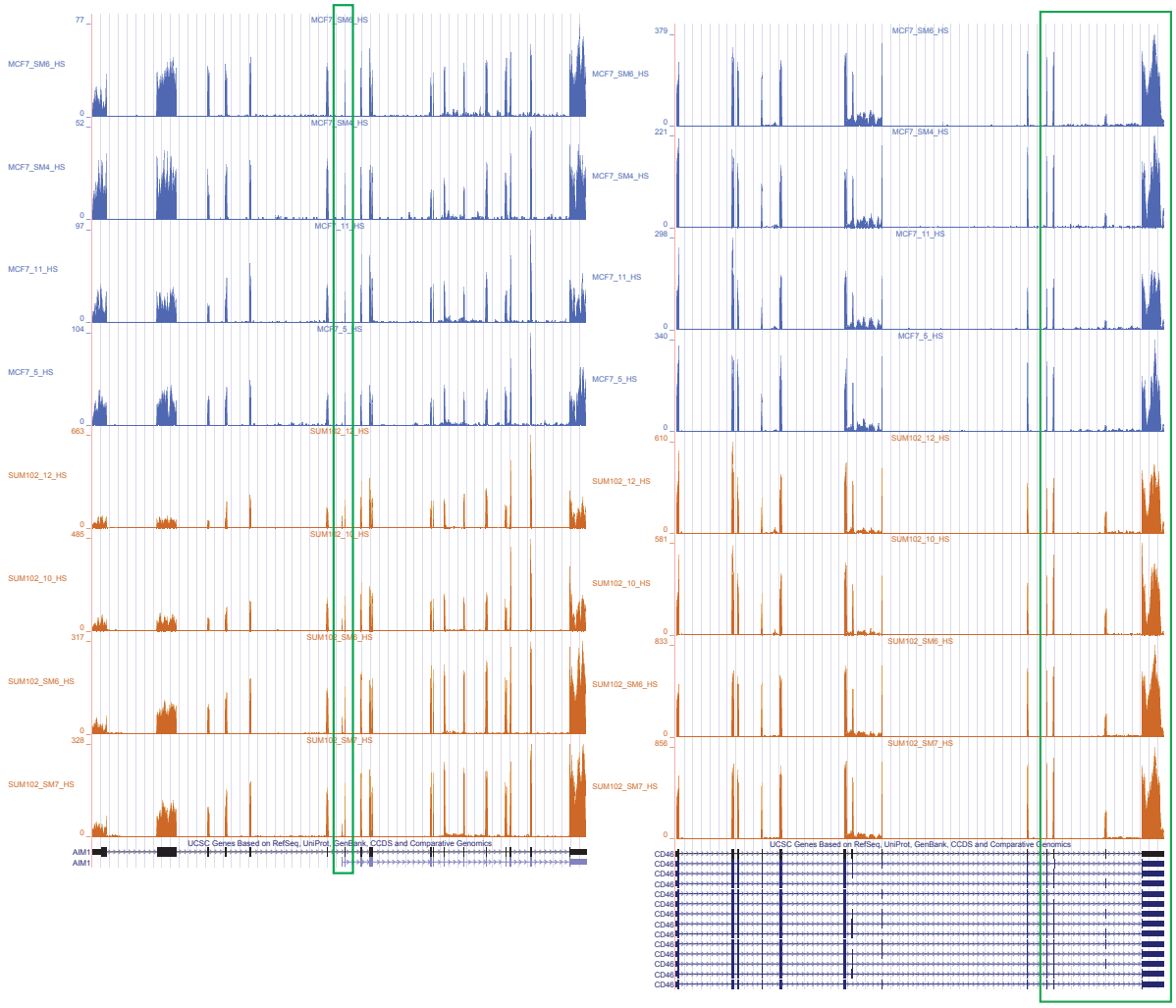


Fig. 7: a-c. Boxplots of the Euclidean distance between estimated transcript proportions and the ground truth under varying number of sampled reads: 6M, 12M, 18M and 24M over a total of 14530 human genes with more than one isoforms. (a),(b) and (c) correspond to the gene set that is identifiable with basic exon structure, identifiable with additional MultiSplice features, and unidentifiable, respectively. d-f: Boxplots of the Euclidean distance between estimated transcript proportions and the ground truth under four circumstances: uniform sampling, sampling with positional bias only, with sequence bias only and with all bias. (d),(e) and (f) correspond to the gene set that is identifiable with basic exon structure, identifiable with additional MultiSplice features, and unidentifiable, respectively.



(a) (b)

Fig. 8: a. The coverage plot of Gene AIM1 in all 8 breast cancer cell line samples. Please note the first exon of the second isoform is barely expressed MCF-7 but its expression significantly increased in the SUM-102 samples. b. The coverage plot of Gene CD46. The exon-skipping event on the 13th exon has been confirmed by qRT-PCR.