Statistics and Its Interface Volume 8 (2015) 437–445

A penalized likelihood approach for robust estimation of isoform expression

Hui Jiang* and Julia Salzman*

Ultra high-throughput sequencing of transcriptomes (RNA-Seq) has enabled the accurate estimation of gene expression at individual isoform level. However, systematic biases introduced during the sequencing and mapping processes as well as incompleteness of the transcript annotation databases may cause the estimates of isoform abundances to be unreliable, and in some cases, highly inaccurate. This paper introduces a penalized likelihood approach to detect and correct for such biases in a robust manner. Our model extends those previously proposed by introducing bias parameters for reads. An L1 penalty is used for the selection of non-zero bias parameters. We introduce an efficient algorithm for model fitting and analyze the statistical properties of the proposed model. Our experimental studies on both simulated and real datasets suggest that the model has the potential to improve isoform-specific gene expression estimates and identify incompletely annotated gene models.

Keywords and phrases: Penalized likelihood, Robust estimation, RNA-Seq, Isoform expression.

1. INTRODUCTION

In eukaryotes, a single gene can often produce more than one distinct transcript isoforms, through an important cell mechanism called alternative splicing. Alternative splicing can greatly enrich the diversity of eukaryote transcriptomes (Wang et al., 2008), especially in developmental and differentiation programs, and can contribute to disease when it is dysregulated (López-Bigas et al., 2005). Study gene expression at specific transcript isoform level is therefore of great importance and interest to biologists.

Ultra high-throughput sequencing of transcriptomes (RNA-Seq) has enabled the accurate estimation of gene expression at individual isoform level (Wang et al., 2008). As of today, modern ultra high-throughput sequencing platforms can generate tens of millions of short sequencing reads from prepared RNA samples in less than a day. For these reasons, RNA-Seq has become the method of choice for assays of gene expression. To analyze increasing amounts of data generated from biological experiments, a number of statistical models and software tools have been developed (Jiang and Wong, 2009; Trapnell et al., 2010; Li and Dewey, 2011). For a review of the methods for transcript quantification using RNA-Seq, see Pachter (2011).

Although these methods have achieved great success in quantifying isoforms accurately, there are still many remaining challenging issues which may hinder their wider adoption and successful application by biologists. Systematic biases introduced during the sequencing and mapping processes (Li, Jiang and Wong, 2010; Hansen, Brenner and Dudoit, 2010; Roberts et al., 2011) can cause the estimates of isoform abundances to be unreliable. Furthermore, recently there have been periods of time where hundreds of new transcripts are discovered every month (Pruitt et al., 2009; Harrow et al., 2012; Karolchik et al., 2014), including occasional examples of thousands of new isoforms being identified in a single study (Salzman et al., 2012, 2013). These incomplete annotations can also lead to unreliable estimates of isoform abundances (Black Pyrkosz, Cheng and Titus Brown, 2013).

This paper introduces a penalized likelihood approach to detect and correct for such biases in a robust manner. Bias parameters are introduced for read abundance, and an L1 penalty is used for the selection of non-zero bias parameters. We introduce an efficient algorithm for fitting this model and analyze its statistical properties. Our experimental studies on both simulated and real datasets show that transcript estimates can be highly sensitive to including or omitting parameters modeling read bias. Together, our results suggest that this method has the potential to improve isoform-specific gene expression estimates and improve annotation of existing gene models.

2. A PENALIZED LIKELIHOOD APPROACH

2.1 The model

We adopt the notation and extend the model in Salzman, Jiang and Wong (2011), which provides a flexible statistical framework for modeling both single-end and paired-end RNA-Seq data, including insert length distributions. To state the model, for a gene g with I annotated distinct transcript isoforms, suppose that the sequencing reads from g are sampled from J possible distinct read types. A read type refers to a group of single-end reads that are mapped to a specific position in a transcript in single-end

arXiv: 1310.0379
*Corresponding author.

sequencing, or a group of paired-end reads that are mapped to a specific fragment in a transcript in paired-end sequencing (Salzman, Jiang and Wong, 2011). We use θ to denote the $I\times 1$ vector representing isoform abundances in the sample and A to denote the $I \times J$ sampling rate matrix with its (i,j)-th element a_{ij} denoting the rate that read type j is sampled from isoform i. Given θ and A, we assume that the $J \times 1$ read count vector n, where n_j denotes the number of reads of type j mapped to any of the I isoforms, follows a Poisson distribution

$$n_j|\theta, A \sim \text{Poisson}\left(\sum_{i=1}^{I}\theta_i a_{ij}\right).$$

The log-likelihood function is therefore

$$l(\theta; n, A) = \sum_{j=1}^{J}\left\{n_j \ln\left(\sum_{i=1}^{I}\theta_i a_{ij}\right) - \sum_{i=1}^{I}\theta_i a_{ij}\right\}, \tag{2.1}$$

where the term $-\ln(n_j!)$ was dropped because it does not contain θ. Model (2.1) is essentially a generalized linear model with Poisson distribution and identity link function, as well as constraints that $\theta_i \geq 0$ for all i. This model is identifiable when $J \geq I$ and A is full rank. It can be shown that the log-likelihood function is concave in θ (Jiang and Wong, 2009) and the MLE can be estimated using either constrained Newton-Raphson or an EM algorithm.

In (Salzman, Jiang and Wong, 2011), the sampling rate matrix A is a set of parameters, assumed to be a known function of the sequencing library and the gene of interest. For single-end RNA-Seq data, the simplest model is to assume uniform sampling and let $a_{ij} = N$ where N is the sequencing depth (proportional to total number of mapped reads) of the experiment if isoform i can generate read type j or let $a_{ij} = 0$ otherwise. For paired-end RNA-Seq data, an insert length model can be assumed such that $a_{ij} = q(l_{ij})N$ if read type j can be mapped to isoform i with insert length (fragment length) l_{ij}, where $q(\cdot)$ is the empirical probability mass based on all the mapped read pairs. Salzman, Jiang and Wong (2011) discusses these sampling rate models in more details.

Although these simplified sampling rate models usually work well in practice, there are systematic biases introduced during the sequencing and mapping processes which may cause biased estimates of the sampling rates and consequently biased estimates of isoform abundances. Several approaches have been developed to model and correct for such biases (Li, Jiang and Wong, 2010; Hansen, Brenner and Dudoit, 2010; Roberts et al., 2011). However, completely removing sampling biases is almost impossible because the technical process of sequencing and read mapping is often too complex to model. Including all possible transcript isoforms (de novo identification) also poses computational challenges and biases. Using all annotated transcripts in the model, many times exceeeding 10 per gene, can introduce non-identifiability of isoforms. However, while the vast majority of human genes have multiple annotated (and likely unannotated) transcripts, most cell types, or single cells, express only a subset of annotated transcripts.

To explore statistical approaches that could improve transcript quantification with RNA-Seq, we present a flexible model to account for all different kinds of biases in estimated sampling rates. We assign a bias parameter β_j to each read type j and reparametrize β_j as $\beta_j = e^{b_j}$ to constrain $\beta_j > 0$. When $\beta_j = 1$ (i.e., $b_j = 0$), there is no bias for read type j. The actual effective sampling rate for read type j from isoform i now becomes $a'_{ij} = a_{ij}\beta_j = a_{ij}e^{b_j}$, and the log-likelihood function is now

$$l(\theta, b; n, A) = \sum_{j=1}^{J}\left\{n_j \ln\left(\sum_{i=1}^{I}\theta_i a_{ij}e^{b_j}\right) - \sum_{i=1}^{I}\theta_i a_{ij}e^{b_j}\right\}. \tag{2.2}$$

Since the number of observations is J, which is smaller than the number of variables $I + J$ in model (2.2), model (2.2) is not identifiable. To solve this problem, we introduce a penalty $p(b)$ on the bias parameters b and formulate an L1-penalized log-likelihood

$$\begin{aligned} f(\theta, b) &= l(\theta, b; n, A) - p(b) \\ &= \sum_{j=1}^{J}\left\{n_j \ln\left(\sum_{i=1}^{I}\theta_i a_{ij}e^{b_j}\right) - \sum_{i=1}^{I}\theta_i a_{ij}e^{b_j}\right\} \\ &\quad - \lambda\sum_{j=1}^{J}|b_j| \end{aligned} \tag{2.3}$$

where $\lambda > 0$ is a tuning parameter. By choosing a large enough λ, all b_j's will be zero and model (2.3) will reduce to model (2.1). By allowing some b_j's to be non-zero, we will still have an identifiable model and in the meanwhile allow robust estimation of θ.

Introducing the L1 penalty shrinks b towards 0, consequently inflates θ compared to fitting a model with sparse but unbiased estimates of b. One way to reduce such bias in the estimation of θ is to use a two-step approach for model fitting: first fit the model with the L1 penalty, then fit the model without the L1 penalty while retaining only non-zero b_j's as model parameters. Clearly, to avoid nonidentifiable issues, the number of non-zero b_j's must be smaller than or equal to $J-I$, which can be achieved by increasing the tuning parameter λ. The statistical properties of the two-step approach is discussed in Section 2.3.

2.2 Optimization

In this section we develop an efficient algorithm for fitting model (2.3), i.e., maximizing the L1-penalized log-likelihood function $f(\theta, b)$.

Proposition 2.1. *The L1-penalized log-likelihood function $f(\theta, b)$ in model (2.3) is biconcave.*

Because $f(\theta, b)$ is biconcave, we use Alternative Concave Search (ACS) to estimate θ and b, by alternatively fixing one of them and optimizing for the other. The sequence of function values generated by the ACS algorithm is monotonically increasing and $f(\theta, b)$ is bounded from above (because it is a penalized log-likelihood function), which guarantees convergence of the ACS algorithm.

Algorithm 2.2. *With b fixed, θ can be estimated with the following EM algorithm*

$$\textit{E-step:} \quad \hat{n}_{ij}^{(k+1)} := \mathbf{E}\left(n_{ij}|n, A, b, \hat{\theta}^{(k)}\right) = \frac{n_j \hat{\theta}_i^{(k)} a_{ij}}{\sum_{i=1}^{I} \hat{\theta}_i^{(k)} a_{ij}}$$

$$\textit{M-step:} \quad \hat{\theta}_i^{(k+1)} = \frac{\sum_{j=1}^{J} \hat{n}_{ij}^{(k+1)}}{\sum_{j=1}^{J} a_{ij} e^{b_j}}$$

Alternatively, θ can be estimated using the more efficient Newton-Raphson algorithm. In our implementation, we only execute one round of the EM iteration each time we optimize θ with b fixed.

Proposition 2.3. *With θ fixed, b_j can be estimated using the following closed-form formula*

$$\hat{b}_j = \ln\left(1 + \frac{S_\lambda(n_j - \sum_{i=1}^{I} \theta_i a_{ij})}{\sum_{i=1}^{I} \theta_i a_{ij}}\right) \tag{2.4}$$

where $S_\lambda(x) = sign(x)(|x| - \lambda)_+$ is the soft thresholding operator, where $(x)_+ = \max(x, 0)$.

For more efficient convergence, we use an analytical property of values of $\hat{\theta}$ and $\hat{b}$ which maximize $f(\theta, b)$:

Proposition 2.4. *There is at least one set of $\hat{\theta}$ and $\hat{b}$ which maximize $f(\theta, b)$ such that median$(\hat{b}_1, \ldots, \hat{b}_J) = 0$. Furthermore, under the assumption that $\hat{b}_j \geq 0$ for all $1 \leq j \leq J$, there must exist at least one j such that $\hat{b}_j = 0$.*

Accordingly, after each iteration which solves (2.4), we include a step which centers the $\hat{b}_j$'s around their median, i.e., $\hat{b}'_j = \hat{b}_j - \text{median}(\hat{b}_1, \ldots, \hat{b}_J)$.

2.3 Statistical properties

Several statistical properties of the two-step approach introduced in Section 2.1 are provided in this section. First, we state an intuitive interpretation of the procedure:

Proposition 2.5. *Fitting model (2.3) using the two-step approach introduced in Section 2.1 is equivalent to fitting model (2.1) after removing all the observed n_j's whose corresponding $\hat{b}_j$'s are non-zero. In another word, the two model-fitting steps essentially perform outlier detection and removal, respectively.*

This observation can be extended to prove that in the case of $I = 1$, with an appropriately chosen λ, the two-step procedure yields a consistent estimate of θ under the assumption that $b_j \geq 0$ for all j. While $I = 1$ may appear to be a trivial case, in fact, this approach is equivalent to considering a subset of the full model introduced in Salzman, Jiang and Wong (2011) where each read is considered if and only if it can be generated by exactly one isoform. Reasonable statistical power can be achieved with this approach, and it is of relatively wide use by biologists.

For convenience, the proposition and proof of Proposition 2.6 are stated for the case where $a_{1j} = N$, but this assumption can be relaxed to allow a_{1j} to be arbitrary. Also, from the proof, it is clear that choosing λ larger than $(\max_j n_j)^{1/2}$ also results in consistent estimates, but perhaps unnecessarily sparse models.

Proposition 2.6. *Under the assumptions that $I = 1$, $\lambda = (\max_j n_j)^{1/2}$, $a_{1j} = N$ and $b_j \geq 0$ for all j, the two-step approach yields a consistent estimate of θ.*

3. EXPERIMENTS

3.1 Simulations

In this section, we use simulation to study our model in various gene structures, relative isoform abundances, bias patterns and sequencing depths. For each simulation replicate, we estimate θ and b using three approaches as follows and compare their estimation accuracies:

1. The conventional approach (Jiang and Wong, 2009; Salzman, Jiang and Wong, 2011) with no bias correction (i.e., fix $b = 0$).
2. Our proposed one-step approach with bias correction (i.e., without doing the second step of estimation θ introduced in Section 2.1).
3. Our proposed two-step approach with bias correction.

Throughout the simulations, we choose $\lambda = (\max_j n_j)^{1/2}$ because of the consistency result (Proposition 2.6) we obtain with this choice.

Example 3.1. *We simulate the case that a gene has a single annotated isoform (i.e., $I = 1$), 5 read types after collapsing (i.e., $J = 5$, e.g., the gene has 5 exons). Suppose that the sampling rate matrix $A = NC$, where $C = (1, 1, 1, 1, 1)$ are the relative sampling rates for the five exons (e.g., each exon has the same length of 1,000 bp), and N is the relative sequencing depth (e.g., $N = 10$ in Table 1 means that there are a total of 10M single-end reads sequenced from the RNA-Seq experiment. We assume that the true parameters are $\theta = 1$ and $b = (2, 0, 0, 0, 0)^T$.*

For Example 3.1 we simulate 100 replicates where read counts for each read type j ($j = 1, \ldots, 5$) are simulated as i.i.d Poisson r.v. with parameter $\theta N C_j e^{b_j}$ where b_j is the bias parameter for read type j, and report the average (and standard deviation) of the estimation errors of θ in L2 distance in Table 1. We also report the average (and standard deviation) of the number of b_j's that are misidentified as zero

Table 1. Estimation accuracy of Example 3.1. Average of 100 replicates, standard deviation reported in parentheses

Sequencing Depth	No Bias Correction	Bias Correction (1-step)	Bias Correction (2-step)	#Misidentified
10	1.32 (0.2)	0.24 (0.14)	0.13 (0.1)	0.03 (0.17)
100	1.26 (0.06)	0.07 (0.04)	0.04 (0.04)	0.09 (0.29)
1000	1.28 (0.02)	0.02 (0.01)	0.01 (0.01)	0.05 (0.22)

Table 2. Estimation accuracy of Example 3.2. Average of 100 replicates, standard deviation reported in parentheses

Sequencing Depth	No Bias Correction	Bias Correction (1-step)	Bias Correction (2-step)	#Misientified
10	3.78 (0.23)	3.41 (1.85)	3.02 (1.71)	0.13 (0.34)
100	3.76 (0.08)	1.82 (1.28)	1.4 (0.9)	0.01 (0.1)
1000	3.77 (0.03)	0.45 (0.32)	0.36 (0.26)	0 (0)

Table 3. Estimation accuracy of Example 3.3. Average of 100 replicates, standard deviation reported in parentheses

Sequencing Depth	No Bias Correction	Bias Correction (1-step)	Bias Correction (2-step)	#Misientified
10	76.8 (471.72)	1.22 (1.42)	0.93 (0.69)	2.17 (1.53)
100	7792.35 (75161.75)	2.56 (17.7)	0.41 (0.41)	2.2 (1.57)
1000	406.35 (1934.52)	0.42 (1)	0.18 (0.41)	2 (1.68)

vs. non-zero. Table 1 shows empirical results confirming our theory: if some $b_j > 0$, without bias correction, θ will not be estimated consistently. While both one-step and two-step approaches achieve consistent estimates of θ, the two-step approach is more efficient. On average, we misidentify less than one nonzero b's.

Example 3.2. *We simulate the case with $I = 2$, $J = 6$ and $C = (1, 2, 1, 2, 3, 2; 1, 2, 0, 2, 3, 2)$, e.g., a gene with six exons and two isoforms differ by the inclusion/exclusion of the third exon. We assume that the true parameters are $\theta = (6, 3)^T$ and $b = (-5, 0, 0, 0, 0, 0)^T$.*

For Example 3.2 we simulate 100 replicates where read counts for each read type j ($j = 1, \ldots, 6$) are simulated as i.i.d Poisson r.v. with parameter $\sum_{i=1}^{2} \theta_i N C_{ij} e^{b_j}$ where b_j is the bias parameter for read type j. The simulation and estimation results for Example 3.2 are shown in Table 2. The performance of the three approaches is similar to that in Example 3.1.

Example 3.3. *We now consider a case with $I = 5$, $J = 20$. For each replicate of the simulation, we randomly generate each element of C as $c_{ij} = I_{u_1<0.1}0 + I_{u_1>=0.1}\, Uniform(0, 1)$, where $u_1 \sim Uniform(0, 1)$. We also randomly generate each element of the true parameters θ and b as $\theta_i \sim Exponential(1)$ and $b_j = I_{u_2<0.9}0 + I_{u_2>=0.9}N(0, 3)$ where $u_2 \sim Uniform(0,1)$.*

For Example 3.3 we simulate 100 replicates where read counts for each read type j ($j = 1, \ldots, 20$) are simulated as i.i.d Poisson r.v. with parameter $\sum_{i=1}^{5} \theta_i N C_{ij} e^{b_j}$ where θ_i, C_{ij} and b_j are a randomly generated expression level, sampling rate and bias parameter as described above. The simulation and estimation results for Example 3.3 are shown in Table 3. The performance of the three approaches is similar to that in Examples 3.1 and 3.2. In particular, the approach without bias correction introduces a huge estimation error in some of the cases (e.g., when b_j is large and a_{ij} is small).

3.2 Real data analysis

We evaluated our model using real RNA-Seq data from the Gm12878 cell line generated by the ENCODE project (ENCODE Project Consortium et al., 2012). A total of 415,630 single-end reads of 75 bp mapped to human chromosome 22 are used in the analysis. We use RefSeq human annotation database (Pruitt et al., 2009) for our analysis. We ran both the conventional approach (without bias correction) and our proposed one-step approach (with bias correction) on this data set. 579 genes have estimated expression level ≥ 1 using the RPKM unit (Mortazavi et al., 2008), and 65 of the 579 genes have at least 2-fold change in their gene expression estimates between the approaches with and without bias correction.

MED15 is an example of a gene with greater than 2-fold change in the total expression of two isoforms with and without bias correction, shown in Figure 1. The center part of the gene has a much greater read density than the 5' or 3' ends. Without bias correction, MED15's expression is estimated as 1487.11 RPKM (with the two isoforms estimated as 54.89 RPKM and 1432.22 RPKM, respectively). Our bias correction approach identifies this bias and down-weights the contribution of reads from the center part of the gene. Consequently, it estimates the gene expression as 702.52 RPKM (with the two isoforms estimated as 53.65 RPKM and 648.87 RPKM, respectively). Another example (IGLL5) is also shown in Figure 1. Many other genes also show a similar pattern of biases.

The observed biases in these genes could be due to mapping artifacts, or preferential amplification of portions of the

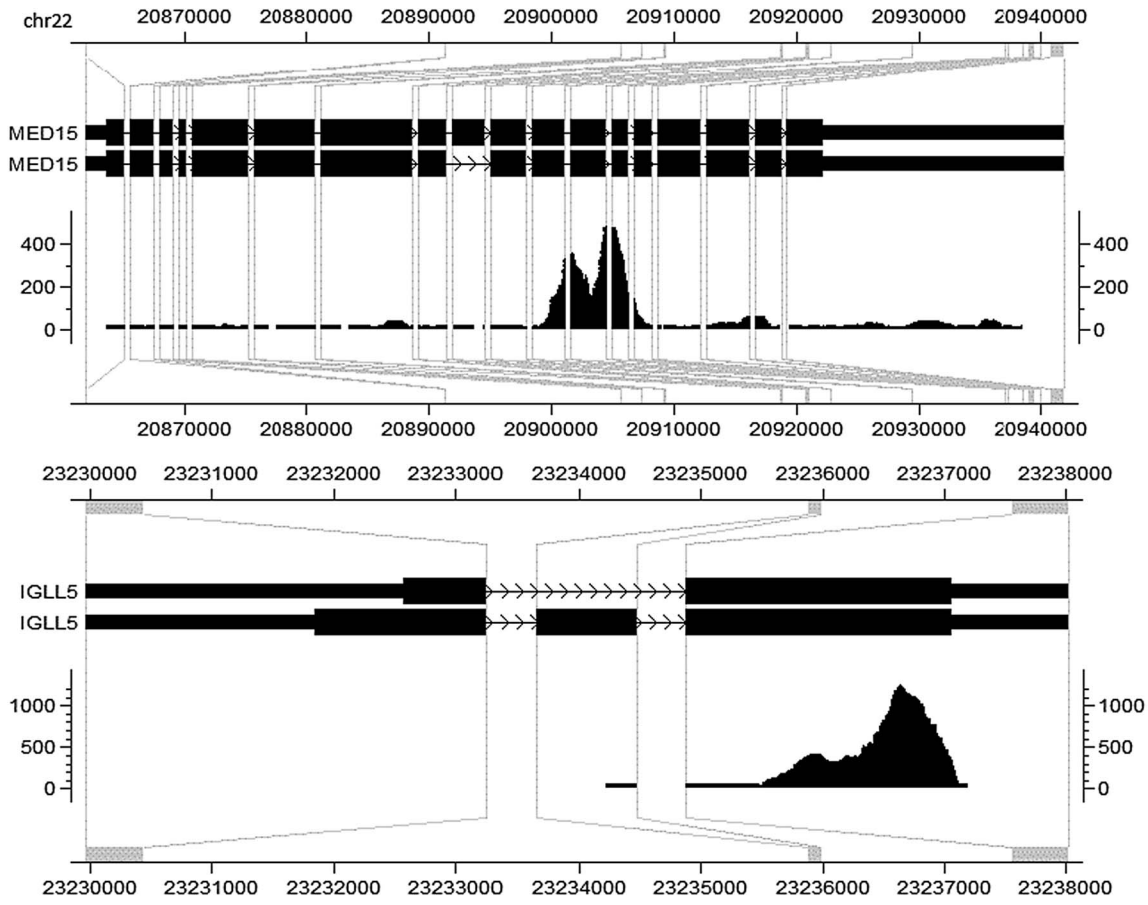

Figure 1. Visualization of RNA-Seq reads mapped to the genes MED15 and IGLL5 on human chromosome 22 in CisGenome Browser (Jiang et al., 2010). From top to bottom for each gene: genomic coordinates, gene structure where exons are magnified for better visualization, coverage of mapped reads. Reads are 75 bp single-end.

gene during RNA-Seq library preparation. Further investigation, including experimental testing may be required to determine if either of these explanations for increased read density are explanatory. Another explanation could be that the gene model used for our experiment, which includes just two isoforms, is incomplete. For example, the observed increased read density could be due to expression of other isoforms of MED15 that include these regions.

4. DISCUSSION

In this paper we choose $\lambda = (\max_j n_j)^{1/2}$, which seems to work reasonably well with both simulated and real data and which we have shown to produce consistent estimates of θ under reasonable assumptions. We believe that more research on statistical properties of different choices of λ may lead to improvement of our model in applied settings. For example, we plan to evaluate a standard approach of choosing λ by cross-validation, although it comes at the cost of more intensive computation. Also, as our proof of consistency shows, choosing values of λ larger than $(\max_j n_j)^{1/2}$ will also yield consistent estimators of θ under the regime analyzed in Proposition 2.6.

To implement our proposed model, because J (the number of distinct read types) is usually very large, especially for paired-end RNA-Seq data, we adopt the collapsing technique introduced in Salzman, Jiang and Wong (2011) and merge read types of proportional sampling rate vectors into read categories (which are minimal sufficient statistics of the model). For instance, in our simulations we group read types from the same exon as read categories. This does not change the model (2.3) except that j now represents a read category rather than a read type. Therefore, in this paper the terms read type and read category are used interchangeably. Salzman, Jiang and Wong (2011) also introduced another data reduction technique which ignores all the read categories with zero read counts by introducing an additional term with the total sampling rates for each isoform $w_i = \sum_{j=1}^{J} a_{ij}$. In this case, the log-likelihood function becomes

$$l(\theta; n, A, w) = \sum_{n_j>0} \left\{ n_j \ln \left(\sum_{i=1}^{I} \theta_i a_{ij} \right) \right\} - \sum_{i=1}^{I} \theta_i w_i. \quad (4.1)$$

For simplicity, we do not discuss model (4.1) in this paper but our approach easily extends to deal with model (4.1).

From (2.4), it is apparent that larger n_j's are relatively less affected by the soft thresholding operator than smaller n_j's. Intuitively, the proposed approach works better when the read categories are of similar group sizes. In our real data experiment, we collapsed reads into exons and junctions to roughly fulfill this condition, and simulation demonstrates that our proposed approach is not very sensitive to how collapsing is performed. An alternative approach is to use n_j as the weight for the corresponding b_j, i.e., by letting $p(b) = \lambda \sum_{j=1}^{J} n_j |b_j|$. All the statistical properties and optimization techniques introduced in the paper can be adapted to this new penalty function with only minor modifications. In simulations (results not shown here), this new penalty function does not perform noticeably better than the current penalty function. There may be advantages and disadvantages to increase penalties for biases corresponding to larger n_j's.

Although the two-step approach appears to be slightly more efficient than the one-step approach in our simulations, it has several critical drawbacks: 1) It requires an increase in computation up to a factor of two; 2) It may introduce a non-identifiable issue in the second step of estimation when the number of nonzero b's identified in the first step of estimation is large; and 3) It makes parameter estimates sensitive to λ. Therefore, we use the one-step approach in our real data experiment and we plan to study the two-step approach in more detail in future work.

The example of MED15 highlights another use of fitting bias parameters. First, in the presence of unannotated isoforms of a gene, correcting for bias in read sampling may be correcting for real biological confounding. In such scenarios, simulation suggests that correcting for bias improves model fit and quantification conditional on the gene models used for the study. For example, the two transcripts of MED15 in Figure 1 are probably more realistically estimated by our bias-corrected model. In addition, screening genes with large estimated bias parameters may be a tool for identifying unannotated transcripts or incomplete models used in the mapping step.

Finally, the approach introduced in this paper is adapted and stated for the isoform expression estimation problem, which is essentially a Poisson regression model with identity link function. Similar ideas have been proposed recently for linear regression (She and Owen, 2011), logistic regression (Tibshirani and Manning, 2013) and unsupervised learning (Witten, 2013). We believe that it may be possible to generalize our approach to other models and other practical applications may exist as well.

ACKNOWLEDGEMENTS

HJ's research was supported in part by an NIH grant 5U54CA163059-02 and a GAPPS Grant from the Bill & Melinda Gates Foundation. JS was supported by NIH grant 1K99CA16898701 from the NCI.

APPENDIX A. APPENDIX SECTION

Proof of Proposition 2.1.

$$\begin{aligned} f(\theta, b) &= \sum_{j=1}^{J} \left\{ n_j \ln \left(\sum_{i=1}^{I} \theta_i a_{ij} e^{b_j} \right) - \sum_{i=1}^{I} \theta_i a_{ij} e^{b_j} \right\} \\ &\quad - \lambda \sum_{j=1}^{J} |b_j| \\ &= \sum_{j=1}^{J} \left\{ n_j \ln \left(\sum_{i=1}^{I} \theta_i a_{ij} \right) \right\} + \sum_{j=1}^{J} n_j b_j \\ &\quad - \sum_{j=1}^{J} \sum_{i=1}^{I} \theta_i a_{ij} e^{b_j} - \sum_{j=1}^{J} \lambda |b_j| \end{aligned}$$

where $n_j b_j$ and $-\lambda|b_j|$ are concave, $n_j \ln(\sum_{i=1}^{I} \theta_i a_{ij})$ is concave because $-\sum_{i=1}^{I} \theta_i a_{ij}$ is concave and $\ln(\cdot)$ is concave and non-decreasing, and $-\theta_i a_{ij} e^{b_j}$ is biconcave because both θ_i and e^{b_j} are convex. □

Proof of Proposition 2.3. Fixing θ, since the L1-penalty is decomposable, $f(b)$ can be written as the sum of J terms $f(b) = \sum_{j=1}^{J} f_j(\theta, b_j)$, where

$$\begin{aligned} f_j(b_j) &= n_j \ln \left(\sum_{i=1}^{I} \theta_i a_{ij} e^{b_j} \right) - \sum_{i=1}^{I} \theta_i a_{ij} e^{b_j} - \lambda |b_j| \\ &= n_j b_j + n_j \ln \left(\sum_{i=1}^{I} \theta_i a_{ij} \right) - e^{b_j} \sum_{i=1}^{I} \theta_i a_{ij} - \lambda |b_j| \end{aligned}$$

Therefore, $b_j = \text{argmax}_{b_j} f_j(\theta, b_j)$. Note that the second term of $f_j(b_j)$ does not contain b_j. Since $|\cdot|$ is non-differentiable, we take the subdifferential of f_j at b_j

$$\partial f_j(b_j) = n_j - e^{b_j} \sum_{i=1}^{I} \theta_i a_{ij} - \lambda s_j$$

where $s_j = \text{sign}(b_j)$ if $b_j \neq 0$ and $s_j \in [-1, 1]$ if $b_j = 0$. It can be verified that (2.4) is the solution to the equation of $\partial f_j(b_j) = 0$. □

Proof of Proposition 2.4. Suppose $\hat{\theta}$ and $\hat{b}$ are such that $(\hat{\theta}, \hat{b}) = \text{argmax}_{(\theta, b)} f(\theta, b)$. Let $\hat{b}'_j = \hat{b}_j - m$ and $\hat{\theta}'_i = \hat{\theta}_i e^m$, where $m = \text{median}(\hat{b}_1, \ldots, \hat{b}_J)$, then

$$\begin{aligned} f(\hat{\theta}', \hat{b}') &= \sum_{j=1}^{J} \left\{ n_j \ln \left(\sum_{i=1}^{I} \hat{\theta}'_i a_{ij} e^{\hat{b}'_j} \right) - \sum_{i=1}^{I} \hat{\theta}'_i a_{ij} e^{\hat{b}'_j} \right\} \\ &\quad - \lambda \sum_{j=1}^{J} |\hat{b}'_j| \end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{J}\left\{ n_j \ln\left(\sum_{i=1}^{I}\hat{\theta}_i e^{m} a_{ij} e^{\hat{b}_j - m}\right)\right. \\
&\quad \left. - \sum_{i=1}^{I}\hat{\theta}_i e^{m} a_{ij} e^{\hat{b}_j - m}\right\} - \lambda\sum_{j=1}^{J}|\hat{b}_j - m| \\
&= \sum_{j=1}^{J}\left\{ n_j \ln\left(\sum_{i=1}^{I}\hat{\theta}_i a_{ij} e^{b_j}\right) - \sum_{i=1}^{I}\hat{\theta}_i a_{ij} e^{\hat{b}_j}\right\} \\
&\quad - \lambda\sum_{j=1}^{J}|\hat{b}_j - m| \\
&\geq \sum_{j=1}^{J}\left\{ n_j \ln\left(\sum_{i=1}^{I}\hat{\theta}_i a_{ij} e^{\hat{b}_j}\right) - \sum_{i=1}^{I}\hat{\theta}_i a_{ij} e^{\hat{b}_j}\right\} \\
&\quad - \lambda\sum_{j=1}^{J}|\hat{b}_j| \\
&= f(\hat{\theta}, \hat{b})
\end{aligned}
$$

That is, $\hat{\theta}'$ and $\hat{b}'$ maximizes $f(\theta, b)$ with $\text{median}(\hat{b}'_1, \ldots, \hat{b}'_J) = 0$. Under the assumption that $\hat{b}_j \geq 0$ for all $1 \leq j \leq J$, equality holds in the above derivation only if at least one $\hat{b}_j = 0$. □

Proof of Proposition 2.5. Without loss of generality, assume $\hat{b}_j \neq 0, (j = 1, \ldots, k)$ and $\hat{b}_j = 0, (j = k+1, \ldots, J)$ after the first step of model fitting with the L1 penalty. In the second step of model fitting without the L1 penalty, we have $(\hat{\theta}, \hat{b}) = \text{argmax}_{(\theta,b)} f(\theta, b)$, where

$$
\begin{aligned}
& f(\theta, b) \\
&= \sum_{j=1}^{k}\left\{ n_j \ln\left(\sum_{i=1}^{I}\theta_i a_{ij} e^{b_j}\right) - \sum_{i=1}^{I}\theta_i a_{ij} e^{b_j}\right\} \\
&\quad + \sum_{j=k+1}^{J}\left\{ n_j \ln\left(\sum_{i=1}^{I}\theta_i a_{ij}\right) - \sum_{i=1}^{I}\theta_i a_{ij}\right\}.
\end{aligned} \tag{A.1}
$$

Solving

$$
\frac{\partial f(\theta, b)}{\partial b_j} = n_j - \sum_{i=1}^{I}\theta_i a_{ij} e^{b_j} = 0
$$

we have

$$
b_j = \log\left(\frac{n_j}{\sum_{i=1}^{I}\theta_i a_{ij}}\right). \tag{A.2}
$$

Plugging (A.2) into (A.1), we have

$$
\begin{aligned}
f(\theta, b) &= \sum_{j=1}^{k}(n_j \ln n_j - n_j) \\
&\quad + \sum_{j=k+1}^{J}\left\{ n_j \ln\left(\sum_{i=1}^{I}\theta_i a_{ij}\right) - \sum_{i=1}^{I}\theta_i a_{ij}\right\}
\end{aligned}
$$

therefore $\hat{\theta} = \text{argmax}_\theta f'(\theta)$ where

$$
f'(\theta) = \sum_{j=k+1}^{J}\left\{ n_j \ln\left(\sum_{i=1}^{I}\theta_i a_{ij}\right) - \sum_{i=1}^{I}\theta_i a_{ij}\right\}
$$

is exactly the log-likelihood after removing all the observation n_j's whose corresponding b_j's are non-zero. □

Proof of Proposition 2.6

Without loss of generality, throughout this section, assume k and J are fixed with $1 < k < J$ and

$$
b_1 = b_2, \ldots, b_k = 0 < b_{k+1} \leq b_{k+2} \leq \ldots b_J.
$$

The key to the proof of Proposition 2.6 is the following Lemma:

Lemma A.1 (Main Lemma)**.** *As $N \to \infty$, $1_{\hat{b_j}=0}$ for some $_{k<j\leq J} \to 0$ a.s.*

The proof requires the following:

Lemma A.2 (Rearrangement inequality)**.** *For any $1 \leq i, j \leq J$ with $n_i < n_j$, if $\hat{\theta}, \hat{b}$ maximize (2.3), $\hat{b_i} \leq \hat{b_j}$.*

Proof of Lemma A.2. The proof is by contradition. Suppose that the likelihood (2.3) is maximized and for some i and j with $n_i < n_j$, $\hat{b_i} > \hat{b_j}$. Let τ be the transposition (i, j). Then, the rearrangement inequality implies that $f(\theta, b) < f(\theta, b_\tau)$. □

Lemma A.3. $\sum_{i,j\ i\leq k, k<j\leq J} 1(n_i > n_j) \to 0$ *a.s.*

Proof of Lemma A.3. If $i \leq k$ and $k < j$, the CLT implies

$$
1(n_i > n_j) \to 0 \text{ a.s.}
$$

as $N \to \infty$ which completes the proof since J is finite. □

Lemma A.4. *Let*

$$
\begin{aligned}
\Omega &= \{\sum_{i\leq k} 1(|n_i - N\theta(1 + e^{b_{k+1}})| > \max_{k<j\leq J}\sqrt{n_J})\} \\
&\quad \cap \{\sum_{i,j\ i\leq k, k<j\leq J} 1(n_i < n_j)\}.
\end{aligned}
$$

Then, $\Omega \to 1$ a.s.

Proof of Lemma A.4. Since $n_1 = Po(N\theta)$, and $b_{k+1} > 0$, the CLT and Lemma A.3 imply $\Omega \to 1$ *a.s.* □

Proof of Lemma A.1. The proof is by contradiction. As above, the dependence of the point estimates of parameters N is repressed.

Suppose

$$
1_{\hat{b_j}=0} \text{ for some } k < j \leq J \qquad i.o.
$$

Since J is not growing with N, without loss of generality, we can assume that for a fixed set $A \subset \{k+1 \dots J\}$,

$$\cup_{j \in A} 1_{\{\hat{b_j}=0\}} \qquad i.o.$$

Then, Lemma A.4 implies that

$$\cup_{j \in A} 1_{\{\hat{b_j}=0\}} \qquad i.o. \text{ on } \Omega.$$

Let Ω_i denote the infinite subsequence of events where

$$\cup_{j \in A} 1_{\{\hat{b_j}=0\}}$$

On the events Ω_i, Proposition 2.5 implies that

$$\hat{\theta} \to \theta(1 + \sum_{j \in A} e^{b_j}).$$

Therefore, on Ω_i, for all j with $1 \leq j \leq k$, Lemma A.4 implies

$$\hat{b_j} = \ln\left(1 + \frac{S_\lambda(n_j - \hat{\theta}N)}{\hat{\theta}N}\right) > 0$$

Now Lemma A.2 and Lemma A.4 imply that on Ω_i, the MLE must satisfy $\hat{b}_i > 0$ for all $1 \leq i \leq J$, which contradicts Proposition 2.4. □

Proof of Proposition 2.6. Lemma A.1 states that as $N \to \infty$,

$$1_{\hat{b_j}=0 \text{ for some } k < j \leq J} \to 0 \ a.s.$$

On the complement of this event, Proposition 2.5 implies $\hat{\theta} \to \theta$ a.s. □

Received 1 October 2013

REFERENCES

Black Pyrkosz, A., Cheng, H. and Titus Brown, C. (2013). RNA-Seq Mapping Errors When Using Incomplete Reference Transcriptomes of Vertebrates. *ArXiv e-prints*.

ENCODE Project Consortium, Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C. and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489** 57–74.

Hansen, K. D., Brenner, S. E. and Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res* **38** e131.

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J. M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigó, R. and Hubbard, T. J. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22** 1760–1774.

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

Jiang, H., Wang, F., Dyer, N. P. and Wong, W. H. (2010). CisGenome Browser: a flexible tool for genomic data visualization. *Bioinformatics* **26** 1781–1782.

Karolchik, D., Barber, G. P., Casper, J., Clawson, H., Cline, M. S., Diekhans, M., Dreszer, T. R., Fujita, P. A., Guruvadoo, L., Haeussler, M., Harte, R. A., Heitner, S., Hinrichs, A. S., Learned, K., Lee, B. T., Li, C. H., Raney, B. J., Rhead, B., Rosenbloom, K. R., Sloan, C. A., Speir, M. L., Zweig, A. S., Haussler, D., Kuhn, R. M. and Kent, W. J. (2014). The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* **42** D764–D770.

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, J., Jiang, H. and Wong, W. H. (2010). Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biol* **11** R50.

López-Bigas, N., Audit, B., Ouzounis, C., Parra, G. and Guigó, R. (2005). Are splicing mutations the most frequent cause of hereditary disease? *FEBS Lett* **579** 1900–1903.

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.

Pachter, L. (2011). Models for transcript quantification from RNA-Seq. *ArXiv e-prints*.

Pruitt, K. D., Tatusova, T., Klimke, W. and Maglott, D. R. (2009). NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res* **37** D32–D36.

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L. and Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol* **12** R22.

Salzman, J., Jiang, H. and Wong, W. H. (2011). Statistical modeling of RNA-Seq data. *Statistical Science* **26** 62–83. MR2849910

Salzman, J., Gawad, C., Wang, P. L., Lacayo, N. and Brown, P. O. (2012). Circular RNAs Are the Predominant Transcript Isoform from Hundreds of Human Genes in Diverse Cell Types. *PLoS ONE* **7** e30733.

Salzman, J., Chen, R. E., Olsen, M. N., Wang, P. L. and Brown, P. O. (2013). Cell-Type Specific Features of Circular RNA Expression. *PLoS Genet* **9** e1003777.

She, Y. and Owen, A. B. (2011). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association* **106**. MR2847975

Tibshirani, J. and Manning, C. D. (2013). Robust Logistic Regression using Shift Parameters. *arXiv preprint arXiv:1305.4987*.

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.

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.

Witten, D. M. (2013). Penalized unsupervised learning with outliers. *Statistics and its Interface* **6** 211. MR3066686

Hui Jiang
Department of Biostatistics
Center for Computational Medicine and Bioinformatics
University of Michigan
Ann Arbor, MI 48109
USA
E-mail address: jianghui@umich.edu

Julia Salzman
Department of Biochemistry
Stanford Cancer Institute
Stanford University
Stanford, CA 94305
USA
E-mail address: julia.salzman@stanford.edu