

OPEN ACCESS Full open access to this and thousands of other papers at http://www.la-press.com.

METHODOLOGY

On Differential Gene Expression Using RNA-Seq Data

Juhee Lee¹, Yuan Ji¹, Shoudan Liang², Guoshuai Cai² and Peter Müller³

¹Department of Biostatistics, UT M.D. Anderson Cancer Center Houston, Texas, USA. ²Department of Bioinformatics and Computational Biology, UT M.D. Anderson Cancer Center Houston, Texas, USA. ³Department of Mathematics, UT Austin, Austin, Texas, USA. Corresponding author email: yuanji@mdanderson.org

Abstract

Motivation: RNA-Seq is a novel technology that provides read counts of RNA fragments in each gene, including the mapped positions of each read within each gene. Besides many other applications it can be used to detect differentially expressed genes. Most published methods collapse the position-level read data into a single gene-specific expression measurement. Statistical inference proceeds by modeling these gene-level expression measurements.

Results: We present a Bayesian method of calling differential expression (BM-DE) that directly models the position-level read counts. We demonstrate the potential advantage of the BM-DE method compared to existing approaches that rely on gene-level aggregate data. An important additional feature of the proposed approach is that BM-DE can be used to analyze RNA-Seq data from experiments without biological replicates. This becomes possible since the approach works with multiple position-level read counts for each gene. We demonstrate the importance of modeling for position-level read counts with a yeast data set and a simulation study. **Availability:** A public domain R package is available from http://odin.mdacc.tmc.edu/~vlji/BMDE/.

Keywords: clustering, false discovery rate, mixture models, next-generation sequencing

Cancer Informatics 2011:10 205-215

doi: 10.4137/CIN.S7473

This article is available from http://www.la-press.com.

© the author(s), publisher and licensee Libertas Academica Ltd.

This is an open access article. Unrestricted non-commercial use is permitted provided the original work is properly cited.

1 Introduction

1.1 RNA-Seq experiments

RNA-Seq is a high-throughput sequencing technology that has recently emerged as a popular methodology to measure gene expression with high accuracy. It generates millions of short reads of mRNA or cDNA. The short reads are mapped to the genome, resulting in a sequence of read counts at millions of genomic positions.^{1,2} RNA-Seq exhibits a high level of reproducibility,¹ and mitigates many limitations of microarrays.³ Consequently, RNA-Seq enables researchers to investigate more complex aspects of the trancriptome, such as allele-specific expression and the discovery of novel promoters and isoforms,⁴ and to develop new approaches to old but fundamental biological questions. An example of the latter is the identification of differentially expressed genes between two conditions.

RNA-Seq experiments produce data on millions of short reads. The data report the base sequence of the reads and the positions on the genome to which the reads are mapped. Most current methods collapse the position-level read counts into a single gene-level summary, such as the number of reads that map per kilobase of exon model per million mapped reads (RPKM), or simply the sum of all the read counts across positions within each gene. Simple hypothesis testing based on the gene-level summaries implements inference on differential gene expression under two biological conditions. We take a different approach. We start by modeling the position-level read counts within a gene. This enables us to account for position outliers among position-level counts. We demonstrate that failure to identify and downweight outliers can bias gene-level summaries. Our hierarchical modeling approach then proceeds with borrowing information across genes for the inference of differential expression. Furthermore, we specifically model systemic biases such as total RNA amount in each experiment to achieve better accuracy in calling differentially expressed genes.

Ji and Liu⁵ illustrated how inference with a Bayesian hierarchical model can improve statistical inference for high-throughput experiments. They also highlighted that borrowing information across loci through a hierarchical model can improve statistical inference even in the case without biological replicates. We follow this advise and propose a Bayesian hierarchical model



to effectively utilizes position-level information and accounts for the variabilities in all the position-level read counts mapped to each gene. We demonstrate the superior performance of the BM-DE method, even when the RNA-seq data are generated from experiments without biological replicates. Due to the still elevated cost of RNA-Seq many studies are carried out without replicates. In such experiments, only one biological sample is prepared per condition for a single run of RNA sequencing. We show that the BM-DE method reduces the false positive findings. Note that this does not imply that the BM-DE method can account for the biological variation in such experiments. This is impossible without replicates. We recommend to use sound and efficient experimental designs⁶ with biological replicates for RNA-Seq experiments. For existing data, some without replicates, the proposed BM-DE approach can be used to increase the precision of calling differentially expressed genes.

1.2 Inference for RNA-Seq data

RNA-Seq data are usually normalized across libraries to adjust for different total read counts by lanes or by samples. In early work, researchers simply used cumulative counts, summing up read counts across positions, followed by minor normalization to account for gene length and the total number of reads.⁷ Recently, more sophisticated normalization methods were proposed. For example, see Robinson and Oshlack⁸ and Balwierz et al.⁹

With the single expression summary per gene per condition, most statistical modeling and inference for differential expression has been based on classical hypothesis testing, such as Fisher's exact test, likelihood ratio tests, or t-tests. For example, Marioni et al¹⁰ modeled read counts with a Poisson distribution, and used a likelihood ratio test to identify differentially expressed genes. Similar to Marioni et al,¹⁰ Wang et al¹ used a Poisson distribution to test differential expression for experiments without biological replicates. Robinson and Smyth¹¹ developed a negative binomial model to account for the variation across replicate samples. They estimated a common dispersion using all tags, and shrinks dispersions of tags toward the estimated common dispersion similar to empirical Bayes approach. edgeR¹² implemented the model for application for RNA-Seq data. Bullard et al¹³ compared the performance of various hypothesis



tests, and found poor performance of the *t*-test, in particular for genes with low counts. They also studied biases introduced by gene-length and the normalization procedure. They observed that the *t*-test tends to yield significant test statistics more frequently for longer genes. This is due to the dependence of the estimated standard error on the mean read counts.

Oshlack and Wakefield¹⁴ further investigated the transcript length bias in RNA-Seq data for differential expression. They illustrated that the standard approaches that use aggregate read counts for each gene in differential expression are subject to significant bias, and that a simple adjustment, dividing by the transcript length, does not entirely remove this bias. Young et al¹⁵ accounted for the transcript length bias in RAN-Seq data, and developed a statistical model for gene ontology analysis.

Bayesian approaches for differential expression in RNA-Seq data have been developed by many researchers, such as Anders and Huber,¹⁶ Hoen et al,³ Taub¹⁷ and Wu et al.¹⁸ Wu et al¹⁸ took an empirical Bayes approach to detect differential expression for RNA-Seq data when biological replicates are not available. They developed a hierarchical model with aggregate counts at gene level to estimate log fold change in gene expression, and mitigated the limitation of experiments without replicate by borrowing strength across all genes.

Differently from the previous approaches, the methods proposed in Jian and Wong,¹⁹ Salzman et al,²⁰ and Li et al²¹ used models to estimate gene expression at the isoform level. Oshlack et al⁴ provided a broad review on current research in preprocessing RNA-Seq data and identifying differentially expressed genes.

In this paper, we propose a novel method for the inference on differential gene expression with three distinct features:

- We explicitly model the read count at each genomic position within a gene. The proposed model can reduce the false positive rate by accounting for the dispersion in the position-specific counts. As another desirable consequence of position-level modeling the length bias disappears. We show significant improvements over existing models that only use gene-level summaries.
- The proposed method does not require prior normalization of the mapped read counts. Instead we

simultaneously carry out the normalization and the inference on differential expression.

• We borrow strength across genes in a hierarchical model. Thus, the detection of differentially expressed genes is informed by the expression measurements in the entire data set.

A related important feature is that borrowing strength across genes in the hierarchical model allows meaningful model-based inference without replicates, if desired.

Section 2 describes the proposed Bayesian model. Section 3 reports the data analysis for the yeast data. Section 4 describes a small simulation study. The last section concludes with a final discussion. The manuscript and R programs with a simple example are available at http://odin.mdacc.tmc.edu/~ylji.

2 Probability Model

RNA-Seq data contains millions of read counts, with each read mapped to a genomic position within a gene. Such count data can be easily assembled from the standard output of upstream read alignment, eg, using SOAP or BOWTIE.²² We consider counts, n_{ij} and m_{ij} , of mapped reads starting at position *j* of gene *i* under two different experimental conditions, 0 and 1, respectively. Here i = 1, ..., I and $j = 1, ..., J_i$. Let $N_{ij} = n_{ij} + m_{ij}$ denote the total count over the two conditions at position *j* of gene *i*. For ad-hoc inference about differential expression we may consider the empirical fraction, $r_{ij} = n_{ij}/N_{ij}$ as the position-level ratio or $r_i = \sum_j n_{ij}/\sum_j N_{ij}$ as the gene-level ratio. The proposed model-based inference improves on these empirical estimates by modeling the position-level read counts.

To start, we characterize sampling variation as binomial sampling. Conditional on the total count N_{ij} , we assume $n_{ij} \sim Bin(N_{ij}, p_{ij})$, independently across positions *j*. Therefore, p_{ij} represents the true proportion of the read count under condition 0 relative to the total read count under both conditions at location *j* of gene *i*. One could use r_{ij} as an empirical estimate of p_{ij} . For example, a value of $r_{ij} = 0.5$ implies that the observed numbers of reads mapped into position *j* of gene *i* are the same across the two conditions. Typically, most r_{ij} 's cluster around a particular value representing a relative expression level of gene *i*. Often the data includes some outliers closer to 0 or 1,



due to random noise. One of our modeling aims is to downweigh these outliers in quantifying the gene expression.

To this end, we introduce a mechanism to downweigh outlying p_{ij} in the inference for differential expression. We achieve this by introducing a latent indicator w_{ij} for each position, with $w_{ij} = 0$ representing an outlier at position *j*. We assume that p_{ij} follows a mixture of beta distributions Ji et al.²³

$$p_{ij} \mid w_{ij}, \alpha_i, \beta_i \stackrel{indep.}{\sim} \begin{cases} \operatorname{Be}(\alpha_i, \beta_i) & \text{if } w_{ij} = 1, \\ \operatorname{Be}(1/2, 1/2) & \text{if } w_{ij} = 0, \end{cases}$$

where Be(*a*, *b*) represents a beta distribution with mean a = (a + b): When $w_{ij} = 0$ the *j*-th position is an outlier, and the expected ratio is given a Be(1/2, 1/2) prior which assigns most probability mass close to 0 or 1. We assume w_{ij} follows a Bernoulli distribution with with probability π_i^w ie, $w_{ij} \sim \text{Ber}(\pi_i^w)$, in which π_i^w represents a gene-specific proportion of outliers. The parameters (α_i , β_i) characterize the expression of gene *i*, excluding the outliers. This formal accounting for outliers in the mixture robustifies inference in critical ways. Later, in the application to a yeast RNA-Seq data set, we will show that failure to downweigh such outliers could even flip the reported inference on differential expression for some genes (Fig. 6).

We reparameterize α_i and β_i for easier interpretation and computation. We follow Robert and Rousseau,²⁴ and let $\eta_i = \log(\alpha_i + \beta_i)$ and $\xi_i = \log(\alpha_i / \beta_i)$. Note that ξ_i is the logit of the mean $\alpha_i/(\alpha_i + \beta_i)$ of the beta distribution. In the (ξ_i, η_i) parametrization an unusually large or small value of ξ_i indicates differential expression, whereas η_i allows for varying levels of heterogeneity across genes. This interpretation leaves ξ_i as the main parameter of interest. Figure 3(b) shows the posterior means of all ξ_i for a yeast RNA-Seq data set (see Section 3). While the cloud in the middle represents the majority of nondifferentially expressed genes, the genes with values ξ_i outside the cloud are those with differential expression. We use a mixture of normal distributions for ξ_i to formalize the notion of differential expression. That is,

$$\begin{aligned} \xi_{i} &| \overline{\xi}, s_{\xi}^{2} \stackrel{iid}{\sim} \pi_{0}^{\lambda} \cdot \mathrm{N}\left(\overline{\xi}, s_{\xi}^{2}\right) + \pi_{-1}^{\lambda} \cdot \mathrm{N}\left(\overline{\xi} - \delta_{-1}, s_{\xi}^{2}\right) \\ &+ \pi_{1}^{\lambda} \cdot \mathrm{N}\left(\overline{\xi}, + \delta_{1}, s_{\xi}^{2}\right). \end{aligned}$$
(1)

We introduce a latent trinary indicator $\lambda_i \in \{0, -1, 1\}$ to represent normal, under-, and over-expression, and rewrite the mixture model (1) as a hierarchical model

$$\begin{split} \boldsymbol{\xi}_{i} | \boldsymbol{\lambda}_{i}, \ \boldsymbol{\overline{\xi}}^{ijd} N \Big(\boldsymbol{\overline{\xi}} + \boldsymbol{\lambda}_{i} \boldsymbol{\delta}_{\boldsymbol{\lambda}_{i}}, s_{\boldsymbol{\xi}}^{2} \Big), \\ \Pr(\boldsymbol{\lambda}_{i} = l) = \boldsymbol{\pi}_{l}^{\boldsymbol{\lambda}}, \ l = -1, \ 0, \ 1 \end{split}$$

We complete the model with priors for $\pi^{w} = (\pi_{1}^{w}, ..., \pi_{l}^{w}), \pi^{\lambda} = (\pi_{-1}^{\lambda}, \pi_{0}^{\lambda}, \pi_{1}^{\lambda}), \delta_{-1}, \delta_{1} \text{ and } s_{\xi}^{2}$. We use a beta distribution $\pi_{i}^{w} \sim \text{Be}(a_{w}, b_{w})$, independently across *i*, a Dirichlet prior $\pi^{\lambda} \sim \text{Dir}(a_{-1}, a_{0}, a_{1})$, and a gamma prior $s_{\xi}^{-2} \sim (a_{s}, b_{s})$. Finally, we use independent gamma priors $\delta_{l} \sim Ga(a_{l}^{\delta}, b_{l}^{\delta}), l = -1, 1$, and $\pi(\overline{\xi}) \propto 1$.

The hyperprior distribution on $\overline{\xi}$, allows for imbalance between the overall counts under the two conditions.

In contrast to fixing $\overline{\xi}$, for example, at $\overline{\xi} = 0.5$, the hierarchical extension with the hyperprior allows for a systematic bias (such as different sequencing depth) across the two conditions. Using possibly different δ_1 and δ_1 allows for varying deviation from the mean $\overline{\xi}^-$ for of over- versus under-expressed genes. For simplicity, we fix η_i in the analysis for the yeast data. If a prior on η_i were desired, one could easily extend the model accordingly, using, for example the prior model from Robert and Rousseau.²⁴ The model is summarized in Figure 1.



Figure 1. Hierarchical model for RNA-Seq data.



3 Yeast Data Analysis

3.1 Data

We illustrate the proposed approach with an RNA-Seq data set from Ingolia et al.²⁵ Specifically, mRNA were extracted from yeast, Saccharomyces cerevisiae strain BY4741, in rich growth medium (YEPD medium) and poor growth medium (amino acid starvation). The goal of the experiment was to identify genes that are differentially expressed between these two biologic conditions. The sequences of short reads were produced using an Illumina Genome Analyzer II. The short reads were mapped using the SOAP method Li et al.²⁶ The data set consists of counts under two different conditions for 1,285 genes.

We considered I=1,089 genes having $J_i \ge 5$ positions for analysis and discarded the remaining 196 for lack of information. The read counts of those 1,089 genes, under the two growth conditions, $\sum_{j=1}^{J_i} .n_{ij}$ and $\sum_{j=1}^{J_i} .m_{ij}$ range from 1 to 9,334 and from 0 to 14,150, respectively. Figure 2 shows histograms of J_i (panel a) and $\sum_{j=1}^{J_i} .N_{ij}$ (panel b) on a logarithm scale (with base 10). Overall, genes have many positions with non-zero counts, and reads per position are small.

3.2 Markov chain Monte Carlo simulations

We estimated and fixed η_i as follows. First, we find $\hat{\alpha}_i$ and $\hat{\beta}_i$ such that and $\hat{\alpha}_i/(\hat{\alpha}_i + \hat{\beta}_i) = r_i$ and

 $\hat{\alpha}_i \hat{\beta}_i / (\hat{\alpha}_i + \hat{\beta}_i)^2 / (\hat{\alpha}_i + \hat{\beta}_i + 1) = \operatorname{var}(r_{ij})$, the sample variance of the r_{ij} . We fix $\eta_i = \log(\hat{\alpha}_i + \hat{\beta}_i)$. We expect that about 5% of all genes are differentially expressed and that about 5% of all positions are outliers. We therefore set $(a_w, b_w) = (19, 1), (a_{-1}, a_0, a_1) = (1, 38, 1),$ $a_{-1}^{\delta}, b_{-1}^{\delta} = (5, 0.11), a_{1}^{\delta}, b_{1}^{\delta} = (5, 0.12), \text{and}(a_{s}, b_{s}) = (3, 0.09).$ We implemented posterior inference using Markov chain Monte Carlo (MCMC) posterior simulations for the proposed model. The implementation is a standard Gibbs sampling algorithm using Metropolis-Hastings transition probabilities with random walk proposals when the complete conditional posterior distribution is not available for efficient random variate generation. We ran the MCMC simulation by iterating over all complete conditionals for 4,500 iterations, discarding the first 500 iterations as burn-in.

3.3 Results

Figure 3(a) plots the posterior probabilities of differential expression, $\hat{p}_i = \Pr(\lambda_1 \neq 0 \mid \text{data})$. Some genes report very large posterior probabilities \hat{p}_i . Figure 3(b) plots the posterior means $\hat{\xi}_i = E(\xi_i \mid \text{data})$. The three dashed horizontal lines mark the posterior means of $(\overline{\xi} + \delta_1), \overline{\xi}$, and $(\overline{\xi} - \delta_{-1})$, respectively. The genes close to or outside the boundary of the lower and upper dashed lines are reported as differentially expressed.









Figure 3. Posterior probability of differential expression, $\hat{p}_i = \Pr(\lambda_i \neq 0 \mid \text{data})$ (panel a) and the posterior mean of relative gene expression over the two conditions, $\hat{\xi}_i = E(\xi_i \mid \text{data})$ (panel b).

Figure 4a plots the marginal posterior probabilities \hat{p}_i against the empirical estimate r_i of relative expression. The plot illustrates that \hat{p}_i agrees with the ad-hoc estimates r_i for most genes. But there are some genes where \hat{p}_i disagrees with (we would argue, improves upon) ad-hoc inference with r_i . In the next two figures

we explore possible reasons for this. Figures 5 and 6 present summaries for some selected genes to illustrate agreement and disagreement of r_i and \hat{p}_i . In both figures, the plots in the first column show N_{ij} (circle) and n_{ij} (cross) along positions. The second column plots r_{ij} along positions. The dashed line indicates



Figure 4. Posterior probabilities $\hat{p}_i = \Pr(\lambda_i \neq 0 \mid \text{data})$ plotted against r_i (panel a). The triangles and squares indicate genes for which posterior inference agrees (triangles) and disagrees (squares) with the inference based on r_i respectively. They will be discussed in Figures 5 and 6. Panel (b) plots posterior expected FDR against the number *D* of genes reported as differentially expressed.



Figure 5. Inference summaries for three genes for which inferences based on r_i and \hat{p}_j agree. The three genes are marked as a triangle in Figure 4a. The first column shows n_{ij} (crosses) and N_{ij} (circles). The second column plots r_{ij} . The dotted line indicates r_i . The dashed line shows the posterior mean $\hat{\xi}_i$ (plotted at logit $-1 \hat{\xi}_i$ to map to the unit scale). The third column plots \hat{w}_{ij} .

the posterior mean $\hat{\xi}_i$, and the dotted line shows the empirical estimate r_i . The line for $\hat{\xi}_i$ is plotted at logit $^{-1}\hat{\xi}_i$ to map to the unit scale. The third column plots the posterior probability $\hat{w}_{ij} = \Pr(w_{ij} = 1 \mid \text{data})$ along positions.

Comparison of the two figures explains the observed discrepancies in r_i and \hat{p}_i . The large r_i in Figure 6 are due to outliers in r_{ij} , including some positions with small total read counts N_{ij} . In contrast, under the posterior inference, many of the \hat{w}_{ij} are imputed with relatively smaller values, leading to a downweighting of the corresponding r_{ij} in the inference for the gene-specific indicators λ_i for differential expression, and thus for \hat{p}_i . Except for these few outliers, most r_{ij} 's are aligned around a value close to 0.5, indicating nondifferential expression. In other words, while r_i is very sensitive to outliers, as desired.

The computation of posterior probabilities $\hat{p}_i = \Pr(\lambda_i \neq 0 \mid \text{data})$ is only half the desired inference. We still need to decide which genes should be reported as differentially expressed. We use a decision rule based on flagging genes with $\hat{p}_i > \kappa$ for some threshold κ . We fix the threshold κ by setting a bound on the false discovery rate (FDR).²⁷ Figure 4b summarizes the FDR implied by decision rules of reporting the genes with highest probability of differential expressed genes. The rule corresponds to a threshold $\kappa = 0.618$.

4 Simulation

We carry out a simulation study to further examine the proposed model. The study investigates the performance of our method in the case where genes have many positions with nonzero counts. In the study, we assume small within-gene variabilities in





Figure 6. Same as Figure 5 for three genes for which inference based on r_i and \hat{p}_i disagree. The genes are marked as rectangles in Figure 4a. Many of the positions are imputed to be possible outliers, and thus downweighted in the inference.

the read counts and large across-gene variabilities. We achieve this by centering η_i around a small value and allowing a relative large variance for ξ_i in our model.

Since the primary goal is inference on ξ_i , we fix η_i at their simulation truth. We place priors on the remaining parameters, $(\overline{\xi}, s_{\xi}^2, \pi^w, \pi^{\lambda}, \delta_{-1}, \delta_1)$ as described in Section 2.

We compare model-based estimates with the simulation truth, and compare the inference under the proposed model to that under two methods: (1) the Analysis of Sequence Counts (ASC) proposed by Wu et al¹⁸ and (2) the MA-plot-based method with random sampling model (DEGseq) proposed in Wang et al.¹

In the ASC, Wu et al model the aggregate read count for each gene under each condition as a binomial random variate, given the total read count summing over all the genes at each condition. The expected proportions in the binomial are compared between the two conditions for each gene. They use δ to denote the difference between the logarithms of the proportions and λ as the sum of the two log proportions. They propose unimodal prior distributions for δ and λ and compute the posterior probability $P(|\delta_i| > \Delta_0 | \text{data})$, where δ_i is log fold change in gene expression of gene *i*, and Δ_0 is a pre-defined threshold for biological significance. In DEGseq, Wang et al. define $M_i = \log_2(C_{0i}) - \log_2(C_{1i})$ and $A_i = (\log_2(C_{0i}) + \log_2(C_{1i}))/2$ where $C_{0i} = \sum_{j=1}^{J_i} n_{ij}$ and $C_{1i} = \sum_{j=1}^{J_i} m_{ij}$. They assume that given $A_i = a, M_i$ approximately follows a normal distribution with mean and variance,

$$E(M_i \mid A_i = a) = \log_2(C_0.) - \log_2(C_1.),$$

$$Var(M_i \mid A_i = a) = 4(1 - \sqrt{2^{2a}/(C_0.C_1.)}(\log_2 e)^2 / \{(C_0.+C_1.)\sqrt{2^{2a}/(C_0.C_1.)}\},$$



where $C_{\kappa} = \sum_{i=1}^{I} C_{\kappa i}$ for $\kappa = 0$, 1. Inference on differential gene expression is then formalized with a z-test. For this simulation study, a normalization for DEGseq and ASC is not necessary for this study since $\overline{\xi}$ is set at 0.

We simulate a sample of I = 1,200 genes. For half of the genes we assumed $J_i = 300$ recorded positions per gene, and for the other half we use $J_i = 100$. We let $\lambda_i = -1$ or 1 for 150 genes and $\lambda_i = 0$ for the remaining 450 genes. Given λ_i , we generate $\eta_i \sim N(\overline{\eta}, s_{\eta}^2)$ and $\xi_i \sim N(\overline{\xi} + \lambda_i \delta_{\lambda_i}, s_{\xi}^2)$, with $\overline{\eta} = 2, s_{\eta}^2 = 0.25^2, \xi = 0,$ $s_{\xi}^2 = 0.1$, and $\delta_{-1} = \delta_1 = 1$. We let $w_{ij} = 0$ or 1 independently with probabilities 0.05 and 0.95, respectively. Conditional on $w_{ij} = 0$ or 1, we respectively generate p_{ij} from either a Be (α_{ij}, β_i) or Be(1/2, 1/2)prior, where $\alpha_i = \exp(\eta_i) \exp(\xi_i)/(1 + \exp(\xi_i))$ and $\beta_i = \exp(\eta_i) = (1 + \exp(\xi_i))$. Finally, we generate $N_{ij} \sim \text{Ga}(1.5, 1/1.5)$ (rounded up to the nearest integer), and $n_{ij} \sim \text{Bin}(N_{ij}, p_{ij})$, independently. We then proceed to estimate ξ_i and P $(\lambda_i \neq 0 \mid \text{data})$ conditional on N_{ij} and n_{ij} under the proposed model.

The receiver operating characteristic (ROC) curve is commonly used to select an optimal method for classification problems. We assume a decision rule that reports genes with posterior probabilities, $p(\lambda \neq 0)$ | data) and P($|\delta| > \delta_0$ data) (in the cases of the proposed approach and ASC) or P-value (in the case of DEGseq) beyond a threshold where we set $\delta_0 = 1.8$. The ROC curve plots true positive rate against the false positive rate as a parametric curve indexed by the threshold. Figure 7 shows the three ROC curves. The ROC curve for the proposed method compares favorably against the alternatives. It demonstrates the limitations of ASC. We believe that this is due to the strong assumptions on the shape of the priors of δ and λ . The simulation truth is that the mean expression of the genes is generated from a mixture of three distributions, which does not agree with the unimodal assumptions of the ASC model.

Regarding the performance of DEGseq, we note that longer genes tend to have larger aggregate counts across positions. Therefore, DEGseq is more likely to declare long genes with small effects as differentially expressed genes since its estimated standard deviation inherently depends on the mean counts. Specifically, we observe that DEGseq tends to produce smaller *p*-values for nondifferentially expressed genes with $J_i = 300$ than those with $J_i = 100$ due to



Figure 7. ROC curves for identification of differential gene expression under the proposed method (black solid line), the DEGseq (red dotted line) proposed by Wang et al²⁸ and the ASC (blue dashed line) proposed by Wu et al¹⁸ in the simulation study.

the gene-length bias (see Fig. 8a). On the other hand, the proposed model accounts for position-specific variability while more information on relative gene expression gets accumulated as the number of positions within a gene increases (see Fig. 8b). Therefore, the proposed method tends to produce smaller posterior probability of differential expression for non-differentially expressed genes with $J_i = 300$ than those with $J_i = 100$. This, coupled with vague position-specific information leads to superior performance of the proposed method for longer genes. This conveys significant implication on statistical inference of differential expression using RNA-Seq data. Since RNA-Seq experiments produce many non-zero count positions within a gene, and many reads per position, the RNA-Seq data enables us to model variability among expression levels on positions within the same gene, and the incorporation of it into a model improves the resulting inference.

We note that if both N_{ij} and J_i are small, modeling the position-level read counts does not significantly improve inference. Also, if there is little variation across position-level counts, then the loss of information under aggregation remains negligible. We found that for cases where short reads are mapped to small number of positions, DEGseq performs well (results not shown). However, such situations are untypical





Figure 8. Boxplots of *P* value under DEGseq (panel a) and \hat{p}_i under the proposed method (panel b) by the number of positions within a gene, $J_i = 100$ or 300.

for large-scale RNA-Seq experiments with usually very noisy data.

5 Discussion

We proposed a Bayesian model-based approach for inference with RNA-Seq data. We introduced a hierarchical structure to model the position-level count data. We demonstrate through a simulation study and the analysis of a yeast experiment that the model effectively downweights outlying observations at the position level and obtains more robust estimates of gene expression.

The model provides a promising framework for further development of statistical models for RNA-Seq data. One possible extension is to relax the parametric assumption for ξ_i . By removing the restriction to a specific parametric family of distributions, one could further robustify inference about gene expression levels. Another important extension is to incorporate dependence across genes. In the current model we assumed that ξ_i are independently and identically distributed. One may achieve more precise estimates and formal inference about dependence structure by generalizing the model to allow for dependence of ξ_i across genes. One could build on available prior information to construct informative priors for dependence at the level of the indicators λ_i . For model with indicators at the gene level similar to λ_i used in our model this is carried out in.²⁹ The binary nature of λ_i greatly simplifies general modeling of dependence structure. For a recent discussion of models for dependent gene expression see, for example, Stingo et al³⁰ or Jones et al,³¹ and references therein. Both references use a model-based Bayesian approach as in this paper.

Finally, while the model was specifically developed for experiments comparing two conditions without biologic replicates, simple modification would allow the use for experiments with replicates or experiments with multiple conditions. The proposed model can be extended for experiments with replicates by replacing the binomial sampling model for n_{ij} by a model for counts across replicates. For experiments with multiple conditions, one may consider a multinomial likelihood with a Dirichlet prior.

Acknowledgements

Yuan Ji and Peter Müller's research is partially supported by NIH R01 CA132897. Shoudan Liang's research is supported by NIH K25 CA123344.

Disclosure

This manuscript has been read and approved by all authors. This paper is unique and is not under consideration by any other publication and has not been published elsewhere. The authors and peer reviewers of



this paper report no conflicts of interest. The authors confirm that they have permission to reproduce any copyrighted material.

References

- Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*. 2009b;10:57–63.
- Li J, Jiang H, Wong WH. Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biology*. 2010b:11.
- Hoen PAC, Ariyurek Y, Thygesen HH, et al. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Research*. 2008; 36:e141.
- 4. Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results. *Genome Biology*. 2010;11:12.
- 5. Ji H, Liu XS. Analyzing' omics data using hierarchical models. *Nature Biotechnology*. 2010.
- Auer PL, Doerge RW. Statistical design and analysis of RNA sequencing data. *The Genetics Society of America*. 2010;185:405–16.
- Mortazavi A, Williams BA, McCue K, Schaefier L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*. 2008;5:621–8.
- 8. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*. 2010;11:3.
- Balwierz PJ, Carninci P, Daub CO, et al. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome Biology*. 2009;10:7.
- Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*. 2008;18:1509–17.
- Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*. 2007;23(21):2881–7.
- Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139–40.
- Bullard JH, Purdom EHKD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*. 2010:11.
- Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*. 2010:4.
- Young MD, Wakefieldand MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*. 2010; 11:2.
- Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biology*. 2010;11:10.
- Taub MA. Analysis of high-throughput biological data: some statistical problems in RNA-seq and mouse genotyping. Ph.D. thesis. *Department of Statistics*, UC Berkeley. 2009.
- Wu Z, Jenkins BD, Rynearson TA, et al. Empirical bayes analysis of sequencing-based transcriptional profiling without replicates. *BMC Bioinformatics*. 2010:11.
- Jian H, Wong WH. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*. 2009;25:1026–32.
- Salzman J, Jiang H, Wong WH. Statistical modeling of RNA-Seq data. Tech. rep., *Division of Statistics*, Stanford University; 2010.
- Li B, Ruotti V, Stewart R, Thomson J, Dewey C. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*. 2010a;26: 493–500.
- Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memoryeficient alignment of short DNA sequences to the human genome. *Genome Biology*. 2009;10:R25.
- Ji Y, Wu C, Liu P, Coombes K. Applications of beta-mixture models in bioinformatics. *Bioinformatics*. 2005;21(9):2118–22.
- Robert CP, Rousseau J. A Mixture Approach to Bayesian Goodness of Fit. Les cahiers du CEREMADE (2002–9). 2004.

- Ingolia N, Ghaemmaghami S, Newman J, Weissman J. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*. 2009;324(5924):218–23.
- Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. *Bioinformatics*. 2008.
- Newton MA, Noueiry A, Sarkar D, Ahlquist P. detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*. 2004;5:155–76.
- Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*. 2009a;26(1):136–8.
- 29. Telesca D, Müller P, Parmigiani G, Freedman RS. Modeling dependent gene expression. Tech. rep., Harvard University; 2010.
- Stingo F, Chen Y, Tadesse M, Vannucci M. Incorporating biological information into linear models: a bayesian approach to the selection of pathways and genes. *Annals of Applied Statistics*. 2011.
- Jones B, Carvalho C, Dobra A, Hans C, Carter C, West M. Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*. 2004;20:388–400.

Publish with Libertas Academica and every scientist working in your field can read your article

"I would like to say that this is the most author-friendly editing process I have experienced in over 150 publications. Thank you most sincerely."

"The communication between your staff and me has been terrific. Whenever progress is made with the manuscript, I receive notice. Quite honestly, I've never had such complete communication with a journal."

"LA is different, and hopefully represents a kind of scientific publication machinery that removes the hurdles from free flow of scientific thought."

Your paper will be:

- Available to your entire community free of charge
- Fairly and quickly peer reviewed
- Yours! You retain copyright

http://www.la-press.com