Published in final edited form as:

Stat Biosci. 2015 May 1; 7(1): 48–67. doi:10.1007/s12561-013-9096-7.

Bayesian Hierarchical Model for Differential Gene Expression Using RNA-seq Data

Juhee Lee[*,1], **Yuan Ji**[2], **Shoudan Liang**, **Guoshuai Cai**[3], and **Peter Müller**[4]

[1]Department of Statistics, The Ohio State University, Columbus, OH, U.S.A.

[2]Cancer Research Informatics, Northshore University HealthSystem, Evanston, Illinois, U.S.A.

[3]Dpt. of Bioinformatics and Computational Biology, M.D. Anderson Cancer Ctr., Houston, Texas, U.S.A.

[4]Department of Mathematics, University of Texas Austin, Austin, Texas, U.S.A.

Abstract

We introduce model-based Bayesian inference to screen for differentially expressed genes based on RNA-seq data. RNA-seq is a high-throughput next-generation sequencing application that can be used to measure the expression of messenger RNA. We propose a Bayesian hierarchical model to implement coherent, fast and robust inference, focusing on differential gene expression experiments, i.e., experiments carried out to learn about differences in gene expression under two biologic conditions. The proposed model exploits available position-specific read counts, minimizing required data pre-processing and making maximum use of available information. Moreover, it includes mechanisms to automatically discount outliers at the level of positions within genes. The method combines gene-level information across replicates, and reports coherent posterior probabilities of differential expression at the gene level. An implementation as a public domain R package is available.

Keywords

Bayes; Differential Gene Expression; FDR; Mixture Models; Next-Generation Sequencing

1 Introduction

1.1 RNA-seq Data and Differential Gene Expression

RNA-seq is a high-throughput sequencing technology to efficiently measure transcript abundance that allows researchers to investigate a wide range of biological questions based on high resolution measurements. Here, transcripts refer to segments of RNA. In RNA-seq experiments, isolated RNA samples are fragmented and synthesized to complimentary DNA (cDNA) in a library preparation step. As a result, millions of short reads are produced. Each short read is sequenced and finally, the reads are mapped into a reference genome based on sequence similarities. Widely used mapping methods include MAQ (Li *et al.*, 2008a), SOAP

[*]Address for Correspondence: 1958 Neil Avenue Cockins Hall, Room 404 Columbus, OH 43210-1247 lee.2597@osu.edu.

(Li *et al.*, 2008b), Bowtie (Langmead *et al.*, 2009) and SHRiMP (Rumble *et al.*, 2009). Counts of mapped reads along genomic positions inform us about transcript abundance (Marioni *et al.*, 2008). Figure 1 shows a stylized graphical representation of RNA-seq data after alignment of the reads.

An important application of RNA-seq is the identification of differentially expressed genes across different biological conditions. We propose hierarchical models to screen genes for differential expression. The model is robust against intragenic variability by explicitly making use of position level data. The proposed inference provides calls for differential expression at the intragenic positions level as well as at the gene level. In other words, our method reveals positions within a gene that have different expression between two biological conditions, and at the same time produces inference on differential expression at the gene level. The position-level calls as well as the gene-level calls borrow information from read counts across different positions within a gene and across different genes. The entire inference is based on a coherent and model-based statistical framework. We do not model expression of isoforms. Isoform calling is a separate and challenging research topic in and of itself. Our goal is to generate a list of genes with strong evidence of differential expression supported by most intragenic read counts at various positions. In some other cases when alternative splicing is irrelevant (such as sequencing experiments for yeast or some plants), our method can be directly applied to obtain differentially expressed genes.

Compared to traditional array-based platforms, the analysis of RNA-seq data introduces the new challenge of modeling the read counts at multiple positions within each gene. To make appropriate inference at the gene level, one must first summarize the position-level counts and form a meaningful gene-level quantification. Another challenge is the typically small number of replicate experiments. The cost of sequencing experiments is still high, and the number of replicate samples remains small for most existing data. Thus, it is important that an inference approach provides stable estimates even with small sample sizes.

Several statistical methods have been proposed for the detection of differential expression using RNA-seq data. Most require pre-processing and normalization of the read count data to make expression measurements comparable across samples. Specifically, the read counts across positions within a gene are usually aggregated to form gene-level summaries. Then gene-level summaries are normalized to account for different library sizes across samples. The normalized gene-level summaries are used as gene expression quantifications for differential analysis. For example, Mortazavi *et al.* (2008) introduced reads per kilobase of exon model per million mapped reads (RPKM) as a normalized gene-level measurement. RPKM aggregates read counts across positions within a gene by simple summation and normalizes the aggregated counts for gene length and for the total read number in the measurement.

Read counts at positions depend on other factors as well as the gene expression levels. For a gene without isoforms such as a gene in yeast, approximately constant read counts along positions are expected to be observed. However, read counts at some positions can be different from those at other positions due to systematic biases. Dohm *et al.* (2008) discussed that GC-riched regions tend to generate larger counts than AT-riched regions.

Similarly, Hillier *et al.* (2008), Li *et al.* (2010), Hansen *et al.* (2010) and Schwartz *et al.* (2011) stated that biases associated with genomic positions are present in RNA-seq data and they have possible effects in downstream analyses. Protocols in experiments can also affect read counts at positions, as pointed in Li *et al.* (2010). Thus, ignoring these biases and using aggregated gene-level read counts can result in inefficient inferences on differential gene expression.

In addition, Bullard *et al.* (2010) and Oshlack and Wakefield (2009) pointed out additional biases that are introduced in these pre-processing steps. In particular, the analysis may suffer from gene-length bias due to the dependence of variance estimates on read counts. That is, long genes with small effects are more likely to be detected than short genes with large effects. Also, due to the domination of read counts by a few, but very abundant genes, normalization with the total counts leaves the detection of differential expression less sensitive. Recently more sophisticated normalization methods to mitigate these problems were proposed by Balwierz *et al.* (2009), Tang *et al.* (2009) and Robinson and Oshlack (2010). We will demonstrate that our proposed model does not require pre-processing or normalization of read counts, and is thus not susceptible to the above biases. In particular, pre-processing and normalization in the proposed model are replaced by model-based adjustments.

Most existing approaches model the read counts with a Poisson or a negative binomial distribution, and conduct statistical hypothesis tests for each gene, using Fisher's exact test, a likelihood ratio test or a t test. Marioni *et al.* (2008) used a Poisson distribution to model gene-level counts in the presence of technical replicates, and carried out a likelihood ratio test to identify differentially expressed genes. Similarly, Wang *et al.* (2009) used a normal approximation to the binomial distribution and developed a package called DEGseq. Robinson and Smyth (2007) developed a negative binomial model to account for overdispersion in biological replicates. They followed an empirical Bayes-like approach to achieve shrinkage in the estimation of counts dispersion. The *edgeR* package (Robinson *et al.*, 2010) implements their method. Anders and Huber (2010) extended the negative binomial model of Robinson and Smyth (2007) by using local regression to link mean and variance parameters. There are few existing Bayesian approaches for inference on differential expression based on RNA-seq data. Hardcastle and Kelly (2010) developed an empirical Bayes method named Bayseq for complex experimental designs having more than two conditions. Wu *et al.* (2010) proposed an empirical Bayes method to identify differentially expressed genes for single replicate data. Differently from the previous approaches, Lee *et al.* (2011) directly modeled position-level counts and proposed a hierarchical Bayes model for single replicate data. Oshlack *et al.* (2010) provided an overview on preprocessing of RNA-seq data and critically review commonly used methods for the differential gene expression.

In this paper, we develop a statistical model to infer differential gene expression across two biological conditions exploiting position level data in a hierarchical model across replicates, genes and positions. Ji and Liu (2010) demonstrated that hierarchical models can yield higher efficiency in analyzing data from high-throughput experiments. They also argue that borrowing information across genes or loci through inference in a hierarchical model

improves inference for any particular gene or locus. In the same spirit, we propose a Bayesian hierarchical model which borrows information across positions, genes, and replicates. A distinct feature of our approach is that we model position-level counts to infer differential expression at the gene level.

Modeling position-level counts makes it possible to downweight the outlying count values at certain positions, therefore leading to more accurate summaries of gene expression. In addition, we do not use a predetermined value to represent nondifferential expression. For example, we do not assume that if a gene is not differentially expressed, the ratio of the expression measurement between two conditions will be 1, as it would be if the sequence data under the two conditions were perfectly normalized. Instead we introduce parameters that represent the ratio of expression when genes are not differentially expressed. These parameters are estimated, thus resulting in more robust inference.

1.2 Data

We analyze RNA-seq data from the yeast experiment presented in Ingolia *et al.* (2009). The mRNA were extracted from a yeast, Saccharomyces cerevisiae strain BY4741, in rich growth medium (YEPD medium) and poor growth medium (amino acid starvation). The fragments of mRNA were sequenced with an Illumina Genome Analyzer II and mapped using the SOAP method (Li *et al.*, 2008b). The data includes two replicates ($K = 2$) and reports position-specific read counts for 1,285 genes.

Section 2 describes the probability model. Section 3 describes two simulation studies. Section 4 reports the data analysis for the yeast data. The last section concludes with a final discussion. The manuscript and R programs with an example are available at http://www.northshore.org/research/investigators/Yuan-Ji-PhD/.

2 Methodology

2.1 Sampling Model of Observed Read Counts

RNA-seq data reports read counts arranged by genomic positions within a gene. For each of the genomic positions, the read count refers to the number of mapped reads starting at that position. See Figure 1 for illustration. We assume that data is recorded under two different experimental conditions denoted as 0 (control) and 1 (experimental condition). We denote the read counts under the two conditions by n_{kij} (experimental condition) and m_{kij} (control), representing the numbers of mapped reads starting at position j of gene i for replicate k, where $i = 1, \ldots, I$, $j = 1, \ldots, J_i$, and $k = 1, \ldots, K(\geq 1)$. Let $N_{kij} = n_{kij} + m_{kij}$ denote the total count at position j of gene i for replicate k, summed over conditions. Implicit in the notation is an assumption that the experiment has the same number of replicates under both conditions and the replicates are randomly paired for an analysis. We note that although it is not necessary for the proposed model, if replicates in RNA-seq data are designed to be paired, then other effects possibly confounded with relative expression gene expression level can be minimized. For example, Auer and Doerge (2010) recommended that experiments should be carried out with a balanced block design, exploiting features of the multiplexing technique that is already employed in most RNA-sequencing devices.

For ad-hoc inference about differential expression one may consider the empirical fraction, $r_{kij} = n_{kij}/N_{kij}$ for each position or $r_{ki} = \Sigma_j n_{kij}/\Sigma_j N_{kij}$ for each replicate. In contrast, we propose model-based adjustments of theses empirical fractions to account for sampling variability across positions within a gene, for outliers and for biological variability across replicates. To start, we assume a Poisson distribution for the read counts in each condition, $n_{kij} \overset{ind}{\sim} \mathrm{Poi}(\phi_{0kij})$ and $m_{kij} \overset{ind}{\sim} \mathrm{Poi}(\phi_{1kij})$, where the counts are independently distributed across conditions, replicates (k), genes (i), and positions (j). Reads are counted at the location corresponding to the starting position of the read, avoiding trivial dependence by overlap of reads across multiple loci. The Poisson models for n_{kij} and m_{kij} imply $N_{kij} \overset{ind}{\sim} \mathrm{Poi}(\phi_{0kij}+\phi_{1kij})$, and $n_{kij}|N_{kij} \overset{ind}{\sim} \mathrm{Bin}(N_{kij}, p_{kij})$ where $p_{kij} = \varphi_{0kij}/(\varphi_{0kij} + \varphi_{1kij})$. Assuming that φ_{0kij} and φ_{1kij} independently follow gamma distributions, it follows that p_{kij} is a priori independent of $(\varphi_{0kij} + \varphi_{1kij})$. We can therefore focus on modeling n_{kij} given N_{kij} and p_{kij}. See Appendix A for more details. Here, p_{kij} represents the true proportion of read counts under condition 0 relative to the total read counts under both conditions at location j of gene i and replicate k. Note that the binomial model is assumed conditional on p_{kij} only. Marginalizing with respect to the following prior on p_{kij} will allow for more flexibility.

2.2 Modeling Position-Level Calls

Examining the counts within a gene, we find that among the values of r_{kij} occasionally some r_{kij}'s are very different from the rest. Figure 2 shows a plot of r_{kij} for a selected gene and replicate from the yeast data. The values r_{kij} are large for most positions, but we also observe some aberrant positions with unusually small r_{kij}'s. Those potential outliers marked as crosses in the plot may inappropriately drive the estimate of the mean relative expression level for this gene toward a smaller value, and result in the misidentification of differential expression. Since yeast genes do not splice alternatively, the outlier values can not be explained by isoforms. We believe that they are caused by systemic biases in the experiments and instruments. In cases when isoforms are possible, these positions might indicate splice junctions. For the yeast data in which isoforms are absent, we want to downweight such position-level outliers in the inference for differential expression at the gene level, and we employ a mixture of beta distributions as a prior distribution for modeling p_{kij}. We introduce a latent indicator w_{kij} for each position, with $w_{kij} = 0$ representing an outlier at position j for replicate k. Let

$$p_{kij}|w_{kij}, \lambda_i, \alpha_{ki}, \beta_{ki} \overset{indep.}{\sim} \begin{cases} \mathrm{Be}(\alpha_{ki}, \beta_{ki}) & \text{if } w_{kij}=1, \\ \mathrm{Be}(\alpha^0_{\lambda_i}, \beta^0_{\lambda_i}) & \text{if } w_{kij}=0. \end{cases} \quad (1)$$

We assume $w_{kij} \sim \mathrm{Ber}(\pi^w_{ki})$, where π^w_{ki} represents a gene-specific proportion of outliers for replicate k. Here $\lambda_i \in \{-1, 0, 1\}$ is a latent indicator for under-, normal- and over-expression of gene i.

The beta mixture models (1) is a key component of the model. First, note that the mixture includes a regression on the latent indicator λ_i of differential gene expression. That is, we let the beta distribution for outliers depend on the status of differential expression. Since λ_i takes three values, there are in fact four components in the beta mixture (1), corresponding

to the distinct values of w_{kij} and λ_i. When $w_{kij} = 0$, i.e., when the position is an outlier, the prior of p_{kij} depends on the value of λ_i which tells us if gene i is differentially expressed. Specifically, if $\lambda_i = 0$, meaning gene i is not differentially expressed, then an outlier position should have p_{kij} values away from the "middle". Recall that p_{kij} characterizes the ratio of count in one condition over the sum of two counts across both conditions. So if a gene is not differentially expressed, the ratio should be close to 0.5. Therefore, an outlier here means that p_{kij} is extreme, closer to 0 or 1. Therefore, we let $\alpha_0^0 = \beta_0^0 = 1/2$, resulting in a Be(1/2, 1/2) prior for p_{kij} that puts large mass on values close to 0 and 1. Similarly if $\lambda_i = -1$, a position that is not an outlier should have p_{kij} close to 0 since the gene is under-expressed. Therefore for outliers, we let $\alpha_{-1}^0 = 1$ and $\beta_{-1}^0 = 1/2$ so that the beta prior puts most mass on values close to 1; finally when $\lambda_i = 1$, we let $\alpha_1^0 = 1/2$ and $\beta_1^0 = 1$.

2.3 Modeling Differential Gene Expression

The main parameters of interest in (1) are $(\alpha_{ki}, \beta_{ki})$. They inform us about the expression of gene i in replicate k, excluding the outliers. The formal accounting for outliers in the mixture (1) robustifies inference in critical ways. Later, in the application to yeast data, we will show how failure to downweight outliers could even flip the reported inference on differential expression for some genes.

The use of a single indicator λ_i for differential expression versus a large number of indicators w_{kij} for possible outliers avoids identifiability concerns related to (falsely) imputing $\lambda_i = 0$ and $w_{kij} = 0$ for all k and j, instead of $\lambda_i \neq 0$ for a differentially expressed gene. The prior probability of the earlier set of indicators is far smaller than for the latter. This effect of prior shrinkage towards the more parsimonious model is known as Ockham's razor (Jefferys and Berger, 1992).

For simplicity and following Robert and Rousseau (2004), we reparameterize α_{ki} and β_{ki} as $\gamma_{ki} = \alpha_{ki} + \beta_{ki}$ and $\mu_{ki} = \alpha_{ki}/(\alpha_{ki} + \beta_{ki})$. After reparameterization, the beta distribution Be(α_{ki}, β_{ki}) is indexed as Be($\gamma_{ki}\mu_{ki}$, $\gamma_{ki}(1 - \mu_{ki})$) where $\gamma_{ki} > 0$ is now a scale parameter, and $0 < \mu_{ki} < 1$ is a location parameter. A second reparameterization to $\eta_{ki} = \log(\gamma_{ki})$ and $\xi_{ki} = \log(\mu_{ki}/(1 - \mu_{ki})) = \text{logit}(\mu_{ki})$ further simplifies computation by removing restrictions on the parameter space. In the (ξ_{ki}, η_{ki}) space, an unusually large or small value of ξ_{ki} indicates differential expression, whereas η_{ki} allows for varying levels of heterogeneity across genes. This interpretation leaves ξ_{ki} as the main parameter of interest. Figure 8bc shows the posterior means of all ξ_{ki} from the analysis for the yeast data. While the central cloud represents the majority of nondifferentially expressed genes, the genes having distant values of ξ_{ki} relative to the genes in the cloud (above or below the cloud) are those with differential expression. We use a mixture of normal distribution for ξ_{ki} to formalize the notion of differential expression. Recall that $\lambda_i \in \{-1, 0, 1\}$ is an indicator for under-, normal- and over-expression. We assume

$$\xi_{ki}|\bar{\xi}_k, s^2_{-1k}, s^2_{0k}, s^2_{1k}, \delta_{1k}, \delta_{-1k}, \lambda_i \overset{iid}{\sim} \begin{cases} \mathrm{N}(\bar{\xi}_k - \delta_{-1k}, s^2_{-1k}) & \text{if } \lambda_i = -1, \\ \mathrm{N}(\bar{\xi}_k, s^2_{0k}) & \text{if } \lambda_i = 0, \\ \mathrm{N}(\bar{\xi}_k + \delta_{1k}, s^2_{1k}) & \text{if } \lambda_i = 1, \end{cases} \quad (2)$$

and $\Pr(\lambda_i = \ell) = \pi^{\lambda}_{\ell}$ for $\ell = -1, 0, 1$. The unusual indexing with 0, 1, −1 is used in anticipation of the upcoming discussion.

The location parameter $\bar{\xi}_k$ is the mean of ξ_{ki} when gene i is not differentially expressed. We keep index k in $\bar{\xi}_k$ to account for a potential difference in library size across different replicate experiments. If the overall counts under the two conditions are equal, then $\bar{\xi}_k$ = logit(0.5) represents non-differential expression. Due to various reason, such as lane effects in RNA-Seq experiments, we keep $\bar{\xi}_k$ random, rather than fixing $\bar{\xi}_k \approx$ logit(0.5). The means of r_{ki} for the yeast data are observed at 0.587 and 0.585 for the two replicates.

The three Gaussian components in the mixture model (2) correspond to normal and over- or under-expression. Note that we use two parameters (δ_{-1k}, δ_{1k}) to allow for different deviation from the mean $\bar{\xi}_k$ for over- or under-expressed genes. We introduce a hyperprior $p(\bar{\xi}_k) \propto 1$. For simplicity, we estimate and fix η_{ki} as shown in Appendix B, based on an empirical Bayes approach. If a prior on η_{ki} were desired, one could follow Robert and Rousseau (2004), and use $p(\gamma_{ki}) \propto \{1 - \exp(-\tau_0 \gamma^{c_0}_{ki})\} \exp(-\tau_1 \gamma^{c_1}_{ki} - \tau_2 / \gamma^{c_2}_{ki})$, where $\gamma_{ki} = \exp(\eta_{ki})$, and c_0, c_1, c_2, τ_0, τ_1, τ_2 are additional hyperparameters. However, this prior is quite informative and results seem to vary with different calibration of the prior. We found that the estimation procedure in Appendix B resulted in better performance in our examples.

We complete the model with priors for $\boldsymbol{\pi}^w_k = (\pi^w_{k1}, \ldots, \pi^w_{kI})$, $\boldsymbol{\pi}^{\lambda} = (\pi^{\lambda}_{-1}, \pi^{\lambda}_0, \pi^{\lambda}_1)$, $(\delta_{-1k}, \delta_{1k})$ and $(s^2_{-1k}, s^2_{0k}, s^2_{1k})$. We use a beta distribution $\pi^w_{ki} \sim \mathrm{Be}(a^w_k, b^w_k)$, independently across k and i, a Dirichlet prior $\boldsymbol{\pi}^{\lambda} \sim \mathrm{Dir}(a_{-1}, a_0, a_1)$, and a gamma prior $s^{-2}_{\ell k} \sim \mathrm{Ga}(a^s_{\ell k}, b^s_{\ell k})$, $\ell = -1, 0, 1$ where $\mathrm{Ga}(a, b)$ is a gamma distribution with shape parameter a and rate parameter b. Finally we use independent gamma priors $\delta_{\ell k} \sim \mathrm{Ga}(a^d_{\ell k}, b^d_{\ell k})$, $\ell = -1, 1$. The hierarchical model is summarized in Figure 3.

2.4 Computation

We implement posterior inference using Markov chain Monte Carlo (MCMC) posterior simulations for the proposed model. The complete conditional posterior distributions for all parameters except ξ_{ki} are available for efficient random variate generation. The implementation is a standard Gibbs sampling algorithm with a Metropolis-Hastings transition probability to update ξ_{ki}.

The computation takes 1.08 hours on 2.67 GHz CPU for the RNA-seq yeast data set in Section 1.2 for 5,000 iterations. Note that the yeast data set has 1,016 genes and the average number of positions within a gene is 27.81.

3 Simulation

3.1 Simulation 1

3.1.1 Simulation 1: Validation—We examine the proposed model in a simulation study. We compare model estimates with the simulation truth and with inference by *edgeR* (Robinson *et al.*, 2010). EdgeR aggregates read counts across positions within a gene by simple summation and normalizes the gene-level read counts using the TMM (trimmed mean of M values) method (Robinson and Oshlack, 2010). It requires a separate normalization procedure and uses the normalized counts to compute *p*-values of differential expression based on a negative binomial exact test.

We simulate a set of $I = 1,000$ genes with 250 genes each having $J_i = 50$, 150, 200, or 300 positions, respectively. We let $\lambda_i = -1$ or 1 for 25 genes among each set of 250 genes and $\lambda_i = 0$ for the remaining 225 genes. We assume three replicates ($K = 3$). Given λ_i, we generate η_{ki} and ξ_{ki} from $\mathrm{N}(\bar{\eta}_k, s^2_{\eta k})$ and $\mathrm{N}(\bar{\xi}_k + \lambda_i \delta_{\lambda_i k}, s^2_{\xi k})$ distributions with $\bar{\eta_k} = 2.5$, $s^2_{\eta k} = 0.25^2$, $\bar{\xi_k} = 0$, $s^2_{\xi k} = 0.45^2$, $\delta_{-1k} = \delta_{1k} = 0.6$. We let $w_{kij} = 0$ or 1 independently with probabilities 0.1 and 0.9, respectively. When $w_{kij} = 1$ we generate $p_{kij} \sim \mathrm{Be}(a_{ki}, \beta_{ki})$, where $a_{ki} = \exp(\eta_{ki})\exp(\xi_{ki})/(1 + \exp(\xi_{ki}))$ and $\beta_{ki} = \exp(\eta_{ki})/(1 + \exp(\xi_{ki}))$. Similar to the previous subsection, when $w_{ij} = 0$ we use $p_{kij} \sim \mathrm{Be}(\alpha^0_{\lambda_i}, \beta^0_{\lambda_i})$ with $(\alpha^0_{\lambda_i}, \beta^0_{\lambda_i}) = (1, 1/10)$ and (1/10, 1) for $\lambda_i = -1$ and 1, respectively. For $\lambda_i = 0$, we let $(\alpha^0_{\lambda_i}, \beta^0_{\lambda_i})$ be (1, 1/10) or (1/10, 1) with equal probability. Finally, we generate $N_{kij} \sim \mathrm{Ga}(20, 1/4)$, $\mathrm{Ga}(40, 1/4)$, and $\mathrm{Ga}(10, 1/2)$ for $k = 1$, 2, 3, respectively (rounded up to the nearest integer), and $n_{kij} \sim \mathrm{Bin}(N_{kij}, p_{kij})$, independently. In particular, the simulation includes a strong replicate effect. We then proceed to estimate ξ_{ki} and $P(\lambda_i \neq 0 \mid N, n)$ under the proposed model.

To proceed with inference on λ_i, we specified priors on the parameters, $(\bar{\xi}_k, s^2_{-1k}, s^2_{0k}, s^2_{1k}, \boldsymbol{\pi}^w_k, \boldsymbol{\pi}^\lambda, \delta_{-1k}, \delta_{1k})$ as described in Section 2. We a priori assumed that 10% of genes are differentially expressed and that 10% of positions are outlying. We estimated and fixed η_{ki} as described in Appendix B. We also initialized ξ_{ki} as in Appendix B. We let $s^{-2}_{-\ell k} \sim \mathrm{Ga}(20, 1)$ for $\ell = -1$, 0, 1 and $k = 1$, 2 and $\delta_{-\ell k} \sim \mathrm{Ga}(20, 1/20)$ for $\ell = -1$, 1 and $k = 1$, 2 where $\mathrm{Ga}(a, b)$ is a gamma distribution with mean ab. We ran the MCMC simulation by iterating over all complete conditionals for 5,000 iterations, discarding the first 3,000 iterations as burn-in.

Figure 4 plots summaries for a selected gene and replicate to illustrate how the model discounts position-level outliers by means of w_{kij}. The selected gene is truly under-expressed, $\lambda_i = -1$. The left panel shows the empirical fractions $r_{kij} = n_{kij}/N_{kij}$ along the positions. The crosses represent outlying positions. The dashed line shows the true mean expression level $\mu_{ki} = \exp(\xi_{ki})/(1 + \exp(\xi_{ki}))$ for the gene. The right panel shows the posterior probabilities $\Pr(w_{kij} = 0 \mid \text{data})$ plotted along position j. Most positions marked with crosses report high posterior probabilities, confirming that the model correctly identifies outlying positions.

Figure 5ab plots the posterior estimates $\hat{\xi_{ki}} = E(\xi_{ki} \mid \text{data})$ and the posterior probabilities of differential gene expression, $\hat{p_i} \equiv Pr(\lambda_i \neq 0 \mid \text{data})$ against the simulation truth. The posterior means $\hat{\xi_{ki}}$ are close to the true values of ξ_{ki}. Panel (a) plots $\hat{\xi_{ki}}$ for replicate $k = 1$. Panel (b) summarizes the posterior probabilities $\hat{p_i}$ of differential gene expression. The plot is arranged by the true values of λ_i. Differentially expressed genes report larger posterior probabilities of differential gene expression, as desired.

3.1.2 Simulation 1: Comparison—We compare inference under the proposed model with inference under *edgeR*, *Bayseq* (Hardcastle and Kelly, 2010), *DEGseq* (Wang *et al.*, 2009) and the overdispersed logistic model (Baggerly *et al.*, 2004). Because of the small number of replicates we use a common dispersion parameter in *edgeR*. The receiver operating characteristic (ROC) curve is commonly used to evaluate classification methods. Figure 6a plots the true positive rate and the false positive rate as the cutpoint under the proposed method and *edgeR* changes. The plot demonstrates the benefit of exploiting position-level data under the proposed method. The other four methods uses gene-level aggregate counts. Therefore, the inference under the other methods is sensitive to outliers. In contrast, the proposed model combines position-level relative gene expression by downweighting some outlying positions as shown in Figure 4. This leads to improved inference when aberrant counts are recorded for some positions under either one of the two conditions. In summary, the ROC curve suggests possible advantages of the proposed method for inference on the differential expression.

We perform further comparison by varying δ_1 and δ_2 and examine how the performance of the five methods changes. We let $\delta = \delta_1 = \delta_2 = 0.4$, 0.6 or 0.8. For each value of δ, we conduct 30 simulations and compare areas under ROC curves. Table 1(a) shows the means and standard deviations of the areas under the ROC curves. For each value of δ, we conduct 30 simulations and compare areas under ROC curves. Table 1 shows the means and standard deviations of the areas under the ROC curves. For all methods in the comparison, area under the curve increases with δ. Among the five methods, the proposed model shows the most favorable performance. Furthermore, we increase the number of replicates ($K = 6$) and the results are summarized in Table 1(b). All the methods appear to be improved in performance with larger K. We note that for small δ increasing the number of replicates without explicitly modeling outlying positions does not offer an large improvement. The comparison highlights the advantage of statistical modeling that can exploit information about position-level variation and improve inference by downweighting positions that are judged to be outliers relative to this variation.

In addition, we examine the impact of the normalization on the reported inference of the differential expression. We first apply the trimmed mean of M-values normalization method (TMM method) proposed by Robinson and Oshlack (2010) to position-level read counts and fix $\bar{\xi_k}$ at 0 (= logit(0.5)). Figure 6b shows the comparison of the resulting inference under the model with fixed $\bar{\xi_k}$ and position-level read counts normalized by the TMM method to the inferences with the proposed method and *edgeR*. From the figure, using TMM for data pre-processing and fixing $\bar{\xi_k}$ deteriorates the performance of the proposed model, although it still outforms the *edgeR*.

3.2 Simulation 2

In this subsection, we do not assume that data are generated from the proposed model. Instead, we simulate expression data based on the yeast RNA-seq data in Section 1.2 and the analysis results from *edgeR*. Specifically, we simulate 400 differentially expressed genes and 1600 non-differentially expressed genes. We first apply *edgeR* to find *p*-values for all genes in the yeast data set. We pool n_{kij} and N_{kij} from 100 genes having the least significant *p*-values, and randomly sample n_{kij} and N_{kij} from this pool to generate counts for genes that are non-differentially expressed under the simulation truth. The number of positions, J_i is determined by random sampling of observed J_i in the yeast data set. For differentially expressed genes, we sample N_{kij} as we do for non-differentially expressed genes, but we generated p_{kij} from beta distributions and sample n_{kij} from Bin(N_{kij}, p_{kij}). For under-expressed genes, the beta distributions are Be(2, 5) and Be(3, 4) for replicate 1 and 2, respectively, and for overexpressed genes, Be(5, 2) and Be(6, 1) for replicate 1 and 2, respectively.

To study how outlying positions affect the statistical inference, we replace the counts n_{kij} for some randomly selected positions with outlier values. We increase the fraction of such outliers gradually from 0% to 10%, 20% and 30%. To generate the outliers, we first sample w_{kij} independently from $\mathrm{Ber}(\pi_0^w)$, $\pi_0^w = 1$, 0.9, 0.8, and 0.7, respectively. Only when $w_{kij} = 0$, we generate $p_{kij} \sim \mathrm{Be}(\alpha_{\lambda_i}^0, \beta_{\lambda_i}^0)$ with $(\alpha_{\lambda_i}^0, \beta_{\lambda_i}^0) = (1, 1/10)$ and (1/10, 1) for $\lambda_i = -1$ and 1, respectively. For $\lambda_i = 0$, we let $(\alpha_{\lambda_i}^0, \beta_{\lambda_i}^0)$ be (1, 1/10) or (1/10, 1) with equal probability. Then we generate outlying n_{kij} from Bin(N_{kij}, p_{kij}) and substitute the outlying n_{kij} for the positions with $w_{kij} = 0$.

To proceed with inference on λ_i, we specified priors assuming that approximately 15% of the positions-specific counts are outliers, and that approximately 20% genes are differentially expressed. We estimated and fixed η_{ki} as described in Appendix B. We fixed the hyperparameters similar to Section 3.1. We carried out posterior MCMC simulation by iterating over all complete conditionals for 5,000 iterations, discarding the first 3,000 iterations as burn-in.

We again compare inference under the proposed model to *edgeR* (Figure 7). Considering the small number of replicates, we again use a common dispersion parameter for *edgeR*.

Comparing ROC curves across simulations with varying proportions of outliers we find that the ROC curves deteriorate with increasingly larger numbers of outliers, as expected. Comparing ROC curves under the proposed model versus *edgeR* for small numbers of outliers (0% outliers) we find that modeling position-level counts in the proposed model does not lead to much improvement compared to modeling gene-level counts of *edgeR*. However, the proposed model performs better as more outlying positions are added. Note that the comparison is biased in favor of *edgeR* since the nondifferentially expressed genes in the simulation truth are determined by the inference under *edgeR*.

4 Yeast Data

We illustrate the proposed model with the RNA-seq yeast data set that was briefly introduced in Section 1.2. Since we observed very low read counts at most positions, we aggregated counts in windows of 50 nt each. We considered $I = 1,016$ genes with $J_i \geq 5$ positions in both replicates for analysis and discarded the remaining 269 for lack of information.

To fit the proposed model to the data, we specified priors as follows: Assuming that around 10% of genes are differentially expressed and around 5% of position-specific counts are outliers, we use $\pi^w_{k\cdot i} \sim \text{Be}(19, 1)$ and $\pi^\lambda \sim \text{Dir}(1, 18, 1)$. We estimated and fixed η_{ki} (see Appendix B). We initialized ξ_{ki} as logit($\hat{\mu}_{ki}$) similar to estimate η_{ki}. We fixed the hyperparameters similar to Section 3.1. We ran the MCMC simulation by iterating over all complete conditionals for 5,000 iterations, discarding the first 3,000 iterations as burn-in.

Figure 8a plots the posterior probabilities of differential expression, $\hat{p}_i = \Pr(\lambda_i \neq 0 \mid \text{data})$. Some genes report large posterior probabilities $\hat{p}_i$ indicating differential expression. Figures 8bc plot the posterior means $\hat{\xi}_{ki} = E(\xi_{ki}|\text{data})$ for each replicate. The three dashed horizontal lines mark posterior means for $(\bar{\xi}_k + \delta_{1k})$, $\bar{\xi}_k$ and $(\bar{\xi}_k - \delta_{-1k})$, respectively. The genes beyond or very close to the lower and upper dashed lines can be considered as differentially expressed. Genes that are reported as differentially expressed are marked by crosses. We will discuss later how the list of reported genes is determined.

Note that the lines for $\bar{\xi}_k$ in panels bc are away from the zero level (equivalent to logit(0.5)). The non-zero values adjust for the imbalance in the total counts across the two conditions separately for each of the two replicates.

Figure 9ab plot the marginal posterior probabilities $\hat{p}_i$ against the empirical estimate of relative expression, r_{ki} for each replicate k. The plot illustrates that $\hat{p}_i$ borrows strength across replicates, and it agrees with ad-hoc estimates r_{ki} for most genes. But there are some genes where $\hat{p}_i$ disagrees with and improves ad-hoc inference with r_{ki} for replicate 1, but agrees for replicate 2. For example, see the gene marked with triangle in Figure 9ab. In Figure 10 we explore possible reasons for this. We present summaries for two selected genes to illustrate how the proposed model combines relative counts across replicates and how $\hat{p}_i$ adjusts r_{ki}. In each panel of the figure, the plots in the first column show r_{kij} along positions. The dashed line indicates the posterior mean $\hat{\xi}_{ki}$, and the dotted line shows the empirical estimate r_{ki}. The line for $\hat{\xi}_{ki}$ is plotted at logit$^{-1}\hat{\xi}_{ki}$ to map to the unit scale. The second column plots the posterior probability $\hat{w}_{kij} = \Pr(w_{kij} = 1 \mid \text{data})$ of position j not being an outlier along positions. The first and second row plot replicate 1 and 2, respectively.

Comparison of the two rows in each panel explains the observed discrepancies in r_{ki} and $\hat{p}_i$ for replicate 1. The relatively larger sample sizes N_{kij} in replicate 2 contain more information about relative expression levels. Using the information from replicate 2, the model concludes that some of the unusual values of r_{ki} under $k = 1$, as shown in Figure 10b, are due to outliers in r_{kij}, including even some positions with large total read counts N_{kij}. The estimated probabilities $\hat{w}_{kij}$ for those positions are reduced, leading to downweighting

of the corresponding r_{kij} in the inference for the gene-specific indicators λ_i for differential expression, and thus for $\hat{p_i}$.

The computation of posterior probabilities $\hat{p_i} = \Pr(\lambda_i \neq 0 \mid \text{data})$ is only half the desired inference. The posterior probabilities do not yet determine the list of genes to be reported for differential expression. The selection of genes to be flagged for differential expression is the second step. Let $d_i \in \{0, 1\}$ denote an indicator for reporting gene i as differentially expressed. A natural decision is to report the genes with highest probability of differential expression, i.e., $d_i = I(\hat{p_i} > \kappa)$, for some threshold κ. In fact, d_i can be shown to be the optimal Bayes rule under several formalizations of the decision problem (Müller *et al.*, 2004). Newton *et al.* (2004) propose to control posterior expected false discovery rate (FDR) to determine the cutoff κ. Let $D = \Sigma_i d_i$ denote the number of reported genes, and let FDR = $(1/D)\ \Sigma_i\ (1 - |\lambda_i|)d_i$ denote the fraction of false positives, i.e., the number of wrongly reported genes, relative to D. Here $|\lambda_i|$ is the (unknown) true indicator of differential expression of gene i. The posterior expected FDR is easily evaluated as $\overline{\text{FDR}} = E(\text{FDR} \mid \text{data}) = (1/D) \sum (1 - \hat{p}_i) d_i$. For clarification,, we note that most authors consider the frequentist expectation E(FDR) with respect to repeat experimentation under an assumed true scenario. See, for example, Müller *et al.* (2004) for more discussion and references. But taking a Bayesian perspective to estimate $\hat{p_i}$ and $\hat{\xi_i}$ it is natural to consider the posterior expectation $\overline{\text{FDR}}$ instead. Newton *et al.* (2004) propose to select κ to achieve a desired level of $\overline{\text{FDR}}$, say $\overline{\text{FDR}} \leq \rho$.

Figure 9c summarizes the FDR implied by decision rules of the type $d_i = I(\hat{p_i} \geq \kappa)$. Reducing κ leads to increasing (posterior) estimated FDR and increasingly larger lists of differentially expressed genes. The figure plots $\overline{\text{FDR}}$, the posterior expected FDR, versus the number of reported genes. For example, for $\overline{\text{FDR}} \leq 0.10$ the rule reports D = 239 differentially expressed genes. The rule corresponds to a threshold κ = 0.53.

We compare the identification of differential gene expression under the propose method to that under *edgeR*. We use the q−values to detect differentally expressed genes with *edgeR* to avoid a problem due to the simultaneous testing of many genes (Efron *et al.*, 2001). For $\overline{\text{FDR}} \leq 0.10$, the decision rule under *edgeR* is to identify genes with q−value less than 0.10 as differentially expressed genes. Following the rule, *edgeR* detects 682 genes as differentially expressed genes. Among them, 238 genes are the genes detected by the proposed method. The result is graphically illustrated in Figure 11a using a Venn Diagram. Figure 11b shows a plot of ranks of $\hat{p_i}$ versus ranks of 1−q−values. The plot demonstrates that a gene with a large posterior probability of differential gene expression under the proposed method tends to have smaller q−values under *edgeR*.

5 Discussion

The model developed in this paper borrows strength across genomic positions within a gene, and across replicates and genes. The model robustifies inference for gene expression through downweighting position-level outliers. Furthermore, in the proposed model the indicator for differential expression, relative expression levels, varying number of reads in different samples, and indicators for position-level outliers are all incorporated in the same model.

Through this integration, the estimate of the relative expression level becomes more robustfied. We illustrate this through a simulation study and the analysis of a yeast experiment.

The proposed model assumes that the two conditions have the same number of replicates, and randomly matches pairs to conduct the analysis. In the case where the number of replicates differs by conditions, we suggest a random split of one sample into two samples.

We focused on the comparison of two conditions in this paper, but the proposed model allows an easy extension for inference with RNA-seq data also under multiple conditions. For this extension, one may use a multinomial likelihood with a Dirichlet distribution as a prior to model read counts from the multiple conditions. Other possible extensions are to incorporate covariates and to relax the normal assumption for ξ_{ki}. To utilize information from covariates on differential gene expression, one can introduce a functional form between the beta/Dirichlet probabilities and the covariate values such as logit or probit model. To accommodate possible overdisperion, one may use heavier tailed distributions or relax the parametric assumption on the distribution of ξ_{ki}.

Acknowledgments

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

Appendix A: Probability model

In this appendix, we present mathematical details discussed in Section 2. For the likelihood we assume that

$$n_{kij} \sim \text{Poi}(\phi_{0kij}), \text{ and } m_{kij} \sim \text{Poi}(\phi_{1kij}),$$

where n_{kij} and m_{kij} represent read counts at position j of gene i for replicate k under condition 0 and 1, respectively. By letting $N_{kij} = n_{kij} + m_{kij}$, we have

$$n_{kij}|N_{kij} \sim \text{Bin}(N_{kij}, p_{kij}), \text{and } N_{kij} \sim \text{Poi}(\phi_{kij}),$$

where $\varphi_{kij} = \varphi_{0kij} + \varphi_{1kij}$ and $p_{kij} = \varphi_{0kij}/(\varphi_{0kij} + \varphi_{1kij})$.

We consider independent gamma priors for φ_{0kij} and φ_{1kij}, i.e., Ga(α_{ki}, θ_{ki}) and Ga(β_{ki}, θ_{ki}). Then we have $p_{kij} \sim$ Be(α_{ki}, β_{ki}) and $\varphi_{kij} \sim$ Ga($\alpha_{ki} + \beta_{ki}$, θ_{ki}), and p_{kij} and φ_{kij} are a priori independent.

The resulting joint posterior factors as

$$\text{P}(p_{kij}, \phi_{kij}|n_{kij}, N_{kij}) = \text{P}(p_{kij}|n_{kij}, N_{kij})\text{P}(\phi_{kij}|N_{kij}).$$

The marginal posterior, $P(p_{kij}|n_{kij}, N_{kij})$, is the model of interest. We therefore focus on modeling of n_{kij} conditional on N_{kij} only.

Appendix B: Estimation of ηkij

For the analyses in Sections 3 and 4, we estimate and fix η_{ki}. In this appendix we illustrate a way of estimating η_{ki}. First, we assume that π^w_{ki} and $\boldsymbol{\pi}^\lambda$ are fixed at their prior mean for simplicity. Thus, $p_{kij} \sim \text{Be}(\alpha_{ki}, \beta_{ki})$ with probability π^w_{ki}. With probability $(1-\pi^w_{ki}), p_{kij}$ follows a mixture of the three beta distributions where beta mixture components are $\text{Be}(\alpha^0_\ell, \beta^0_\ell)$, $\ell \in \{-1, 0, 1\}$ in Equation (1) and their weight are π^λ_ℓ. Then we equate sample moments based on empirical ratios, r_{ki} and r_{kij}, with γ_{ki} and μ_{ki}, and find $\hat{\gamma_{ki}}$ and $\hat{\mu_{ki}}$ as follows:

$$r_{ki} = \pi^w_{ki}\hat{\mu}_{ki} + (1-\pi^w_{ki})d,$$

and

$$\begin{aligned}
\text{var}(r_{kij}) &= \text{var}(\text{E}(r_{kij}|p_{kij})) + \text{E}(\text{var}(r_{kij}|p_{kij})) \\
&= (1-\tfrac{1}{N_{kij}})\text{var}(p_{kij}) + \tfrac{1}{N_{kij}}\text{E}(p_{kij})(1-\text{E}(p_{kij})) \\
&= (1-\tfrac{1}{N_{kij}})\{\text{var}(\text{E}(p_{kij}|w_{kij})) + \text{E}(\text{var}(p_{kij}|w_{kij}))\} + \tfrac{1}{N_{kij}}\text{E}(p_{kij})(1-\text{E}(p_{kij})) \\
&= (1-\tfrac{1}{N_{kij}})\{(\hat{\mu}_{ki}-d)^2\pi^w_{ki}(1-\pi^w_{ki}) + (1-\pi^w_{ki})e + \pi^w_{ki}\hat{\mu}_{ki}(1-\hat{\mu}_{ki})\tfrac{1}{\hat{\gamma}_{ki}+1}\} + \tfrac{1}{N_{kij}}\text{E}(p_{kij})(1-\text{E}(p_{kij})),
\end{aligned}$$

where $\text{var}(r_{kij})$ is the sample variance of the r_{kij} across positions j within each gene (i) and replicate (k), and $d = \pi^\lambda_{-1}1/3 + \pi^\lambda_0 1/2 + \pi^\lambda_1 2/3$ and $e = \pi^\lambda_{-1}4/45 + \pi^\lambda_0 1/8 + \pi^\lambda_1 4/45$ are the mean and the variance of p_{kij} with $w_{kij} = 0$, respectively. We estimate $\text{E}(p_{kij})$ with r_{ki} and solve for $\hat{\gamma_{ki}}$ and $\hat{\mu_{ki}}$. We fix $\eta_{ki} = \log(\gamma_{ki})$. We use $\text{logit}(\hat{\mu_{ki}})$ as an initial value of ξ_{ki}. One may exclude r_{kij} with very small N_{kij} to obtain more reasonable estimates of η_{ki}. Specifically, for simulation in Section 3.1, we use r_{kij} with $N_{kij} > 3$. For simulation in Section 3.2 and the yeast data analysis in Section 4 we use r_{kij} with $N_{kij} > 1$ only.

References

Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology. 2010; 11(10)

Auer PL, Doerge RW. Statistical Design and Analysis of RNA Sequencing Data. The Genetics Society of America. 2010; 185:405–416.

Baggerly KA, Deng L, Morris JS, Aldaz CM. Overdispersed logistic regression for sage: Modelling multiple groups and covariates. BMC Bioinformatics. 2004; 5

Balwierz PJ, Carninci P, Daub CO, Kawai J, Hayashizaki Y, Belle WV, Beisel C, van Nimwegen E. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. Genome Biology. 2009; 10(7)

Bullard JH, Purdom E, H KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010; 11

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput dna sequencing. Nucleic Acids Research. 2008; 36:16.

Efron B, Tibshirani R, Storey J, Tusher V. Empirical Bayes Analysis of a Microarray Experiment. Journal of the American Statistical Association. 2001; 96:1151–1160.

Hansen KD, Brenner SE, Ducoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Research. 2010; 38

Hardcastle TJ, Kelly KA. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010

Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JI, Hickenbotham M, Huang W, Magrini VJ, Richt RJ, Sander SN, Stewart DA, Stromberg M, Tsung EF, Wylie T, Schedl T, Wilson R, Mardis E. Whole-genome sequencing and variant discovery in c. elegans. Nature Methods. 2008; 5:183–188. [PubMed: 18204455]

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):218–223. [PubMed: 19213877]

Jefferys W, Berger J. Ockham's razor and bayesian analysis. American Scientist. 1992

Ji H, Liu XS. Analyzing 'omics data using hierarchical models. Nature Biotechnology. 2010

Langmead B, Trapnel C, Pop M, Salzberg SL. Ultrafast and memoryefficient alignment of short dna sequences to the human genome. Genome Biology. 2009; 10

Lee, J.; Müller, P.; Lian, S.; Cai, G.; Ji, Y. Tech rep. Department of Biostatistics; UT M.D. Anderson: 2011. On differential gene expression using rna-seq data.

Li H, Ruan J, Durbin R. Mapping short dna sequencing reads and calling variants using mapping quality scores. Genome Research. 2008a; 18(11):1851–1858. [PubMed: 18714091]

Li J, Jiang H, Wong WH. Modeling non-uniformity in short-read rates in RNA-Seq data. Genome Biology. 2010; 11

Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. Bioinformatics. 2008b

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; 18:1509–1517. [PubMed: 18550803]

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods. 2008; 5:621–628. [PubMed: 18516045]

Müller P, Parmigiani G, Robert C, Rousseau J. Optimal Sample Size for Multiple Testing: the Case of Gene Expression Microarrays. Journal of the American Statistical Association. 2004; 99:990–1001.

Newton MA, Noueiry A, Sarkar D, Ahlquist P. Detecting Differential Gene Expression with a Semiparametric Hierarchical Mixture Method. Biostatistics. 2004; 5:155–176. [PubMed: 15054023]

Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results. Genome Biology. 2010; 11(12)

Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. Biology Direct. 2009; 4

Robert CP, Rousseau J. A Mixture Approach to Bayesian Goodness of Fit. Les cahiers du CEREMADE(2002-9). 2004

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26(1):139–140. [PubMed: 19910308]

Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology. 2010; 11(3)

Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics. 2007; 23(21):2881–2887. [PubMed: 17881408]

Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M. Shrimp: accurate mapping of short color-space reads. PLOS Computational Biology. 2009; 5(5):e1000386.10.1371/journal.pcbi.1000386 [PubMed: 19461883]

Schwartz S, Oren R, Ast G. Detection and removal of biases in the analysis of next-generation sequencing reads. PLoS ONE. 2011; 6:1.

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch B, Siddiqui A, Lao K, Surani M. mRNA-Seq whole-transcriptome analysis of a single cell. Nature Methods. 2009

Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics. 2009; 26(1):136–138. [PubMed: 19855105]

Wu Z, Jenkins BD, Rynearson TA, Dyhrman ST, Saito MA, Mercier M, Whitney LP. Empirical bayes analysis of sequencing-based transcriptional profiling without replicates. BMC Bioinformatics. 2010; 11

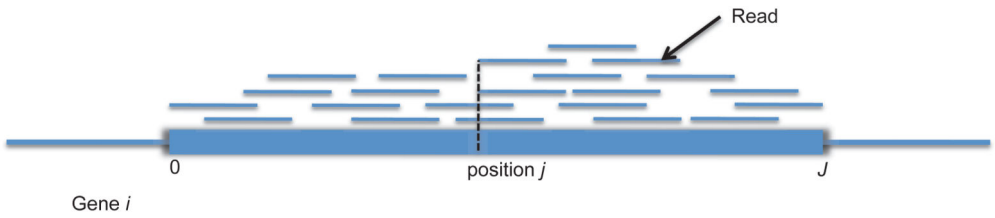


Figure 1.

RNA-seq data. The data is summarized as counts of mapped reads. The read counts at position *j* corresponds to the number of reads whose starting position is position *j*. For example, gene *i* has *J* genomic positions (or bases), and position *j* of gene *i* has count 2.

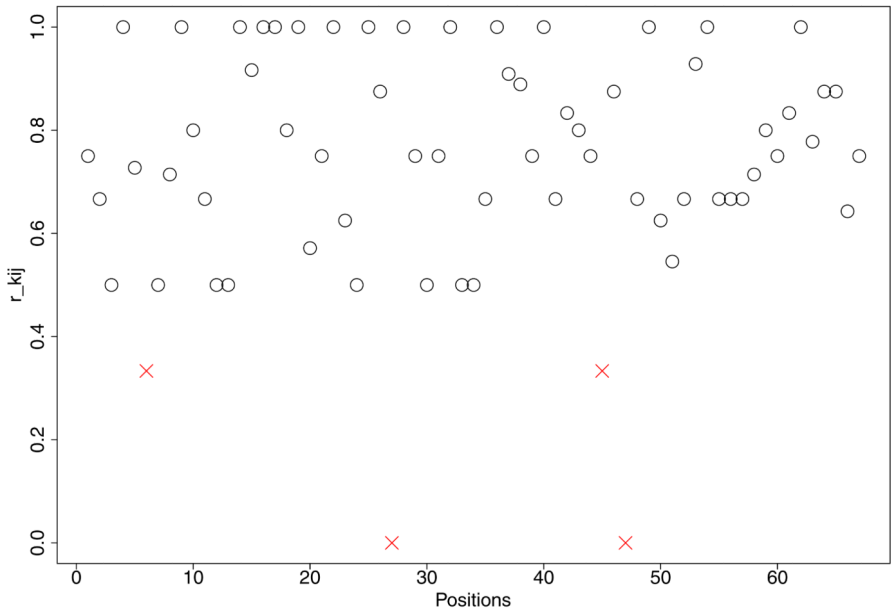


Figure 2.

Plot of r_{kij}, $j = 1, \ldots, J_i$ for gene $i = 16$ and replicate $k = 2$. Potential outliers are marked as crosses. Inference downweights outliers in r_{kij} by introducing indicators w_{kij}.

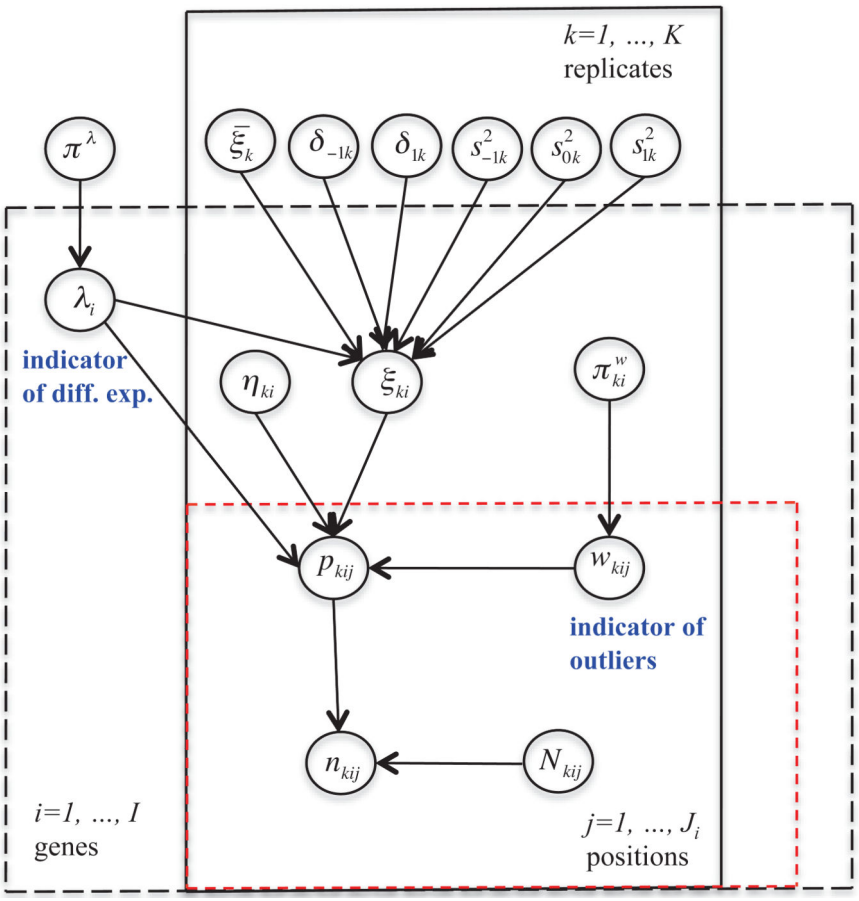
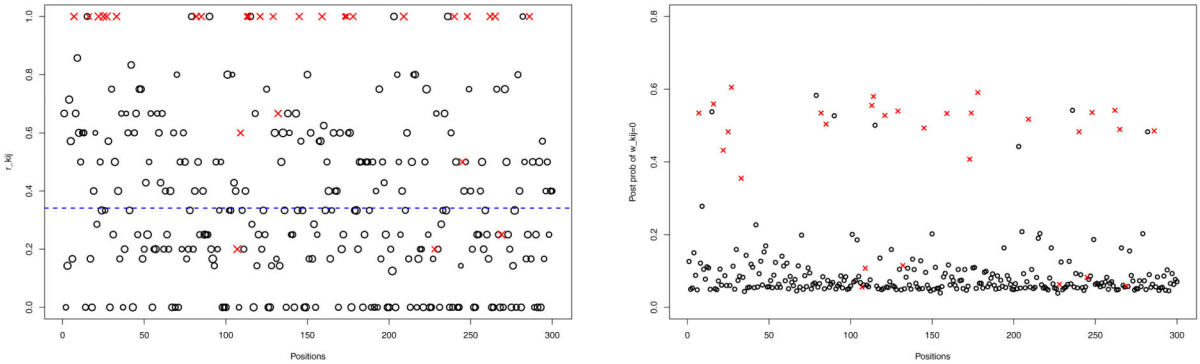


Figure 3.
The proposed hierarchical model for RNA-seq data.



(a) Plot of r_{kij} (b) Plot of $\Pr(w_{kij} = 0 \mid \text{data})$

Figure 4.

Simulation 1: We plot the empirical fraction r_{kij} (left) and posterior probabilities $\Pr(w_{kij} = 0 \mid \text{data})$ of being an outlier (right) for each of 300 positions. This gene is truly under-expressed. Therefore, the outlier positions are represented by crosses with large values of r_{kij} and large posterior probabilities of $w_{kij} = 0$. The dashed line on the left panel represents the true value of $\mu_{ki} = \exp(\xi_{ki})/(\exp(\xi_{ki}) + 1)$.

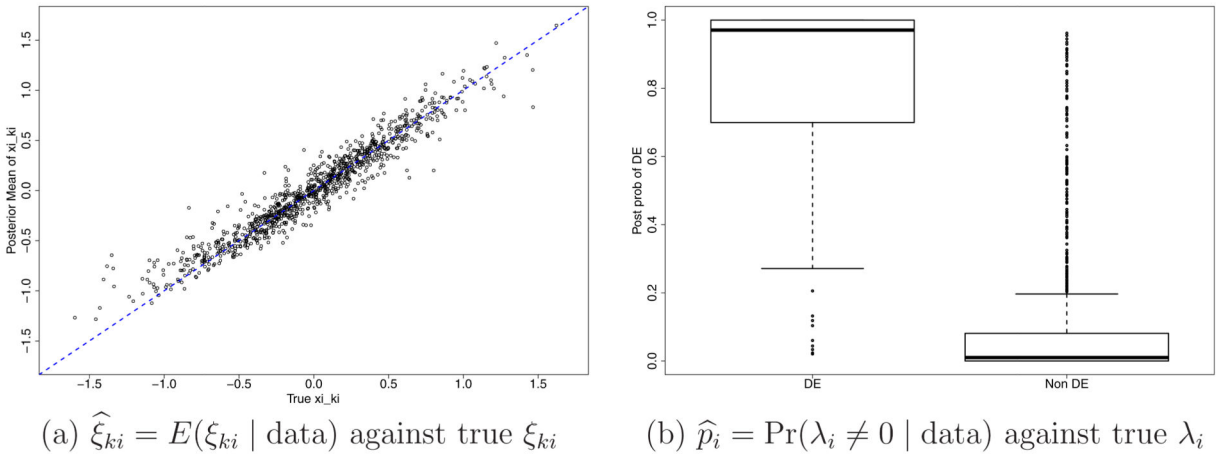
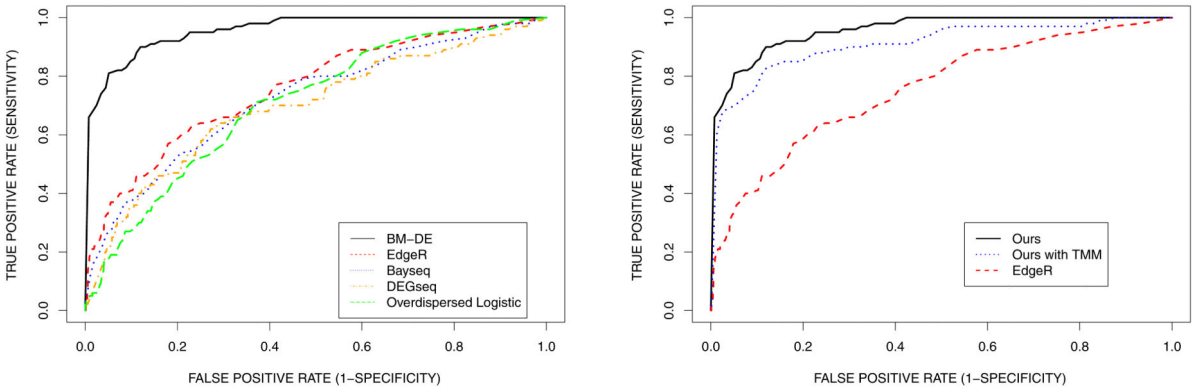


Figure 5.

Simulation 1: Posterior means $\hat{\xi_{ki}} = E(\xi_{ki} \mid \text{data})$ and posterior probability of differential gene expression $\hat{p_i} = \Pr(\lambda_i \neq 0 \mid \text{data})$ in simulation 2. In panel (a) the estimated $\hat{\xi_{ki}}$ for replicate $k = 1$ are close to the truth ξ_i on the 45 degree line (dotted line). Panel (b) shows boxplots of $\hat{p_i}$ by the truth of differential expression.



(a) Comparison of the five methods (b) ROC curve with fixed $\bar{\xi}_k$

Figure 6.

Simulation 1: (a) ROC curves for identification of differential gene expression under the proposed method, *edgeR*, *Bayseq*, *DEGseq* and the overdispersed logistic regression model. (b) ROC curves under the proposed method but with fixed $\bar{\xi}_k$ and position-level read counts normalized with the TMM method (dotted line), the proposed method (solid line) and *edgeR*(dashed line).

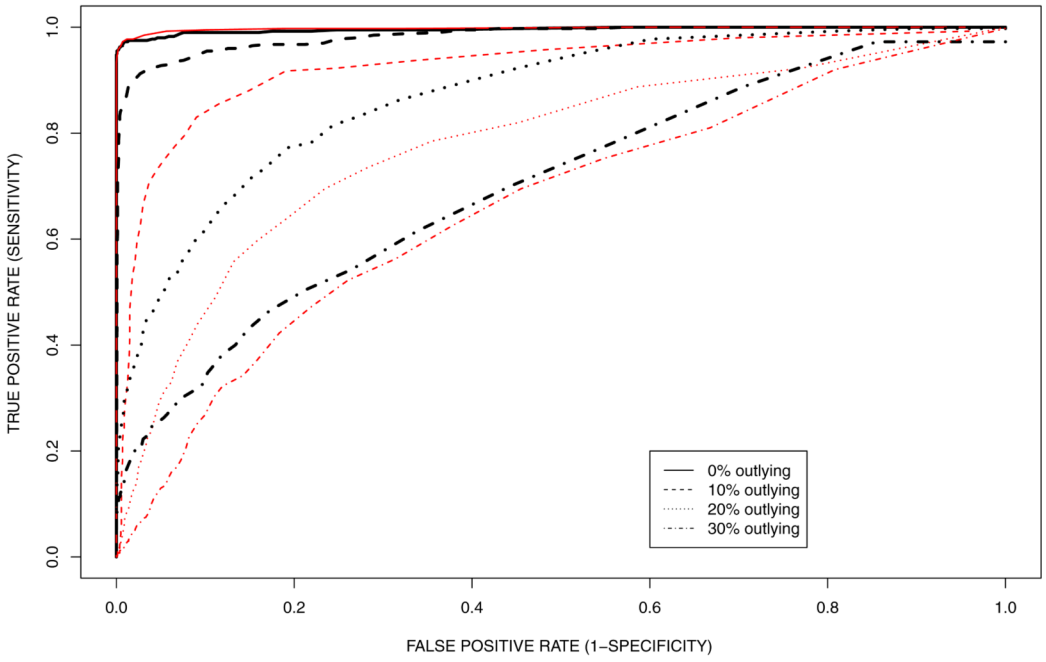
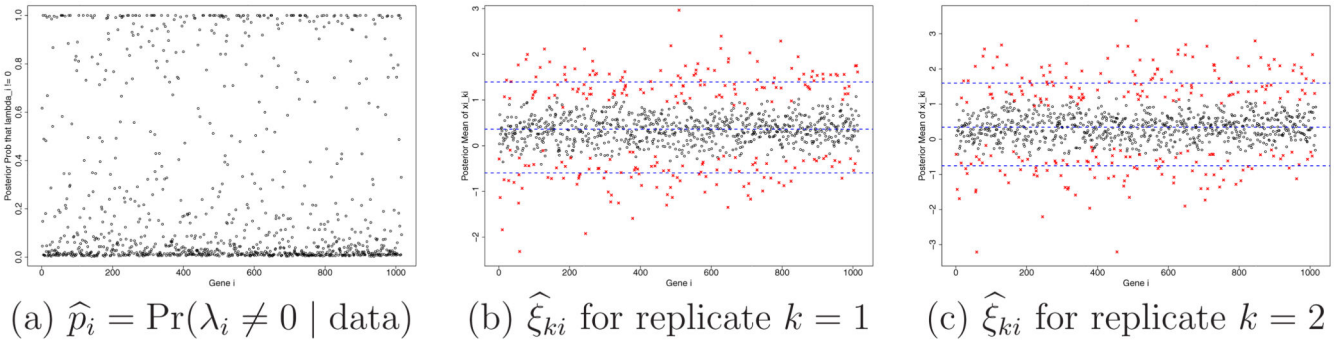


Figure 7.

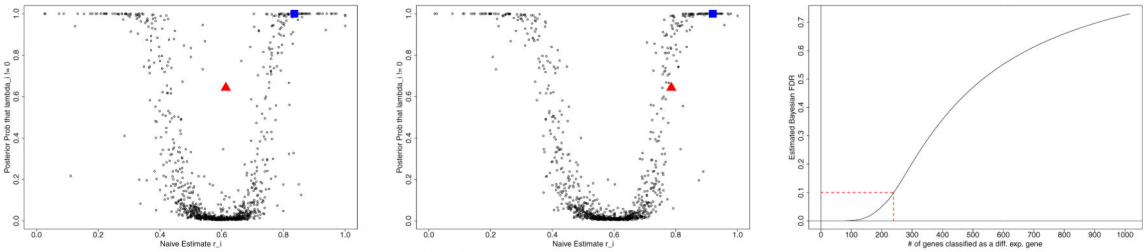
Simulation 2: ROC curves for identification of differential gene expression under the proposed method (thick black lines) and *edgeR* (thin red lines) proposed by Robinson et al. (2010). In the simulation, outlying positions are added gradually from 0% position to 30% positions by 10%.



(a) $\widehat{p}_i = \Pr(\lambda_i \neq 0 \mid \text{data})$ (b) $\widehat{\xi}_{ki}$ for replicate $k = 1$ (c) $\widehat{\xi}_{ki}$ for replicate $k = 2$

Figure 8.

Yeast data: Posterior probability of differential expression, $\widehat{p_i} = \Pr(\lambda_i \neq 0 \mid \text{data})$ (panel a) and the posterior mean of relative gene expression over the two conditions, $\widehat{\xi_{ki}} = E(\xi_{ki} \mid \text{data})$ for $k = 1, 2$ (panels b and c). The genes marked in crosses are genes identified as differentially expressed at $\overline{\text{FDR}} \leq 0.10$.



(a) $\widehat{p}_i$ versus r_{ki} for $k = 1$ (b) $\widehat{p}_i$ versus r_{ki} for $k = 2$ (c) $\overline{\text{FDR}}$ versus D

Figure 9.

Yeast data: Posterior probabilities $p_i \hat{} = \Pr(\lambda_i \neq 0 \mid \text{data})$ plotted against r_{ki} (panels a and b). The genes indicated with triangle and square are discussed in Figure 10. Panel (c) plots posterior expected FDR against the number D of genes reported as differentially expressed.

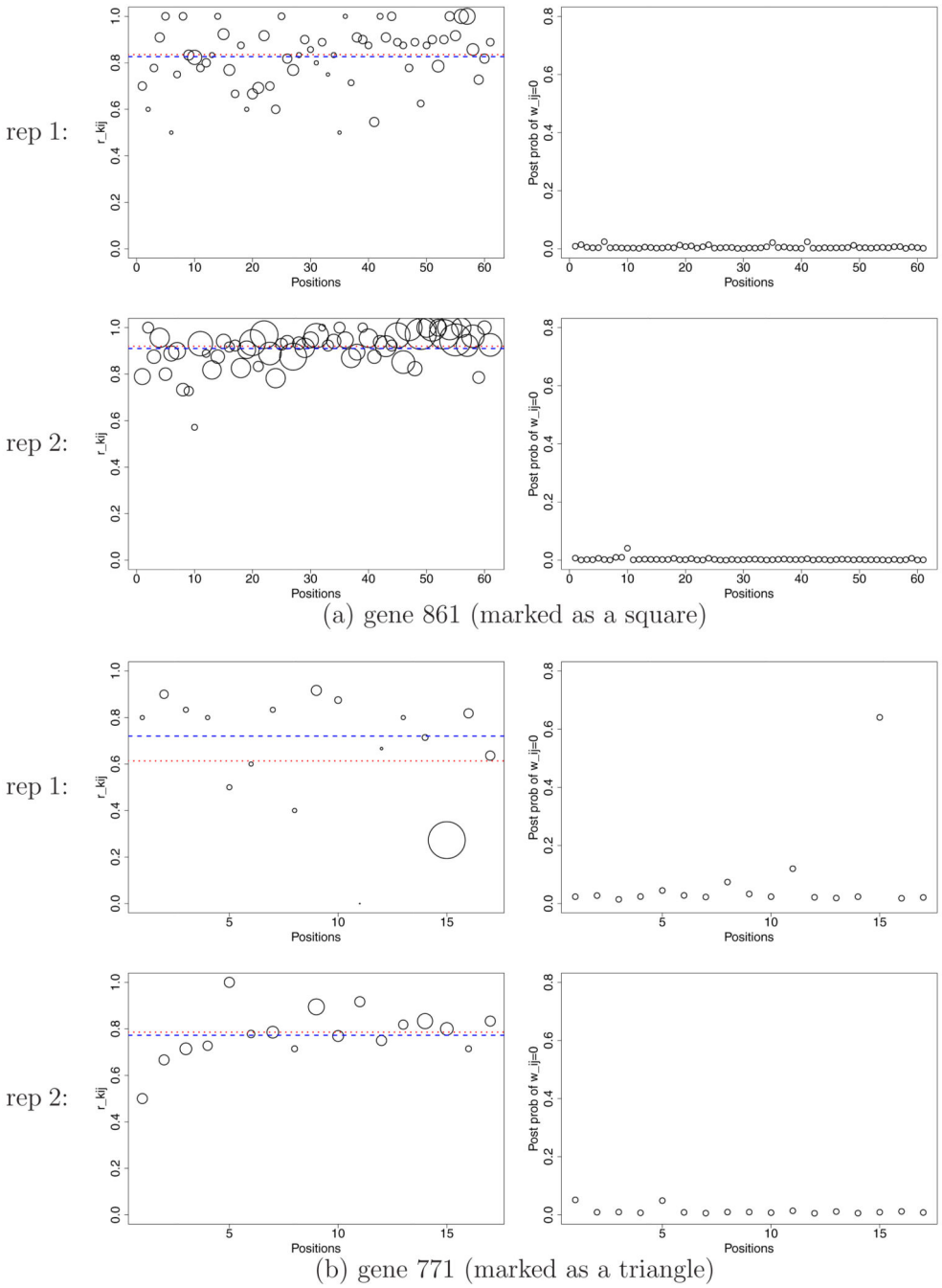
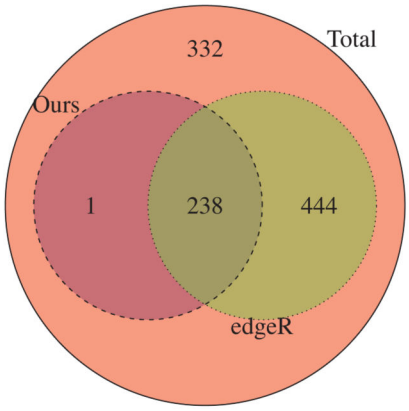
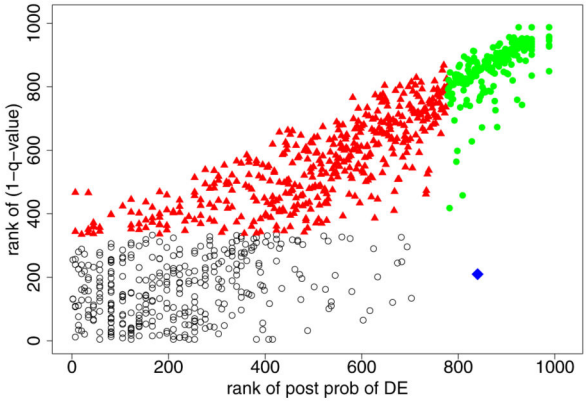


Figure 10.

Yeast data: Inference summaries for the genes marked as a square and a triangle in Figure 9 (a) and (b). In each panel, the first column plots r_{kij} where the areas of the circles are proportional to N_{kij}. The dotted line indicates $^{31}r_{ki}$. The dashed line shows the posterior mean $\hat{\xi_{ki}}$ (plotted at $\text{logit}^{-1}\hat{\xi_{ki}}$ to map to the unit scale). The second column plots $\hat{w}_{kij}$. The first and second rows are plots for replicate 1 and 2, respectively. Note the discrepancy between posterior inference and r_{ki} for replicate 1 of gene 771. This discrepancy is mainly due to outlying positions such as position 15.



(a) Venn Diagram



(b) ranks of $\hat{p}_i$ vs ranks of $q-$values

Figure 11.

Yeast data: (a) Venn Diagram of differentially expressed genes in the yeast data set under the propsed model and *edgeR*. (b) Scatterplot of ranks of posterior probability of differential expression under the BM-DE vs ranks of (1-$q-$values) under *edgeR*. Genes in circle are genes not detected as differentially expressed under neither of the two methods, genes in triangle under *edgeR* only, genes in filled circles under both, genes in diamonds under the proposed method only.

Table 1

Simulation 1: The mean of areas under ROC curves for each of the five methods in comparison with its standard deviation. The numbers are based on 30 iterations.

δ	BM-DE	EdgeR	Bayseq	DEGseq	Overdispersed logistic
0.4	0.923 (0.019)	0.582 (0.034)	0.574 (0.035)	0.547 (0.035)	0.573 (0.033)
0.6	0.947 (0.014)	0.756 (0.023)	0.729 (0.023)	0.677 (0.023)	0.714 (0.024)
0.8	0.970 (0.009)	0.894 (0.024)	0.856 (0.023)	0.800 (0.024)	0.833 (0.024)
(a) $K = 3$					
0.4	0.975 (0.009)	0.659 (0.035)	0.653 (0.031)	0.570 (0.033)	0.650 (0.031)
0.6	0.983 (0.009)	0.874 (0.021)	0.857 (0.020)	0.712 (0.021)	0.847 (0.020)
0.8	0.993 (0.005)	0.975 (0.009)	0.961 (0.011)	0.837 (0.021)	0.951 (0.012)
(b) $K = 6$					