Published in final edited form as:

J Agric Biol Environ Stat. 2015 December ; 20(4): 555–576. doi:10.1007/s13253-015-0227-0.

A Semi-parametric Bayesian Approach for Differential Expression Analysis of RNA-seq Data

Fangfang Liu, Chong Wang, and Peng Liu

Department of Statistics Iowa State University Ames, IA 50011

Abstract

RNA-sequencing (RNA-seq) technologies have revolutionized the way agricultural biologists study gene expression as well as generated a tremendous amount of data waiting for analysis. Detecting differentially expressed genes is one of the fundamental steps in RNA-seq data analysis. In this paper, we model the count data from RNA-seq experiments with a Poisson-Gamma hierarchical model, or equivalently, a negative binomial (NB) model. We derive a semi-parametric Bayesian approach with a Dirichlet process as the prior model for the distribution of fold changes between the two treatment means. An inference strategy using Gibbs algorithm is developed for differential expression analysis. The results of several simulation studies show that our proposed method outperforms other methods including the popularly applied edgeR and DESeq methods. We also discuss an application of our method to a dataset that compares gene expression between bundle sheath and mesophyll cells in maize leaves.

Keywords

Bayesian; Differential expression; Dirichlet process; Posterior probability; RNA-seq

1. INTRODUCTION

In recent years, next-generation sequencing (NGS) technologies have become widely applied in genomic studies. The application of NGS technologies to studying the abundance of messenger RNA (mRNA) molecules is called RNA sequencing (RNA-seq). Since the first set of RNA-seq papers published in 2008, this technology has revolutionized the way agricultural biologists study gene expression as well as generated a tremendous amount of data waiting for analysis.

In a typical RNA-seq experiment, all the mRNA molecules of each sample are extracted, converted to a library of complementary DNA fragments, and then sequenced with a high-throughput sequencing platform, such as Illumina HiSeq. Millions of short sequences, or reads, are obtained from each sample and then mapped to a reference genome or transcriptome. The count of reads mapped to a given gene measures the expression level of this gene. Detecting differentially expressed (DE) genes, whose expression levels change across treatments/conditions, is a fundamental step and often the major goal of the analysis

of RNA-seq data. RNA-seq experiments generate count data rather than continuous data. Methods developed for microarray data analysis based on Gaussian distributions have been used for log-transformed RNA-seq data, such as MAANOVA in Reeb and Steibel (2013) and limma (Law *et al.* 2014). More commonly, RNA-seq count data are modeled directly using Poisson distributions or negative binomial distributions. It has been reported that Poisson models fit data well when there are only technical replicates (Bullard *et al.* 2010). However, the presence of biological replicates is required to draw meaningful conclusions, and biological variation in addition to technical variation existing in data with biological replicates results in an over-dispersion problem, e.g. the count data show variation greater than expected from Poisson random variables. The negative binomial (NB) model has been proposed to deal with over-dispersed data. Several methods for differential expression have been proposed based on NB models and popularly applied for the analysis of RNA-seq data, including methods by Robinson and Smyth (2007), Anders and Huber (2010) and Hardcastle and Kelly (2010).

In the literature of RNA-seq data analysis, the negative binomial model has been parameterized using a parameter (μ) for the mean and a dispersion parameter (ϕ) so that the variance is modeled as $\mu + \phi\mu^2$. Robinson and Smyth (2007) proposed the edgeR method, in which an estimator for the dispersion parameter is derived that shrinks the dispersion parameter for each individual gene using a weighted likelihood approach. The DESeq method (Anders and Huber 2010) uses a local regression model for the variance as a function of the mean that gives a fitted relationship useful for estimating dispersion parameters. DESeq and edgeR take a classical frequentist hypothesis testing approach and generate p-values, while baySeq (Hardcastle and Kelly 2010) was developed within an empirical Bayesian framework. The baySeq approach also assumes NB distributions for the count data. Additionally, it assumes that the parameters of the NB distributions follow some prior distribution and uses the empirical distribution based on the parameter estimates from all genes as the prior. Then baySeq estimates the posterior probabilities of the model under the null hypothesis and uses these posterior probabilities instead of p-values as measurements of significance for differential expression.

Although there have been simulation studies that compared the above-mentioned popular methods for RNA-seq data analysis (Kvam, Liu, and Si 2012; Sonesson and Delorenzi 2013), the optimality of these testing procedures was not studied until Si and Liu (2013). Si and Liu (2013) provided a theoretical framework to derive optimal tests, where optimal tests were defined as the tests that achieve the maximum of the power averaged across all genes for which null hypotheses are false among those tests that control FDR at the same level. Si and Liu (2013) also derived such an optimal test within their framework, and proposed an approximation to the optimal test. The test they proposed is called the AMAP (approximated most average powerful) test, and was derived through an empirical Bayesian method under parametric priors for the mean parameters in the NB model. In the case of two-treatment comparison, Si and Liu (2013) re-parameterized the two treatment means for each gene by the geometric mean and the ratio of the two means. For simplicity, the ratio of the two treatment means is referred to as the fold change (ρ_g). Si and Liu (2013) used a K-component mixture distribution to model the joint distribution of parameters corresponding to the geometric mean and the logarithm of the fold change. Mixture modeling is very often

used for density estimation and is able to represent mixtures of populations that underlie many real-world datasets. It is, however, quite tricky to pick the number of components in the mixture. Besides, the empirical distribution of all genes could be very irregular and difficult to model by a finite mixture of parametric models. Figure 1 shows the histogram of the empirically estimated log fold change ($\log \hat{\rho}_g$) from a maize dataset (Tausta *et al.* 2014). The shape of the displayed empirical distribution of $\log(\hat{\rho}_g)$ is roughly symmetric, but many local modes with high frequencies spike out of the overall pattern which makes the whole distribution difficult to model with parametric distributions.

Bayesian nonparametric modeling is a sound alternative for density estimation and is often applied to avoid critical dependence on parametric assumptions. Several Bayesian nonparametric methods have been developed (Muller and Fernando 2004), as well as consistency theories of nonparametric Bayes density estimation (Wu and Ghosal 2010). By far the most popular Bayesian nonparametric methods use Dirichlet process priors, and such methods have been adopted for the analyses of differentially expressed genes. For example, Kim *et al.* (2009) constructed a Dirichlet process mixture model for the distribution of gene intensities under two conditions and discussed an application to a microarray dataset.

In this paper we use a Dirichlet process as a nonparametric prior to model the distribution of fold changes between two treatment conditions. Furthermore, we model RNA-seq count data using a Poisson-Gamma hierarchical model, or equivalently, a negative binomial (NB) model. This paper shows how our semi-parametric modeling framework can be adapted to provide meaningful posterior probabilities of sharp or composite null hypothesis. The resulting posterior probabilities have a very natural interpretation and are used to estimate false discovery rate. We also show that the calculated posterior probabilities can be viewed as a different approximation for the optimal test in Si and Liu (2013). Hence, similar to the AMAP test, our procedure is also an approximated optimal test.

The remainder of the paper is organized as follows. Section 2 describes our proposed modeling framework and the prior model. In Section 3 we discuss the MCMC algorithm for posterior inference and FDR estimation. Using several simulation studies, in Section 4, we compare our proposed method to some existing methods and demonstrate that our proposed method improves the detection of differentially expressed genes under simulation settings. In Section 5, we apply our method to a dataset that compares gene expression between bundle sheath and mesophyll cells in maize leaves. In Section 6 we discuss possible extensions of our method and conclude the paper.

2 A SEMI-PARAMETRIC BAYESIAN MODEL

2.1 A Poisson-Gamma Mixture model

We consider data analysis from a general RNA-seq experiment with independent biological replicates for each treatment condition. The negative binomial distribution has been the most popular model for such data analysis, due to its ability to accommodate over-dispersion. In our development of a Bayesian framework, we use a Poisson-Gamma mixture model which is marginally equivalent to the negative binomial model. We use the Poisson-Gamma mixture because it is much easier to obtain conjugate priors in this form than a direct

negative binomial parametrization. The Gamma-Poisson mixture model provides a natural hierarchy for interpreting the generation of sequencing counts. The Gamma distribution models the unobserved true expression levels in each biological sample, and conditioning on the true expression level, the count of reads obtained from the RNA-seq experiment follows a Poisson distribution.

Suppose an RNA-seq experiment is conducted to compare gene expression levels among T treatment conditions and there are n_i biological replicates for the i -th condition, $i = 1, \dots, T$. The resulting RNA-seq dataset includes count data for G genes. For $i = 1, \dots, T, j = 1, \dots, n_i, g = 1, \dots, G$, let Y_{gij} denote the number of reads mapped to gene g from replicate j of condition i . We use the following Poisson-Gamma mixture model to model Y_{gij} .

$$Y_{gij} | \lambda_{gij} \sim \text{Poisson}(S_{ij} \lambda_{gij}), \quad (2.1)$$

$$\lambda_{gij} | \alpha_g, \beta_g, \rho_{gi} \sim \text{Gamma}(\alpha_g, \beta_g \rho_{gi}). \quad (2.2)$$

Here the first treatment is set as a reference treatment, i.e., $\rho_{g1} = 1$. In this model, α_g is the shape parameter, and the product of β_g and ρ_{gi} serves as the rate parameter, so the mean of the Gamma distribution in (2.2) is $\alpha_g / (\beta_g \rho_{gi})$. The relationship between the variance of Y_{gij} and expectation of Y_{gij} is given by

$$\text{var}(Y_{gij}) = E(Y_{gij}) + [E(Y_{gij})]^2 / \alpha_g. \quad (2.3)$$

So the dispersion parameter ϕ_g is the reciprocal of the shape parameter α_g , i.e., $\phi_g = 1/\alpha_g$. S_{ij} is a normalization factor that adjusts for varying sequencing depth and potentially other technical effects across the replicates. There are several proposed methods to normalize RNA-seq data that involve estimating the normalization factor in different ways. For example, we could estimate S_{ij} for the ij -th sample using the median count ratio to a pseudoreference (Anders and Huber 2010), the trimmed mean of M-values (TMM) (Robinson and Oshlack 2010), or the 75th percentile of the non-zero counts (Bullard *et al.* 2010).

In the simplest setting where there are only two groups, the goal of differential expression analysis is to test, for each gene, whether the mean expressions are identical in both groups. The model above can be simplified as,

$$\begin{aligned} Y_{gij} | \lambda_{gij} &\sim \text{Poisson}(S_{ij} \lambda_{gij}), \\ \lambda_{g1j} | \alpha_g, \beta_g &\sim \text{Gamma}(\alpha_g, \beta_g), \text{ and} \\ \lambda_{g2j} | \alpha_g, \beta_g, \rho_g &\sim \text{Gamma}(\alpha_g, \beta_g \rho_g). \end{aligned} \quad (2.4)$$

In this model, ρ_g is the fold change between the two treatment means. In our paper, we consider the case when there are two groups. Nevertheless, the method can be easily generalized to accommodate more than two groups.

2.2 Hypothesis Testing

For each $g = 1, \dots, G$, we are interested in testing the following hypothesis regarding the fold change parameter ρ_g :

$$H_0^g: \rho_g \in \Delta_0 \text{ vs. } H_1^g: \rho_g \in \Delta_1, \quad (2.5)$$

where Δ_0 and Δ_1 correspond to the null and alternative sets of values for ρ_g , respectively, and are assumed to be a partition of the positive real line R^+ ($\Delta_0 \cup \Delta_1 = R^+$, $\Delta_0 \cap \Delta_1 = \emptyset$). The null space Δ_0 can be defined in different ways depending on the biological questions of interest. For example, we set $\Delta_0 = \{1\}$ for H_0^g if we are interested in knowing whether the mean expression levels in the two treatments are equal. If we are interested in whether the mean expression is lower in the second treatment than in the first, we set $\Delta_0 = (0, 1]$.

2.3 Prior Specification

Because our main interest is to test the hypotheses in (2.5) for each gene, it is very important to specify an appropriate prior distribution for the fold change parameter ρ_g . The empirical distribution of the fold change varies from study to study and can be of very irregular shapes. As mentioned in Section 1, Bayesian nonparametric modeling is an alternative method for density estimation and is often used to avoid critical dependence on parametric assumptions. The most commonly used Bayesian nonparametric method is the Dirichlet process (DP). The definition of the Dirichlet process is as follows. A random probability distribution F is generated by a Dirichlet process if for any partition $A_1, \dots, A_k$ of some sample space Ω , the vector of random probabilities $(F(A_1), \dots, F(A_k))$ follows a Dirichlet distribution $D(M \cdot F_0(A_1), \dots, M \cdot F_0(A_k))$, where F_0 is a base measure and M is a total mass parameter. We denote this by $F \sim DP(M, F_0)$. For any set B in the sample space Ω , $E\{F(B)\} = F_0(B)$ and $var\{F(B)\} = F_0(B)(1 - F_0(B))/(M + 1)$. So the base measure defines the expectation and M defines the variance when the base measure is set. The properties of the Dirichlet process can be found in Ferguson (1973) and Antoniak (1974). The realization of a Dirichlet process $F \sim DP(M, F_0)$ can also be represented as

$$F(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{\{\mu_h\}}(\cdot), \quad (2.6)$$

with stick breaking priors

$$\mu_h \stackrel{iid}{\sim} F_0, \quad w_h = U_h \prod_{j < h} (1 - U_j) \text{ with } U_h \stackrel{iid}{\sim} \text{Beta}(1, M),$$

where $\delta_{\{x\}}(\cdot)$ denotes a point mass at x , the location parameters μ_h in the point masses are samples from F_0 , and the random weights w_h are generated as rescaled independent beta draws. In words, a realization of a DP can be represented as an infinite mixture of point masses, so a DP is almost surely discrete.

To provide maximal flexibility, we take a Bayesian nonparametric approach to model the uncertainty on the distribution of ρ_g . More specifically, we assume the fold change parameters $(\rho_1, \dots, \rho_G)$ follow a distribution that is drawn from a DP. For our Poisson-Gamma mixture model, using the Gamma distribution as the base distribution ensures the conjugacy of the base distribution with respect to the likelihood for the λ_{g2j} parameters. However, to readily accommodate sharp null hypotheses that we might be interested in, we use a mixture of a point-mass and a Gamma distribution as the base distribution in the DP prior for the distribution of the fold change parameters. In our paper, the sharp null hypothesis we are most interested in is $H_0^g: \rho_g=1$. Therefore, for each $g=1, \dots, G$, the DP prior can be expressed as

$$\begin{aligned} \rho_g | F &\stackrel{iid}{\sim} F, \\ F &\sim DP(M, F_0), \\ F_0 &\sim p_0 \delta_{\{1\}} + (1-p_0) \text{Gamma}(\alpha_0, \beta_0). \end{aligned} \quad (2.7)$$

As a default choice that is widely-used in applications (Kim *et al.* 2009; Green and Richardson 2001; Kalli *et al.* 2011), we let $M=1$. The parameter p_0 can be interpreted as the chance of being differentially expressed. We set $p_0=0.5$ in this paper to give no prior preference to either differential expression or non-differential expression. We discuss some simulation results on the robustness to changes of M and p_0 in Section 6.

We use a Gamma distribution as the prior distribution for β_g again because of its conjugacy. There is no conjugate prior for α_g , however. To reduce the complexity of computation of the posterior distribution, we use an exponential distribution as the prior for α_g .

$$\alpha_g \sim \text{Exp}(r), \quad (2.8)$$

$$\beta_g \sim \text{Gamma}(a_0, b_0), \quad (2.9)$$

where r , a_0 , b_0 and α_0 , β_0 are hyperparameters. We set $r=0.01$, $a_0=0.1$, $b_0=0.1$, $\alpha_0=0.1$, $\beta_0=0.1$ so that the priors are not informative and the inference for α_g and β_g mainly relies on the observed data. For computational convenience, we set the priors for α_g 's, β_g 's and ρ_g 's to be independent.

3. POSTERIOR INFERENCE

3.1 Markov Chain Monte Carlo Simulation

After specifying the prior distributions, data will be used to update these prior distributions to obtain the posterior distributions for inference. Markov chain Monte Carlo (MCMC) methods (Tierney 1994) that are used to obtain posterior samples allow Bayesian inference for highly complex models in which realistic distributional assumptions can be made. We use an MCMC-based full Bayesian approach, which has the advantage of providing joint posterior distributions. Gibbs sampling is the most common tool used to perform MCMC simulations for Bayesian hierarchical models. The Gibbs algorithm was introduced to machine learning by Geman and Geman (1984), and its applicability to statistical modeling was demonstrated by Gelfand and Smith (1990). The application of Gibbs sampling is quite convenient and straightforward for fully conditional distributions especially for fully conjugate Bayesian models. The algorithm proceeds by alternately sampling from each of the full conditional distributions, which are essential for obtaining the posterior samples using the Gibbs sampler.

For Dirichlet process mixture models, typical MCMC schemes are based on integrating out the Dirichlet process via convenient Polya urn representations within a Gibbs sampler to obtain posterior samples (Blackwell and MacQueen 1973; Escobar 1994). Based on this approach, we can use the generalized Polya urn scheme to obtain posterior samples for ρ_g by integrating unknown F in (2.7) out. In this way, the prior of ρ_g 's can be expressed as:

$$\begin{aligned}\rho_g | \alpha_0, \beta_0 &\sim PU(M, F_0), \\ F_0 &\sim p_0 \delta_{\{1\}} + (1-p_0) \text{Gamma}(\alpha_0, \beta_0),\end{aligned}$$

where PU is short for the Polya urn representation. The successive conditional distribution of PU can be written as below,

$$\rho_g | \boldsymbol{\rho}_{-g} \sim \frac{1}{G-1+M} \sum_{k \neq g} \delta_{\{\rho_k\}} + \frac{M}{G-1+M} F_0,$$

where $\boldsymbol{\rho}_{-g}$ is the vector of $(\rho_1, \dots, \rho_G)^T$ after deleting ρ_g . However, the algorithm based on PU may not be very efficient since ρ_g 's change only when they are reallocated to new components and convergence to the posterior distribution may be very slow (Neal 2000).

In order to ease the computational burden and improve the efficiency of the MCMC algorithm, a second Gibbs sampling method has been proposed by Neal (2000). In this algorithm, configuration indicators are defined and used. Let K denote the number of distinct values or clusters in the vector $\rho_1, \dots, \rho_G$ and assume the distinct values are $\rho_1^*, \dots, \rho_K^*$ respectively. Let $\xi = (\xi_1, \dots, \xi_G)^T$ be the configuration indicators defined by

$$\xi_g = k \text{ if and only if } \rho_g = \rho_k^* = \rho_{\xi_g}^*. \quad (3.1)$$

Therefore, the prior model for ρ_g 's is reparameterized with ρ_k^* 's and ξ_g 's, and the priors for parameters ρ_k^* 's and ξ_g 's are

$$\begin{aligned} \rho_k^* &\stackrel{iid}{\sim} F_0, \\ F_0 &\sim p_0 \delta_{\{1\}} + (1-p_0) \text{Gamma}(\alpha_0, \beta_0), \\ (\xi_1, \dots, \xi_G) | M &\sim \text{CRP}(M), \end{aligned}$$

where priors for ρ_k^* 's and ξ_g 's are independent and CRP is short for Chinese Restaurant Process. CRP is a random distribution and the full conditional urn for ξ_g 's can be written as

$$\xi_g | \xi_l, M \sim \sum_{k=1}^{K^{(-g)}} \frac{n_k^{(-g)}}{G-1+M} \delta_{\{k\}} + \frac{M}{G-1+M} \delta_{\{K^{(-g)}+1\}}, \quad (3.2)$$

where $K^{(-g)}$ denotes the number of distinct values in the vector $(\rho_1, \dots, \rho_G)^T$ after deleting ρ_g , and $n_k^{(-g)}$ denotes the size of the k^{th} cluster after removing ρ_g .

The MCMC procedure using Gibbs algorithm is elucidated below after reparameterization. For notational convenience, we write “.” in the conditioning set to indicate the data and all other parameters except the parameter(s) before the conditioning bar. Given starting values for the parameters, our MCMC sampling scheme updates each of the following parameters, one at a time and iterates between the following steps. Details on how we generate the full conditionals for all parameters are available in the Appendix.

1. Obtain draws of λ_{gij} 's from their full conditional distributions,

$$\begin{aligned} p(\lambda_{g1j} | \cdot) &\sim \text{Gamma}(Y_{g1j} + \alpha_g, \beta_g + S_{1j}), \\ p(\lambda_{g2j} | \cdot) &\sim \text{Gamma}(Y_{g2j} + \alpha_g, \beta_g \rho_g + S_{2j}). \end{aligned}$$

2. Draw samples of β_g 's from their full conditional distributions,

$$p(\beta_g | \cdot) \sim \text{Gamma} \left(\alpha_g(n_1 + n_2) + a_0, \sum_{j=1}^{n_1} \lambda_{g1j} + \rho_g \sum_{j=2}^{n_2} \lambda_{g2j} + b_0 \right).$$

3. A closed-form full conditional for α_g 's is not available. Since the conditional posterior distributions for α_g 's is a log-concave function with respect to α_g 's, we draw samples of α_g 's using the adaptive rejection algorithm (Gilks 1992).

4. Obtain posterior samples for ρ_g 's by getting the Markov chain for $(\xi_1, \dots, \xi_G)$ and $(\rho_1^*, \dots, \rho_K^*)$ as follows.

- i. Update the configuration vector $(\xi_1, \dots, \xi_G)$.

- If $\xi = \xi_l$ for some $l \neq g$,

$$P(\xi_g = \xi | \xi_{-g}, \text{rest}) = c n_{\xi}^{-g} \prod_{j=1}^{n_2} \frac{\beta_g^{\alpha_g} (\rho_{\xi}^*)^{\alpha_g}}{\Gamma(\alpha_g)} \lambda_{g2j}^{\alpha_g-1} \exp(-\beta_g \rho_{\xi}^* \lambda_{g2j}). \quad (3.3)$$

- Otherwise,

$$\begin{aligned} & P(\xi_g \neq \xi_l \text{ for all } l \neq g | \xi_{-g}, \text{rest}) \\ &= c M (1-p_0) \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \frac{\beta_g^{n_2 \alpha_g}}{[\Gamma(\alpha_g)]^{n_2}} \prod_{j=1}^{n_2} \lambda_{g2j}^{\alpha_g-1} \frac{\Gamma(n_2 \alpha_g + \alpha_0)}{(\beta_0 + \beta_g \sum_{j=1}^{n_2} \lambda_{g2j})^{n_2 \alpha_g + \alpha_0}} \\ &+ c M p_0 \prod_{j=1}^{n_2} \left\{ \frac{\beta_g^{\alpha_g}}{\Gamma(\alpha_g)} \lambda_{g2j}^{\alpha_g-1} \exp(-\beta_g \lambda_{g2j}) \right\}. \end{aligned}$$

Here c is a common proportionality constant to ensure that the probabilities add up to 1. K is the number of distinct values of the configuration vector $(\xi_1, \dots, \xi_G)$. As a result, K is automatically updated.

- ii. Update $(\rho_1^*, \dots, \rho_K^*)$ using the mixture of point mass and Gamma distribution as follows.

$$p(\rho_k^* | \cdot) \propto p_0 \exp \left\{ - \left(\sum_{\{g: \xi_g = k\}} \sum_{j=1}^{n_2} \beta_g \lambda_{g2j} \right) \right\} \delta_{\{1\}} + c_0 \text{Gamma} \left(n_2 \sum_{\{g: \xi_g = k\}} \alpha_g + \alpha_0, \beta_0 + \sum_{\{g: \xi_g = k\}} \sum_{j=1}^{n_2} \beta_g \lambda_{g2j} \right)$$

where

$$c_0 = (1-p_0) \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \frac{\Gamma(n_2 \sum_{\{g: \xi_g = k\}} \alpha_g + \alpha_0)}{(\beta_0 + \sum_{\{g: \xi_g = k\}} \sum_{j=1}^{n_2} \beta_g \lambda_{g2j})^{n_2 \sum_{\{g: \xi_g = k\}} \alpha_g + \alpha_0}}.$$

3.2 Bayesian FDR for Multiple Hypothesis Testing

Multiple testing procedures are primarily about controlling the number of false significant results obtained while simultaneously testing a large number of hypotheses. Benjamini and Hochberg (1995) introduced the false discovery rate (FDR) as an error criterion in multiple testing, and the FDR is defined as the expected proportion of falsely rejected null hypotheses among all rejected null hypotheses. The Bayesian FDR is one way to estimate the FDR, and it has been proposed and discussed by several authors including Genovese and Wasserman

(2003) and Newton *et al.* (2004). The Bayesian FDR can be obtained by using posterior probabilities of the null hypothesis.

For each $g(1 \leq g \leq G)$, we denote the posterior probability that the g^{th} null hypothesis is true by $P(\rho_g \in \Delta_0 | Y_g)$. When the null hypothesis of interest is $H_0^g: \rho_g \in \{1\}$, $P(\rho_g \in \Delta_0 | Y_g)$ is in fact the posterior probability that g^{th} gene is not differentially expressed. $P(\rho_g \in \Delta_0 | Y_g)$ can be estimated by the posterior samples obtained from MCMC through v_g , the proportion of the posterior samples for ρ_g that is within the null set Δ_0 , i.e.,

$$v_g = \hat{P}(\rho_g \in \Delta_0 | Y_g) = \frac{1}{N} \sum_{m=1}^N I(\rho_g^m \in \Delta_0 | Y_g), \quad (3.4)$$

where N is the number of posterior samples. We reject H_0^g if the estimated posterior probability v_g is smaller than a threshold c . The choice of the optimal threshold c will be based upon controlling the estimated false discovery rate (FDR) below a target level γ , say, 0.05,

$$c = \sup \{c: \widehat{\text{FDR}}(c) < \gamma\}. \quad (3.5)$$

So the Bayesian FDR is calculated by

$$\widehat{\text{FDR}}(c) = \frac{\sum_{g=1}^G v_g D_g}{\sum_{g=1}^G D_g}, \quad (3.6)$$

where $D_g = I(v_g < c)$ is the decision on the g^{th} hypothesis testing and $\sum_{g=1}^G D_g$ is the total number of rejections.

4. SIMULATION STUDY

In this section, we examine the effectiveness of our proposed semi-parametric Bayesian (SBA) approach and compare its performance with the performance of other popularly applied tests for differential expression analysis of RNA-seq data. The RNA-seq data were simulated from negative binomial (NB) distributions based on which the methods of edgeR (Robinson and Oshlack 2010), DESeq (Anders and Huber 2010), and baySeq (Hardcastle and Kelly 2010) were derived. To mimic the distributions of real datasets, the parameters of the NB distributions used to simulate data were obtained from the estimates of a maize dataset (Li *et al.* 2010). Two sets of simulation studies (A and B) were conducted, and they differ by how the fold change parameters were generated. Within each set, we have a couple of simulation settings by changing the number of replicates per treatment group. For each simulation setting, 30 independent RNA-seq datasets were simulated with each dataset

containing 10,000 genes, 2 treatment groups, and n replicates for each treatment group, where n is either 3 or 6. Using simulated datasets, the performance of our proposed test and other tests are evaluated in terms of the effectiveness of ranking DE genes and FDR estimation. For our proposed method, we burn in the MCMC sampler for 3000 iterations, followed by another 5000 iterations to calculate the estimated posterior probabilities. To check the adequacy of our chains in simulations A and B discussed below, we chose 100 genes at random and obtained the efficient sample size (ESS) for ρ_g 's using the R package CODA (Plummer *et al.* 2006). The results indicated that our MCMC sampling was sufficient.

4.1 Simulation A

To obtain the parameter values of the NB distributions, we first estimated the means of one treatment group and the dispersion parameters for the maize dataset analyzed in Li *et al.* (2010). The “tagwise” option in edgeR package was used to estimate the dispersion parameter for each gene. Given the number of replicates per group (n), the procedure we used to simulate one dataset is as follows.

1. Randomly sample 10,000 genes out of all (25426) genes in the maize dataset without replacement. Each of the corresponding 10,000 pairs of mean and dispersion parameters will be used as the true mean expression level for the control group (μ_g) and the true dispersion parameter (ϕ_g), respectively, for simulated gene g .
2. The RNA-seq count data for the hypothetical control group are drawn from $NB(\mu_g, \phi_g)$ for simulated gene g .
3. Randomly select 5,000 genes out of the 10,000 to be equivalently expressed (EE). And the RNA-seq count data for the hypothetical treatment group are also drawn from $NB(\mu_g, \phi_g)$.
4. The remaining 5,000 genes are simulated to be DE genes. The fold-change (ρ_g) is set to be 4, 8, 0.25, or 0.125, so that we have 1250 genes for each ρ_g value. Then the RNA-seq count data for the hypothetical treatment group are drawn from $NB(\mu_g \rho_g, \phi_g)$.

4.2 Simulation B

The procedure to generate a simulated dataset in Simulation B is the same as that in Simulation A except the simulation process for the ρ_g 's in step 4. Instead of setting ρ_g to be one of the values 4, 8, 0.25, and 0.125, we simulated ρ_g for the 5000 DE genes from a two-component mixture of lognormal distributions, $\log(\rho_g) \sim 0.5Normal(\log(4), 1) + 0.5Normal(-\log(4), 1)$. The distribution of these simulated ρ_g 's values are shown in Figure 2. Note that our simulation method partially preserves the empirical dispersion-mean relationship in the simulated data for both simulation studies.

4.3 Simulation Results for Testing DE genes

To evaluate the test performance without the influence of different normalization methods, we used the same normalization factors for all the five methods. Specifically, we set all normalization factors to be 1.

The performances of different methods in terms of ranking DE genes are evaluated in two types of plots. First, the receiver operating characteristic (ROC) curve, i.e., the plot of the true positive rate versus false positive rate, was generated based on either the posterior probabilities or p-values for each method, and the area under curve (AUC) was also calculated. A better performing method would have a higher ROC curve and a larger value of AUC compared with a worse performing method. Second, the number of false discoveries was plotted against the number of top ranked genes selected as DE, and we call such plots false discovery (FD) plots. The ranking was obtained based on either the posterior probabilities or p-values for each method. For a given number of DE genes, the fewer FD on the list, the better the performance of the method.

ROC curves that plot the TPR versus the FPR are shown in Figure 3. These curves are results of averaging over 30 datasets. We plotted the curves over the FPR values between 0 and 0.1 because the range of small FPR values are of the most practical importance. In addition, we calculated the AUC values for the same range of FPR. The average values and standard errors of the AUC from the 30 simulated datasets are reported in the legends in Figure 3. The AUC values presented in all ROC plots are the percentages of 0.1, where 0.1 is the total area for the plotted range of FPR. The AUC value provides an overall measure of the ability to rank truly DE genes ahead of truly non-DE genes. Figure 3 shows that the SBA method generated the highest ROC curve and largest AUC value among all tests in Simulation A. For Simulation B, the SBA method and baySeq method were similar and they performed better than others when $n = 3$. When $n = 6$, the SBA method and the AMAP method performed better than others.

Figure 4 presents the false discovery plots. When sample size increases from 3 to 6, the number of false positives decreases for all methods, as expected. Comparing different methods, it is obvious that the list of DE genes given by SBA would provide fewer false positives if we declare the same number of genes as DE for all methods in Simulation A. The difference in terms of the number of false positives can be big between SBA and others. For Simulation B, the SBA method was always among the methods that provided the smallest number of false positives for both sample size settings. Overall, the proposed SBA method performed either the best or among the best in all simulation settings.

We also evaluated the estimation of FDR using the posterior probabilities as introduced in Section 3.2. Figure 5 shows that in most cases, our estimate can be conservative, especially in the range of FDR smaller than 0.05, a level often being controlled in practice. It has been reported that FDR may not be controlled for the currently popularly applied method such as edgeR and DESeq (Si and Liu 2013; Sonesson and Delorenzi 2013), generating more false positives in the results than desired. Compared with liberal FDR control, a conservative control generates a more reliable list of DE genes.

Based on all the results, we conclude that the proposed SBA method generated better rankings of genes in most simulation settings and provided conservative estimation for FDR, and hence provided reliable lists of DE genes at a desired level of FDR.

4.4 Simulation Results for Testing: $|\log FC| > \log 1.5$

This section shows the results for testing: $|\log FC| > \log 1.5$ for Simulation B. We applied our proposed method (SBA) and the AMAP directly to do this hypothesis testing problem. And the two step procedure described in Si and Liu (2013) was used when the methods edgeR, DESeq and baySeq were applied. More specifically, The first step was to select a list of DE genes by testing $\rho_g = 1$ while controlling FDR at a certain level. The second step was to select the genes that have large enough fold changes, such as $|\log FC| > \log 1.5$, among the list of DE genes identified in the first step. The comparison of ROC curves for different methods is shown in the upper panel of Figure 6. Our proposed method outperformed all other four methods. Based on the lower panel of Figure 6, we find that our estimate was conservative in the range of FDR smaller than 0.1.

5. MAIZE DATA ANALYSIS

In this section, we analyze a maize RNA-seq dataset published by Tausta *et al.* (2014). Bundle sheath and mesophyll are two cell types in maize leaves that have different functions in the C4 photosynthesis mechanism utilized by corn plants. Tausta *et al.* (2014) applied the laser microdissection technique to capture each type of cells from the tips of the maize leaves and conducted an RNA-seq experiment to identify genes that are DE between the two types of cells. Data from two independent biological replicates were obtained for each cell type. The dataset is available online at NCBI Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo) with accession number GSE 54272.

The goal of the analysis is to identify genes that are differentially expressed between bundle sheath and mesophyll cells. Because there were biological replications, we assumed NB models for the expression counts observed for each gene. We performed our proposed method and controlled Bayesian FDR as described in Section 2. Two other methods (BaySeq, AMAP) that utilize Bayesian FDR were applied to analyze the data as well. The numbers of detected DE genes at different FDR levels are shown in Figure 7. When the FDR was controlled at 0.1%, the AMAP test detected the most DE genes among all three testing methods, while when the FDR was controlled at 1%, our proposed method detected the most DE genes. Moreover, the majority of genes detected by other methods were also identified by our proposed method. For example, as shown in Figure 7, when the FDR was controlled at 1%, the SBA test detected 2363 of the 3760 genes detected by AMAP.

6. DISCUSSION

In this paper, we use the Poisson-Gamma hierarchical structure to model the over-dispersed read counts from RNA-seq experiments and control the FDR based on the posterior probabilities corresponding to the null hypotheses we are interested in. We performed two simulation studies and compared our proposed method with some of the commonly used

methods (e.g., edgeR, DESeq), and the simulation studies show that our method outperformed the other methods.

The dependence of results on the choices of hyperparameters may be a potential problem for Bayesian methods. In this paper, we used very weakly informative priors for α_g and β_g while the parameter M and p_0 were set for the DP prior. The value of 1 for M is a common choice (Kim *et al.* 2009), and the value p_0 denotes the chance of being non-differentially expressed. We performed simulation studies to check the sensitivity of test results to the choices of M and p_0 and found that the results remain mostly the same between the four analyses using different combinations of M and p_0 (M being 1 or 10, and p_0 being 0.5 or 0.8). In practice, we suggest large values for p_0 (e.g., 0.9 or 0.95) for experiments where not many genes are expected to be differentially expressed and relatively small choices for p_0 (e.g. 0.5) for experiments where different treatments might affect many genes. If we do not have such prior knowledge, a non-informative prior distribution may be put on p_0 , but this will introduce more computational complexity.

Most existing methods in the current literature for DE analysis focus on the sharp null hypothesis that the fold change between two groups is 1. Our method can be used to test other sharp null hypotheses, for example, to check whether the treatment mean for the first condition is twice as large as the treatment mean for the second condition. Furthermore, our method provides the flexibility to deal with other kinds of hypothesis testing, say, testing whether the fold change is within a certain interval or not.

In our proposed method, we set the first treatment condition as the reference condition (i.e., baseline) and impose Dirichlet process as the prior for the distribution of fold changes between two treatment conditions for each gene. In a typical study design involving only one treatment and one control group, the control group can be selected as reference so that the estimated fold changes can be interpreted as effects of applied treatments. To extend our proposed Bayesian nonparametric method to more complex experimental designs, say, to those involving more than two groups, we could use Dirichlet process as the prior for the distribution of fold change ratios between each of the treatment groups and the baseline group. In case there is no obvious choice of a reference group based on the experimental design, the analysis results may vary depending on which group is used as the baseline. Based on our simulation results on two analyses that swapped the treatment and control groups, we found that the declared differential expression status remained the same for more than 90% of genes. An alternative method to deal with this potential problem is to use a symmetric base distribution for the Dirichlet process such as the log-normal base. In this case, it is difficult to use Gibbs sampling to generate MCMC posterior samples due to lack of conjugacy. Neal (2000) introduced some methods to obtain posterior samples for such non-conjugacy situations. However, the resulting approach becomes much more computationally complicated than using our proposed method.

Acknowledgments

This research was partially supported by National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health and the joint National Science Foundation/NIGMS Mathematical Biology Program under award

number R01GM109458 and by the National Science Foundation Plant Genome Research Program under award number IOS-1127017.

References

- Anders S, Huber W. Differential Expression Analysis for Sequence Count Data. *Genome Biology*. 2010; 11:R106. [PubMed: 20979621]
- Antoniak C. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*. 1974; 2(6):1152–1174.
- Blackwell D, MacQueen BJ. Ferguson Distributions Via Polya Urn Schemes. *The Annals of Statistics*. 1973; 1(2):353–355.
- Bullard JH, Purdom EA, Hansen KD, Dudoit S. Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments. *BMC Bioinformatics*. 2010; 11:94. [PubMed: 20167110]
- Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B*. 1995; 57:289–300.
- Do KA, Muller P, Tang F. A Bayesian Mixture Model For Differential Gene. *Applied Statistics*. 2005; 54:627–644.
- Escobar MD. Estimating Normal Means With a Dirichlet Process Prior. *Journal of the American Statistical Association*. 1994; 89:268–277.
- Ferguson TS. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*. 1973; 1:209–230.
- Gelfand AE, Smith FM. Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*. 1990; 85(410):398–409.
- Geman S, Geman D. Stochastic Relaxation, Gibbs distributions, and the Bayesian Restoration of Images. *Pattern Analysis and Machine Intelligence, IEEE Transactions*. 1984; 6:721–41.
- Genovese C, Wasserman L. Bayesian and Frequentist Multiple Testing. *Bayesian Statistics*. 2003; 7:145–161.
- Gilks WR. Adaptive Rejection Sampling for Gibbs Sampling. *Applied Statistics*. 1992; 41:337–348.
- Green PJ, Richardson S. Modelling Heterogeneity With and Without the Dirichlet Process. *Scandinavian Journal of Statistics*. 2001; 28:355–375.
- Hardcastle TJ, Kelly KA. baySeq: Empirical Bayesian Methods for Identifying Differential Expression in Sequence Count Data. *BMC Bioinformatics*. 2010; 11(422)
- Kalli M, Griffin J, Walker S. Slice Sampling Mixture Models. *Statistics and Computing*. 2011; 1:93–105.
- Kvam VM, Liu P, Si Y. A Comparison of Statistical Methods for Detecting Differentially Expressed Genes from RNA-Seq Data. *American Journal of Botany*. 2012; 99(2):248–256. [PubMed: 22268221]
- Kim S, Dahl DB, Vannucci M. Spiked Dirichlet Process Prior for Bayesian Multiple Hypothesis Testing in Random Effects Models. *Bayesian Analysis*. 2009; 4:707–732. [PubMed: 23950766]
- Law CW, Chen Y, Shi W, Smyth GK. Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-seq Read Counts. *Genome Biology*. 2014; 15:R29. [PubMed: 24485249]
- Li P, Ponnala L, Gandotra N, Wang L, Si Y, Austa SL, Kebrom TH, et al. The Developmental Dynamics of the Maize Leaf Transcriptome. *Nature Genetics*. 2010; 42:1060–1067. [PubMed: 21037569]
- Muller P, Fernando AQ. Nonparametric Bayesian Data Analysis. *Statistical Science*. 2004; 19(1):95–110.
- 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]
- Neal RM. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*. 2000; 9(2):249–265.

- Plummer M, Best N, Cowles K, Vines K. CODA: Convergence Diagnostics and Output Analysis for MCMC. *R News*. 2006; 6:7–11.
- Reeb PD, Steibel JP. Evaluating Statistical Analysis Models for RNA Sequencing Experiments. *Front Genet*. 2013; 4:178. [PubMed: 24062766]
- Robinson MD, Oshlack A. A Scaling Normalization Method for Differential Expression Analysis of RNA-seq Data. *Genome Biology*. 2010; 11:R25. [PubMed: 20196867]
- Robinson MD, Smyth GK. Moderated Statistical Tests for Assessing Difference in Tag Abundance. *Bioinformatics*. 2007; 23:2881–2887. [PubMed: 17881408]
- Robinson MD, Smyth GK. Small-sample Estimation of Negative Binomial Dispersion, with Applications to SAGE Data. *Biostatistics*. 2008; 9:321–332. [PubMed: 17728317]
- Si Y, Liu P. An Optimal Test with Maximum Average Power While Controlling FDR with Application to RNA-seq Data. *Biometrics*. 2013; 69:594–605. [PubMed: 23889143]
- Smith AM, Roberts GO. Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. *Journal Royal Statistical Society B*. 1993; 55:3–23.
- Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*. 2013; 14:91.
- Tausta SL, Li P, Si Y, Gandotra N, Liu P, Sun Q, Brutnell TP, Nelson T. Developmental Dynamics of Kranz Cell Transcriptional Specificity in Maize Leaf Reveals Early onset of C4-related Processes. *Journal of Experimental Botany*. 2014; 65(13):3543–55. [PubMed: 24790109]
- Tierney L. Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics*. 1994; 22(4): 1701–1728.
- Wu Y, Ghosal S. The L_1 - consistency of Dirichlet Mixtures in Multivariate Bayesian Density Estimation on Bayes Procedures. *Journal of Mutivariate Analysis*. 2010; 101:2411–2419.

APPENDIX DETAILS ABOUT FULL CONDITIONAL DISTRIBUTIONS

Details for how to generate full conditional distributions for the Gibbs sampling.

1. Obtain full conditionals for λ_{gij} 's:

$$\begin{aligned} p(\lambda_{g1j}|\cdot) &\propto p(Y_{g1j}|\lambda_{g1j})p(\lambda_{g1j}|\alpha_g, \beta_g) \\ &\propto \exp(-S_{1j}\lambda_{g1j}) \frac{\lambda_{g1j}^{Y_{g1j}}}{Y_{g1j}!} \lambda_{g1j}^{\alpha_g-1} \exp(-\beta_g \lambda_{g1j}) \\ &\propto \lambda_{g1j}^{Y_{g1j}+\alpha_g-1} \exp(-S_{1j}\lambda_{g1j}-\beta_g \lambda_{g1j}), \end{aligned}$$

so the conditional distribution for the λ_{gij} parameter is,

$$p(\lambda_{g1j}|\cdot) \sim \text{Gamma}(Y_{g1j}+\alpha_g, \beta_g+S_{1j}). \quad (\text{A.1})$$

Similarly, the conditional distribution for the λ_{g2j} parameter is

$$p(\lambda_{g2j}|\cdot) \sim \text{Gamma}(Y_{g2j}+\alpha_g, \beta_g\rho_g+S_{2j}). \quad (\text{A.2})$$

2. Obtain full joint conditional for β_g 's:

$$\begin{aligned} p(\beta_g|\cdot) &\propto \pi(\beta_g) \prod_{j=1}^{n_1} p(\lambda_{g1j}|\alpha_g, \beta_g) \prod_{j=1}^{n_2} p(\lambda_{g2j}|\alpha_g, \beta_g, \rho_g) \\ &\propto \beta_g^{\alpha_g(n_1+n_2)+\alpha_0-1} \exp(-\beta_g \sum \lambda_{g1j} - \beta_g \rho_g \sum \lambda_{g2j} - \beta_g b_0), \end{aligned}$$

so the full conditional distribution for the parameter β_g is,

$$p(\beta_g|\cdot) \sim \text{Gamma} \left(\alpha_g(n_1+n_2)+a_0, \sum_{j=1}^{n_1} \lambda_{g1j} + \rho_g \sum_{j=2}^{n_2} \lambda_{g2j} + b_0 \right). \quad (\text{A.3})$$

3. Obtain full joint conditional for α_g 's:

$$\begin{aligned} p(\alpha_g|\cdot) &\propto \pi(\alpha_g) \prod_{j=1}^{n_1} p(\lambda_{g1j}|\alpha_g, \beta_g) \prod_{j=1}^{n_2} p(\lambda_{g2j}|\alpha_g, \beta_g, \rho_g) \\ &\propto \pi(\alpha_g) \frac{\prod_{i=1}^2 \prod_{j=1}^{n_i} \lambda_{gij}^{\alpha_g-1}}{\Gamma(\alpha_g)^{n_1+n_2} \beta_g^{-\alpha_g(n_1+n_2)} \rho_g^{-n_2\alpha_g}}, \\ &\propto \frac{\prod_{i=1}^2 \prod_{j=1}^{n_i} \lambda_{gij}^{\alpha_g-1}}{\Gamma(\alpha_g)^{n_1+n_2} \beta_g^{-\alpha_g(n_1+n_2)} \rho_g^{-n_2\alpha_g}} \exp(-r\alpha_g). \end{aligned}$$

Since a closed-form full conditional distribution for α_g is not available, we update α_g using an adaptive rejection sampling method, which is one convenient 'black box' random variate generation technique. We can prove that $p(\alpha_g|\cdot)$ is a log-concave function since the Gamma function is log-convex. So we can use adaptive rejection sampling method (Gilks 1992) to obtain the posterior samples. The adaptive rejection sampling is a generalization of standard non-adaptive rejection sampling method, which needs only one condition that the conditional posterior distribution is a log-concave function with respect to the parameter. So it is good to use this sampling strategy to deal with non-conjugacy that might cause complicated computation in applications of Gibbs sampling.

$$\log p(\alpha_g|\cdot) = \alpha_g \sum_{i=1}^2 \sum_{j=1}^{n_i} \log(\lambda_{gij}) - r\alpha_g - (n_1+n_2) \log \Gamma(\alpha_g) + (n_1+n_2)(\log \beta_g) \alpha_g + n_2 \alpha_g \log \rho_g + \text{const.}$$

Thus

$$\frac{\partial \log p(\alpha_g|\cdot)}{\partial \alpha_g} = \sum_{i=1}^2 \sum_{j=1}^{n_i} \log(\lambda_{gij}) + (n_1+n_2)(\log \beta_g) + n_2 \log \rho_g - r - (n_1+n_2) \frac{d \log \Gamma(\alpha_g)}{d \alpha_g}.$$

The second derivative of $\log p(\alpha_g|\cdot)$ is

$$\frac{\partial^2 \log p(\alpha_g|\cdot)}{\partial^2 \alpha_g} = -(n_1+n_2) \frac{d^2 \log \Gamma(\alpha_g)}{d^2 \alpha_g}.$$

And we could derive the following expansion series for the 2nd derivative of the Gamma function,

$$\frac{d^2 \log \Gamma(x)}{d^2 x} = \sum_{k=0}^{\infty} \frac{1}{(x+k)^2}.$$

The equation above implies that the second derivative of $\log \Gamma(x)$ is positive for all real and positive x . That is, the Gamma function is log-convex for positive values of x .

4. Update ρ_g 's by obtaining posterior samples for configuration vector $(\xi_1, \dots, \xi_G)$ and $(\rho_1^*, \dots, \rho_K^*)$.

- i. Update configuration vector $(\xi_1, \dots, \xi_G)$.

- If $\xi = \xi_l$ for some $l \neq g$:

$$\begin{aligned} P(\xi_g = \xi | \xi_{-g}, \text{rest}) &= c \cdot n_{\xi}^{(-g)} \prod_{j=1}^{n_2} p(\lambda_{g2j} | \alpha_g, \beta_g, \rho_{\xi}^*) \\ &= c \cdot n_{\xi}^{(-g)} \prod_{j=1}^{n_2} \frac{\beta_g^{\alpha_g} (\rho_{\xi}^*)^{\alpha_g}}{\Gamma(\alpha_g)} \lambda_{g2j}^{\alpha_g-1} \exp(-\beta_g \rho_{\xi}^* \lambda_{g2j}). \end{aligned}$$

- Otherwise,

$$\begin{aligned} P(\xi_g \neq \xi_l \text{ for all } l \neq g | \xi_{-g}, \text{rest}) &= cM \int \prod_{j=1}^{n_2} p(\lambda_{g2j} | \alpha_g, \beta_g, \rho) F_0(\rho) d\rho \\ &= cM(1 \\ &\quad - p_0) \int_0^{\infty} \prod_{j=1}^{n_2} \left\{ \frac{\beta_g^{\alpha_g} \rho^{\alpha_g}}{\Gamma(\alpha_g)} \lambda_{g2j}^{\alpha_g-1} \exp(-\beta_g \rho \lambda_{g2j}) \right\} \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \rho^{\alpha_0-1} \exp(-\beta_0 \rho) d\rho \\ &\quad + cM p_0 \int_0^{\infty} \prod_{j=1}^{n_2} \left\{ \frac{\beta_g^{\alpha_g} \rho^{\alpha_g}}{\Gamma(\alpha_g)} \lambda_{g2j}^{\alpha_g-1} \exp(-\beta_g \rho \lambda_{g2j}) \right\} \delta_{\{1\}} d\rho \\ &= cM(1 \\ &\quad - p_0) \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \frac{\beta_g^{n_2 \alpha_g}}{[\Gamma(\alpha_g)]^{n_2}} \prod_{j=1}^{n_2} \lambda_{g2j}^{\alpha_g-1} \frac{\Gamma(n_2 \alpha_g + \alpha_0)}{(\beta_0 + \beta_g \sum_{j=1}^{n_2} \lambda_{g2j})^{n_2 \alpha_g + \alpha_0}} \\ &\quad + cM p_0 \prod_{j=1}^{n_2} \left\{ \frac{\beta_g^{\alpha_g}}{\Gamma(\alpha_g)} \lambda_{g2j}^{\alpha_g-1} \exp(-\beta_g \lambda_{g2j}) \right\}. \end{aligned}$$

Here c is the common proportionality constant to ensure that the probabilities add up to 1. K is the number of distinct values of the configuration vector $(\xi_1, \dots, \xi_G)$, as a result K is automatically updated.

- ii. Update $(\rho_1^*, \dots, \rho_K^*)$.

$$\begin{aligned}
p(\rho_k^*|\cdot) &\propto \Pi_{\{g:\xi_g=k\}} \Pi_{j=1}^{n_2} p(\lambda_{g2j}|\alpha_g, \beta_g, \rho_g) \times F_0(\rho_k^*) \\
&\propto \Pi_{\{g:\xi_g=k\}} \Pi_{j=1}^{n_2} p(\lambda_{g2j}|\alpha_g, \beta_g, \rho_k^*) \times F_0(\rho_k^*) \\
&\propto \Pi_{\{g:\xi_g=k\}} \Pi_{j=1}^{n_2} \frac{\beta_g^{\alpha_g} (\rho_k^*)^{\alpha_g}}{\Gamma(\alpha_g)} \lambda_{g2j}^{\alpha_g-1} \exp(-\beta_g \rho_k^* \lambda_{g2j}) \times \{p_0 \mathbf{1}_{\{\rho_k^*=1\}} + (1-p_0) \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} (\rho_k^*)^{\alpha_0-1} \exp(-\beta_0 \rho_k^*)\} \\
&\propto p_0 \exp\left\{-\left(\sum_{\{g:\xi_g=k\}} \sum_{j=1}^{n_2} \beta_g \lambda_{g2j}\right)\right\} \delta_{\{1\}} + (1-p_0) (\rho_k^*)^{\alpha_0-1} + \sum_{\{g:\xi_g=k\}} \sum_{j=1}^{n_2} \alpha_g \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \exp\left\{-(\beta_0 + \sum_{\{g:\xi_g=k\}} \sum_{j=1}^{n_2} \beta_g \lambda_{g2j})\right\} \\
&\propto p_0 \exp\left\{-\left(\sum_{\{g:\xi_g=k\}} \sum_{j=1}^{n_2} \beta_g \lambda_{g2j}\right)\right\} \delta_{\{1\}} + c_0 \text{Gamma}(n_2 \sum_{\{g:\xi_g=k\}} \alpha_g + \alpha_0, \beta_0 + \sum_{\{g:\xi_g=k\}} \sum_{j=1}^{n_2} \beta_g \lambda_{g2j}),
\end{aligned}$$

where

$$c_0 = (1-p_0) \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \frac{\Gamma(n_2 \sum_{\{g:\xi_g=k\}} \alpha_g + \alpha_0)}{(\beta_0 + \sum_{\{g:\xi_g=k\}} \sum_{j=1}^{n_2} \beta_g \lambda_{g2j})^{n_2 \sum_{\{g:\xi_g=k\}} \alpha_g + \alpha_0}}.$$

When $\rho_g = \rho_k^*$ has multiplicity just 1 (i.e., if

$n_k^{(-g)} = n_{\xi_g}^{(-g)} = 0$), for simplicity, the posterior conditional distribution for ρ_k^* 's is

$$p(\rho_k^*|\cdot) \sim p_0 \exp\left\{-\left(\sum_{j=1}^{n_2} \beta_g \lambda_{g2j}\right)\right\} \delta_{\{1\}} + (1-p_0) c_1 \cdot \text{Gamma}(n_2 \alpha_g + \alpha_0, \beta_0 + \sum_{j=1}^{n_2} \beta_g \lambda_{g2j}),$$

$$\text{where } c_1 = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \frac{\Gamma(n_2 \alpha_g + \alpha_0)}{(\beta_0 + \sum_{j=1}^{n_2} \beta_g \lambda_{g2j})^{n_2 \alpha_g + \alpha_0}}.$$

Figure 1. Histogram of the empirical estimated log fold changes ($\log \hat{\rho}_g$) from one maize data set (Tausta *et al.* 2014).

Figure 2.

Histogram for the logarithm of the simulated fold change ρ_g 's. The dashed line denotes the estimated density from the simulated log fold changes and the solid line denotes the density plot for the mixtures of the lognormal distributions with mean $-\log(4)$ and standard deviation 1 and the lognormal distributions with mean $\log(4)$ and standard deviation 1.

Figure 3.

ROC curves resulting from Simulations A and B. For each level of FPR, the TPRs were averaged across the 30 simulated datasets. The percentage annotated for each method is the average AUC, represented as the percentage of the total area 0.1 in the range of $FPR < 0.1$, and the percentage in each set of parentheses is the standard error of the estimated AUC.

Figure 4.

False discovery curves, depicting the average number of false positives encountered among the top-ranked genes by the five evaluated methods. Genes were ranked according to their P-value or posterior probabilities, and the number of false positives was averaged over 30 simulated datasets.

Figure 5.

Plot of the actual FDR versus the nominal level of FDR. Using our proposed SBA method, the proportion of false discoveries among the declared DE genes was calculated for each dataset when we controlled FDR at nominal levels, and the actual FDR was calculated by averaging such proportions over 30 simulated datasets at each nominal FDR level.

(a) ROC curves for testing $FC > 1.5$ (b) FDR estimation for testing $FC > 1.5$ **Figure 6.**

Results for testing $FC > 1.5$ from Simulation B. (a) ROC curves resulting from Simulation B. For each level of FPR, the TPRs were averaged across the 30 simulated datasets. The percentage annotated for each method is the average AUC, represented as the percentage of the total area 0.1 in the range of $FPR < 0.1$, and the percentage in each set of parentheses is the standard error of the estimated AUC. (b) Plot of the true FDR versus the nominal level of FDR.

Figure 7.

The Venn diagram on the left shows the results of detected differentially expressed genes from the Bayseq method, the AMAP test and our proposed method when the FDR was controlled at 0.1%, the Venn diagram on the right shows the corresponding results when FDR was controlled at 1%.