A HIERARCHICAL BAYESIAN APPROACH FOR DETECTING DIFFERENTIAL GENE EXPRESSION IN UNREPLICATED RNA-SEQUENCING DATA

Sanvesh Srivastava
R. W. Doerge

Follow this and additional works at: https://newprairiepress.org/agstatconference

Part of the Agriculture Commons, and the Applied Statistics Commons

This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License.

Recommended Citation

This is brought to you for free and open access by the Conferences at New Prairie Press. It has been accepted for inclusion in Conference on Applied Statistics in Agriculture by an authorized administrator of New Prairie Press. For more information, please contact cads@k-state.edu.
A Hierarchical Bayesian Approach for Detecting Differential Gene Expression in Unreplicated RNA-Sequencing Data

Sanvesh Srivastava and R.W. Doerge*

Department of Statistics, Purdue University, West Lafayette, IN 47907

*Corresponding Author:

R.W. Doerge

Department of Statistics

Purdue University

250 N. University St.

West Lafayette, IN 47907

e-mail: doerge@purdue.edu

phone: 765-494-6030

fax: 765-494-0558
Abstract

Next-generation sequencing technologies have emerged as a promising technology in a variety of fields, including genomics, epigenomics, and transcriptomics. These technologies play an important role in understanding cell organization and functionality. Unlike data from earlier technologies (e.g., microarrays), data from next-generation sequencing technologies are highly replicable with little technical variation. One application of next-generation sequencing technologies is RNA-Sequencing (RNA-Seq). It is used for detecting differential gene expression between different biological conditions. While statistical methods for detecting differential expression in RNA-Seq data exist, one serious limitation to these methods is the absence of biological replication. At present, the high cost of next-generation sequencing technologies imposes a serious restriction on the number of biological replicates. We present a simple parametric hierarchical Bayesian model for detecting differential expression in data from unreplicated RNA-Seq experiments. The model extends naturally to multiple treatment groups and any number of biological replicates. We illustrate the application of this model through simulation studies and compare our approach to existing methods for detecting differential expression such as, Fisher's Exact Test.

**Keywords:** Hierarchical Bayesian modeling, microarrays, next-generation sequencing, Poisson distribution, differential gene expression, generalized linear models, Gibbs sampling.
1. Introduction

Next-generation sequencing technologies have emerged as a promising approach for exploring the cell organization and functionality, and are used in a variety of fields, including genomics, epigenomics, and transcriptomics (Hayden, 2009, Metzker, 2009, Ng et al., 2010, and Roach et al., 2010). Unlike data from earlier technologies such as microarrays, data from next-generation sequencing technologies are highly replicable with little technical variation (Marioni et al., 2008). Data from next-generation sequencing technologies are in the form of discrete gene counts that represent the relative amount of expression of each gene in the genome. When this technology is used to detect differential gene expression between different biological conditions, it is referred to as RNA-Sequencing (RNA-Seq). Similar to other high-throughput data, RNA-Seq data are high-dimensional data, and typically involve limit number of samples (that is, number of individuals analyzed) compared to the number of predictors (that is, genes); a problem known as “big p small n” or “curse of dimensionality”.

Research in high dimensional data first gained momentum with the analysis of microarray data. It has lead to significant advancements in the theory of multiple hypotheses testing (Efron et al., 2001), variable selection (Zou and Hastie, 2005), and the use of false discovery rates (FDR) for multiple testing problems (Benjamini and Hochberg, 1995, Storey, 2003). In order to model such data Efron (2010) recommends approaches such as empirical Bayesian methods that take advantage of information-borrowing across genes to compensate for limited availability of samples. Also, there are Bayesian approaches (Baldi and Long, 2001, Ibrahim et al., 2002) and
penalized-likelihood based approaches (Tibshirani et al., 2004, Ma and Huang, 2007) that take advantage of information-borrowing amongst genes. Many of these ideas have been applied to RNA-Seq data to determine differential gene expression with the central themes of calculating gene-wise test statistics, shrinking them towards a common value, and using FDR adjusted p-values for the modified test statistics to determine differentially expressed genes. Interestingly, RNA-Seq data pose two main non-trivial problems that do not arise when dealing with microarray data. First, due to the discrete nature of the data there are no equivalents of a t-test or an F-test (Casella and Berger, 2001); rather, the distribution of the test statistic is determined by the asymptotic likelihood distribution approximations (Anders and Huber, 2010, Robinson and Smyth, 2007, 2008). Second, due to overdispersion, small counts, and zero inflation which are very common in RNA-Seq data, the assumption of a Poisson distribution on gene counts may not be justified (Vêncio et al., 2004, Thygesen, 2006, Hardcastle and Kelly, 2010). Currently, RNA-Seq data represent a subsample of gene counts that are obtained from the original population of genes in the sample assessed by next-generation sequencing technologies. The total number of genes in the sample assessed by next-generation sequencing technologies is called the library size of the sample. The library size may vary depending on the sample. The effect of differences due to library size is discussed in Robinson and Oshlack (2010).

One of the important issues in RNA-Seq experiments is determining differentially expressed genes. Accurate modeling of gene abundance is crucial for determining differential gene expression. Gene abundance is defined as the population mean from which the gene count is sampled (i.e., the sample assessed by the next-generation sequencing technology). The gene
counts are modeled as a Poisson random variable, and are assumed to be independent of the size of the population (i.e., the total number of gene counts in the sample assessed by the next-generation sequencing technology). Presently, very little attention has been paid to identifying differentially expressed genes in unreplicated experiments mainly because of lack of reliable statistical inference in unreplicated experiments and reliable asymptotic theory. But many unreplicated experiments are conducted by biologists for the purpose of surveying an organism, for preliminary analysis, or because of the high cost of next-generation sequencing technologies. Here we present a simple parametric hierarchical Bayesian model for detecting differential gene expression in data from unreplicated RNA-Seq experiments. Our method borrows information across genes to compensate for the missing information about variation within a treatment group. The model determines the differential expression of each gene through their posterior probability distribution, and extends naturally to multiple treatment groups and any number of biological replicates. Simulation studies are employed to compare the results of our approach to currently used methods for detecting differential expression in unreplicated RNA-Seq data such as, Fisher's Exact Test (Agresti, 2002).

2. Hierarchical Bayesian Modeling Framework

We use a hierarchical Bayesian model (Gelman et al., 2003, Gelman and Hill, 2007) to determine differential gene expression in RNA-Seq data. The hierarchical Poisson model, shown later (equation 2.2 – 2.6), facilitates estimation of the posterior distributions of gene-wise differential expression from the observed data through Markov chain Monte Carlo (MCMC)
simulations. If 0 is not included in the 95% credible interval (CI) as determined from the posterior distribution of a gene, we conclude that the gene is differentially expressed with 95% probability.

Let \( \eta_g \) be the observed count of gene \( g \) in the sample with treatment \( t \), and let \( \theta_g \) denote the expected value of \( \eta_g \). The library size of a particular treatment group \( t \) is defined as the total gene count in the original population of genes in the sample assessed by the next-generation sequencing technology from which gene counts are obtained. This is denoted as \( n_j \) and it is not known apriori. The gene counts depend on the library size, since a large library size implies high gene counts. Our aim is to estimate the posterior distribution for gene abundance, \( \lambda_g \), which is independent of library size. From these posterior distributions we will then obtain the posterior distribution for gene-wise differential expression. Specifically, for a particular \( g \) and \( t \), we obtain \( \lambda_g \) from \( \theta_g \) by dividing it by \( n_j \). Equation 2.1 shows the relationship between \( \lambda_g \), \( n_j \), and \( \theta_g \).

\[
\eta_{gt} \theta = \text{ where } 1 \text{ and } 1, 2 \tag{2.1}
\]

The statistical model for detecting gene-wise differential expression is a two-level hierarchical Bayesian model

\[
n_{gt} | \theta_{gt} \sim \text{Poisson} \left( \right) \tag{2.2}
\]
In the first level (2.1) – (2.2), the model assumes that $\eta_g$, given $\mu_i$ and $\lambda_g$, are independent for different genes $g$ in a particular treatment group $t$, and follow Poisson distribution with mean parameters $\mu_i^\lambda_g$, respectively. It is assumed that $\mu_i^\lambda_g$ corresponds to the mean of $\eta_g$, $\theta_g$. The second level of the hierarchy models the library size for gene $g$ in treatment $t$ as a draw from Poisson distribution with mean $\mu_i$. Note that the mean parameter for abundance of gene $g$ in treatment $t$ is $\lambda_g$. Information borrowing and overdispersion are modeled in the second level (see: equation 15.6, Section 15.1, Gelman and Hill, 2007).

When modeling $\lambda_g$ (2.2), the random parameter $\rho_i$ promotes borrowing of information among all the genes in the treatment group $t$, and $\rho_i$ makes all the genes in treatment $t$ correlated. Hence, the posterior distribution for $\theta_g$ depends on all the gene counts in treatment $t$, including $\eta_g$ (see: pages 333-334, Section 9.3.3, Ntzoufras, 2009). This assumption implies the exchangeability of
gene counts within a treatment. Note that we could have borrowed information across the treatment groups by including a random parameter that depends on genes, but we prefer the exchangeability of gene counts within a treatment, rather than across treatments, since it is more informative. However, depending on the application, borrowing information between treatments may be relevant. Specifically, under the sparsity assumption only few genes respond to a biological treatment, borrowing information across treatment groups is reasonable and is likely to lead to more robust results. The other random term, $\delta_\gamma$ (2.2), models the overdispersion in gene counts. Since random error contributes towards overdispersion, it is not modeled separately.

In order to model the true library size for gene $g$ in treatment $t$, we model the sum of all gene counts, $\sum_{g=1}^{G} n_{g}^t$, as the baseline for the mean library size because the sample variant of library size (i.e., the sum of all gene counts), represents a small fraction of the library size. The actual library size for treatment $t$ is greater than $\sum_{g=1}^{G} n_{g}^t$. The error term, $\varepsilon_t$, models the treatment-wise variation in library size. Parameters $\beta_t$, $\varepsilon_t$, and $\delta_\gamma$ are assumed to follow a normal distribution with zero mean. A layer of weakly informative hyperpriors are imposed on the variances of these distribution to estimate the posterior distribution of $\theta_\gamma$. The hierarchical model is fit using the software package JAGS (Plummer, 2003) via the rjags package (Plummer, 2009) in R (R Development Core Team, 2011). We obtain samples from the posterior distributions of $\theta_\gamma$ through the Gibbs sampler (Gelfand and Smith, 1990, Ntzoufras, 2009) as implemented in JAGS.
The posterior distribution for the gene abundance, $\lambda_g$, is obtained by normalizing the samples from the posterior distribution of $\theta_g$ using the estimate of corresponding library size

$$
\hat{n}_t = \mu_{\|\theta\|} \sum_{g=1}^{G} \eta_{g_t}
$$

(2.6) 

where $s$ denotes the samples drawn from the posterior distribution of the parameters, $\lambda_g$, $\theta_g$, and $\mu_t$. The median, $\mu_{\|\theta\|}$, is obtained from the posterior draws of coefficients of the baseline measure for library size for treatment $t \left( \sum_{g=1}^{G} \eta_{g_t} \right)$. The total number of samples, $S$, is chosen large enough so that the MCMC chains mix well, and we obtain 95% CI for determining differential gene expression with reasonable coverage. The mixing of MCMC chains is tested in JAGS through the Gelman-Rubin diagnostic statistic and trace plots (Gelman and Rubin, 1992). Since the posterior distribution of $\lambda_g$ is based on the posterior distribution of $\theta_g$, it also borrows information from all the gene counts in treatment $t$ to estimate posterior distribution of the abundance of gene $g$ in treatment $t$. Even if there is limited availability of samples, the posterior distribution of $\lambda_g$ gains its strength from borrowing of information through the levels of the hierarchical model. Intuitively, this implies that an increase in number of genes and samples will result in an increase in the reliability of the estimate of the posterior distribution.
We assume that treatment 1 is the baseline category and obtain posterior distribution of differential gene expression of gene $g$ at MCMC sample $s$, $\Delta_{g^s}$, as follows:

\[
\Delta_{g^s} = \frac{w_2}{1-w_2} \lambda_{21}
\]

\[
w_2 = \frac{\left| \frac{n_{g^2}}{\|g^2\|} \right|^2}{\sum_{g_{12}} \left( \frac{n_{g_{12}}}{\|g_{12}\|} \right)^2 - \sum_{g_{21}} \left( \frac{n_{g_{21}}}{\|g_{21}\|} \right)^2}
\]

where $w_2$ is a measure of signal in gene counts from treatment 2. The signal is measured by the fraction of median of gene counts in treatment 2, $\frac{n_{g^2}}{\|g^2\|}$, with respect to the sum of medians across two treatment groups. The strength of the signal in gene counts from a sample is measured by the median of the gene counts in the sample. Large gene counts are assumed to have more signal compared to low gene counts. Therefore, we scale up the sample draws of $\lambda_{g2}$ depending on $w_2$. The measure, $\left| \frac{w_2}{1-w_2} \right|$, is the relative strength of signal in treatment 2 with respect to treatment 1. This scaling factor of $\lambda_{g2}$ is greater than 1 if there is more signal in treatment 2 with respect to treatment 1, and vice-versa. We obtain a symmetric 95% credible interval (CI$_{g}$) for the differential expression of gene $g$ from the posterior draws of $\Delta_{g}$.
\[
\text{CI}_{g} = \left[ \Delta^{0.025}, \Delta^{0.975} \right]_{g}, \quad 1, \ldots, G
\]

(2.8)

2.5% quantile of the posterior draws of \( \Delta^{0.025} \)
97.5% quantile of the posterior draws of \( \Delta^{0.975} \)

0 \notin \text{CI}_{g} \Rightarrow \text{Gene } g \text{ is differentially expressed.}

We can identify all the differentially expressed genes in the RNA-Seq data using equation 2.7. If 0 does not belong to \( \text{CI}_{g} \), then we conclude that gene \( g \) is differentially expressed with 95% probability. Otherwise, we conclude that the gene is not differentially expressed.

3. Simulations and Results

We rely on simulated data to illustrate an application of hierarchical Bayesian modeling (equation 2.2) for determining differential gene expression from unreplicated RNA-Seq data. We also compare the results of the hierarchical model with Fisher's Exact for detecting differential gene expression in unreplicated experiments.

3.1 Simulation Setting

Typically, the total number of genes involved in any transcriptome experiment involves at least thousands of genes. We assume that the total number of genes sampled is 1000, and that there
are two treatment levels. Treatment 1 is assumed to be the baseline case. The gene counts \( n_{gt} \) are generated from Poisson distribution with mean parameters \( \theta_g \).

\[
\text{(3.1)}\quad n_{gt} \sim \text{Poisson}(\theta_g)
\]

We simulate the mean parameters \( (\theta_g) \) of the gene counts \( (n_{gt}) \) from a gamma distribution dependent on library size by first generating \( z_{gt} \) from gamma distribution, and then scaling \( z_{gt} \) by the appropriate library size (equation 3.3). The shape and rate parameters \( (\alpha \text{ and } \beta) \) of the gamma prior are chosen to match the mean and variance of the simulation setting of Bioconductor package (Gentleman et al., 2004), edgeR (Section 12, Robinson et al., 2010). This is different from the modeling assumption in equation 2.2, which assumes \( \theta_g \) to be log-normal.

\[
\theta_g = \begin{cases} 
400 \text{ for } z_{g1} \leq 1000 & g \ldots \\
50 \text{ for } 1011000 & g \ldots \\
10050 \text{ for } z_{g2} \geq 1100 & g \text{ (differentially expressed genes)} 
\end{cases}
\]

\[
\text{(3.2)}
\]

\[
\text{for } g = 001 \text{ and } 1,2
\]

\[
\text{(3.3)}
\]

\[
\sum_{g \in \text{independent}} \text{Gamma}(10,100000)
\]

\[
\text{(3.4)}
\]

\[
z | \text{Gamma}(\alpha, \beta) \Rightarrow \frac{\beta^\alpha}{\Gamma(\alpha)} e^{\beta z}
\]
The library sizes ($4 \times 10^5$ and $5 \times 10^5$ for treatment 1 and 2, respectively in equation 3.2) are identical to the edgeR simulation setting. This makes the simulation closer to real life scenario where a large number of gene counts in RNA-Seq data are low; low counts are typically less than or equal to 5. We further assume that 10% of the genes are differentially expressed to make the simulation study close to reality. It is accomplished by making the first 10% of the genes in treatment 2 to have $\theta_{g2}$ higher than the $\theta_{g1}$ by 100 units (equation 3.2). The difference of 100 units is arbitrary; if the difference is increased, it is easier to estimate the differentially expressed genes accurately.

3.2 Results

We fit the hierarchical Bayesian model (equation 2.2 – 2.6) on the simulated RNA-Seq data using R and JAGS. We sampled 2000 draws from the posterior distributions for 4 parallel MCMC chains. The mixing was proper and confirmed through Gelman-Rubin diagnostic statistic and trace plots. Figures 1 and 2 summarize the results of simulation. Figure 1 illustrates the density plot for the estimated and true differential gene expression parameters for all the genes pooled together. Since there are two subsets of genes, one differentially expressed and the other not-differentially expressed, we expect a bimodal density estimate. Also, since the means of differentially expressed genes differ from the non-differentially expressed genes by 100 units, after normalizing for library effects we expect the bimodal density estimate peaks around 0 and
0.0002 i.e., \( \frac{0.100}{500} = 0.0002 \). This is validated by the bimodal density estimated from the posterior draws of differential gene expression parameters for all the genes that match closely with the true density. The small fraction of the differentially expressed genes have an estimated mean around 0.0002.

**Figure 1:** The density plot compares the accuracy of the hierarchical Bayesian Model (equation 2.2 – 2.6) in estimating true differential gene expression for all the 1000 genes. The samples from the posterior distribution of differential gene expression for all the genes are pooled together and the density is estimated. The estimated density (solid curve) of the true parameters (dotted curve) agree closely. A small fraction of the genes, that are differentially expressed, have
an estimated mean around the true value (0.0002); it is more prominent for the true density plot. The estimated and true densities are both bimodal with closely matching peaks.

Figure 2 illustrates a proposed method to visualize the posterior distributions for the differential gene expression parameters. Instead of visualizing the results for all the genes in the data, we visualize the 95% CI of the posterior distribution for a subset of genes. These subsets are chosen in a non-random manner according to a measure of differential gene expression parameter. We divide the 1000 genes into 10 subsets according to the median of their posterior distribution for differential expression. The first subset contains genes with the 100 highest posterior medians. The last subset contains the genes with 100 lowest posterior medians. We randomly sample five genes from these ten subsets and plot the 95% CIs for these genes on the y-axis and assign colors according to the subset membership of the genes. In Figure 2, the y-axis represents the 95% CIs and the x-axis contains the corresponding gene names. Furthermore, we order the genes according to their posterior medians to make the pattern clearer. From this we observe that the CIs for all the sampled genes from the first subset do not contain 0, and conclude they are differentially expressed. The CI for genes from remaining subsets includes 0, hence a majority of genes in these subsets are not differentially expressed. We also detect false positives, that is, genes that are not differentially expressed, but are declared differentially expressed. Specifically, gene numbers 896, 249, and 250 in the second subset. Notably, the fraction of false positives is close to the actual number of false positives (see: Section 3.3). Figure 2 suggests an effective method of visualizing the results of differential gene expression analysis and represents an overall summary of the model fitting procedure for determining differential gene expression.
Figure 2: Results of gene-specific differential expression using 95% credible intervals (CI; equation 2.7), that contain the true differential expression parameters with 95% probability. The y-axis represents the 95% CIs and the x-axis contains the corresponding genes. The genes are divided into 10 subsets depending on the posterior median; the first subset has the genes with highest 100 posterior medians, and the tenth subsets has the genes with the lowest 100 posterior medians. Five genes are randomly sampled from these subsets, respectively (panels represent subsets). The minimum and maximum values of the 95% CI are mentioned in each panel. The sampled genes are arranged in the decreasing order of their posterior median to make the pattern clearer. Noticeably, the differentially expressed genes in subset 1 separate out from the remaining genes. Further, false positives are detected in subset 2, namely, gene numbers 896, 249, and 250.
The hierarchical Bayesian approach (equation 2.2 – 2.6) detects 152 differentially expressed genes among the 1000 simulated gene profiles. Specifically, we detect the differentially expressed genes with 100% accuracy and detect 52 false positives. Since these results are based on the 95% credible intervals of differential expression, we expect to arrive at a wrong conclusion for approximately 50 genes. Therefore, the hierarchical Bayesian method performs well, even in unreplicated experiments, given the prior assumptions are valid.

3.3 Comparisons of hierarchical Bayesian model with Fisher's Exact Test

We employed the Fisher's Exact Test (FET) to detect differentially expressed genes, and to compare to the results of hierarchical Bayesian modeling. We chose FET because it is the most commonly used method for estimating differential gene expression in unreplicated RNA-Seq data (Marioni et al., 2008). We used R (R Development Core Team, 2011) to perform gene-wise testing and to obtain p-values for all the genes. The p-values were adjusted using the False Discover Rate (FDR) multiple comparison procedure (Benjamini and Hochberg, 1995). FET detects 89 differentially expressed genes at a FDR of 5%. Out of these, only one is a false positive and the remaining 88 genes are true positives, (i.e., they are differentially expressed), and these 89 genes are a subset of the genes detected as differentially expressed by the hierarchical Bayesian modeling approach (Section 3.2). By comparison, FET is conservative when compared to the results of the hierarchical Bayesian model.
In our simulation setting the difference between the means of differentially and non-differentially expressed genes is considerably large by practical standards. Because the gene abundances of the differentially and non-differentially expressed genes will not be well-separated in real data (i.e., a mixture of distribution issue), the results of FET in real data will be even more conservative. In this setting the real benefit of hierarchical Bayesian approach can be seen, since it gains power from the information borrowing among the genes in a treatment. The inference can be further strengthened by borrowing information across treatments among similar genes. This said, the hierarchical Bayesian approach is not without its limitations. Since the results of the hierarchical Bayesian model are based on strong prior assumptions that may not be true in general, validation of such assumptions is required. Obviously, in cases where the prior assumptions can be justified, hierarchical Bayesian will be more powerful. However, for filtering genes for further exploration with good accuracy in unreplicated experiments, we suggest including genes that are declared significant by both FET and hierarchical Bayesian model.

4. Discussion

Statistically, the lack of replication imposes a serious restriction on the detection of differentially expressed genes based on classical approaches. The hierarchical Bayesian model (equation 2.2-2.6) provides an option for detecting differentially expressed genes in unreplicated RNA-Seq experiments. It is more powerful than classical approaches such as Fisher's Exact Test if the prior assumptions are justified. This said, we must remark that this method is based on many, possibly strong, assumptions. First, the assumption of gamma prior distribution imposed on the mean
parameters of gene counts may not be justified, and second, there might be other factors affecting the gene counts that are not included in the second level of hierarchy in equation 2.2. Despite these drawbacks, the model is an effective method of modeling information borrowing and improving inference about differential gene expression that is not possible using the classical approaches. The hierarchical Bayesian method of estimating differentially expressed can be extended to the detection of differentially expressed genes for increasing numbers of replicates and treatment group which will lead to better estimation of within group variation, and thus better overall inference about differential expression of genes. We have implemented the methods discussed here in *JAGS* and *R*. The code can be used to analyze any RNA-Seq data using hierarchical Bayesian model (equation 2.2 – 2.6) with minor modifications.

**Acknowledgement**

We thank the anonymous reviewers and Doug Baumann for helpful comments on an earlier version of this manuscript. This work is funded in part by a National Science Foundation (DBI-0733857) grant to RWD and her colleagues.

**References**


