bioRxiv preprint doi: https://doi.org/10.1101/012823; this version posted December 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

# Shrinkage of dispersion parameters in the binomial family, with applications to genomic sequencing

Sean Ruddy[1], Marla Johnson[2], and Elizabeth Purdom[1,*]

[1]Department of Statistics, UC Berkeley
[2]Department of Biostatistics, UC Berkeley

April 11, 2014

**Abstract**

The prevalence of sequencing experiments in genomics has led to an increased use of methods for count data in analyzing high-throughput genomic data to perform analyses. The importance of shrinkage methods in improving the performance of statistical methods remains. A common example is that of gene expression data, where the counts per gene are often modeled as some form of an over-dispersed Poisson. In this case, shrinkage estimates of the per-gene dispersion parameter have lead to improved estimation of dispersion in the case of a small number of samples. We address a different count setting introduced by the use of sequencing data: comparing differential proportional usage via an over-dispersed binomial model. Such a model can be useful for testing differential exon inclusion in mRNA-Seq experiments or differential allele frequencies in re-sequencing data. In this setting there are fewer such shrinkage methods for the dispersion parameter. We introduce a novel method that is developed by modeling the dispersion based on the double binomial distribution proposed by Efron (1986), also known as the exponential dispersion model (Jorgensen, 1987). Our method (WEB-Seq) is an empirical bayes strategy for producing a shrunken estimate of dispersion and effectively detects differential proportional usage, and has close ties to the weighted-likelihood strategy of edgeR developed for gene expression data (Robinson and Smyth, 2007; Robinson *et al.*, 2010). We analyze its behavior on simulated data sets as well as real data and show that our method is fast, powerful and controls FDR compared to alternative approaches. We provide implementation of our methods in an R package.

## 1 Introduction

In genomic studies, a common approach to high dimensional data is to marginally examine the effect of each feature with a simple statistical test in order to find the most promising feature. A wide-spread example of this type of marginal testing is that of gene expression studies, where each sample consists of measurements of the mRNA levels of tens of thousands of genes from the sample. In this setting, there are generally few samples (sometimes on the order of 10 or less) and thousands of features. A common analysis is to perform a statistical test separately for each gene, for example, a t-test to determine if there is a difference between two groups of samples. In such a paradigm, it has been found that shrinkage of the individual parameter estimates or test statistics greatly improves the results and a great deal of work has been done in different settings to this purpose.

The setting for many of these shrinkage routines has been in the context of continuous, roughly log-normal intensity data from microarray experiments. The growth of relatively cheap sequencing technologies has resulted in sequencing becoming preferred over the previous generation of microarray technologies. The result of sequencing experiments, unlike the continuous intensity measurements of microarray experiments, is often a count of the number of sequences matching a criteria, such as the number of sequences from a particular gene. As a result there has been great interest in how to most effectively use discrete distributions

*To whom correspondence should be addressed. Email: epurdom@stat.berkeley.edu

bioRxiv preprint doi: https://doi.org/10.1101/012823; this version posted December 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

for common tasks that previously relied on normal data. For the setting of marginal testing approaches, this includes appropriate use of over-dispersed models and how to similarly provide shrinkage methods for marginal test-statistics.

A common type of question in this setting is to compare the proportions of sequences measured across different conditions. One setting in which this is common is the case of finding differences in mRNA levels of a gene in different conditions. In this case, the counts per gene are the number of sequenced mRNA that come from that gene. There are different amounts of total sequences collected from different samples, so that the question of interest is whether the proportion of counts allocated to a given gene varies across conditions. In the gene setting, however, the total number of sequences in a sample is in the millions and are spread across thousands of genes, and so the proportions are quite small. For this reason it is common to use a Poisson distribution to model the counts with an offset parameter equal to the total number of sequences (Marioni *et al.*, 2008). Generally an over-dispersed model is preferred, and the prominent modeling technique for over-dispersion has been to use a negative-binomial (Robinson and Smyth, 2007), though some methods have incorporated over-dispersed binomial distributions such as the beta-binomial and the extra-binomial variation of Williams (1982). For all of these gene-expression methods, there has also been focus on creating shrinkage estimators of the dispersion parameter which greatly improve the performance of the methods in small sample sizes (Robinson and Smyth, 2007; Anders and Huber, 2010; Zhou *et al.*, 2011; Yang *et al.*, 2012; Wu *et al.*, 2013; Yu *et al.*, 2013; Leng *et al.*, 2013).

We are interested in a slightly different setting, namely when the Poisson approximation is not valid and the proportions can take on the full range of 0 to 1. In this case, an over-dispersed binomial model is required, and there is much less work in this setting. Our motivating example comes from the question of measuring alternative splicing – when the gene can produce multiple versions of mRNA that included different combinations of the exons of a gene. One simple approach to finding differences in alternative splicing across samples is to measure the number of sequences including the exon and compare it to the number excluding the exon (Shen *et al.*, 2012; Wu *et al.*, 2011), in which case differences in the exon usage appear as a question of comparing proportions across conditions. Another example can be found in resequencing of tumors where the mutations can be present at different proportions in a sample and the question is to compare proportions of mutation inclusion across conditions.

Our focus, like that of the more standard gene expression setting, is on shrinkage estimates of the dispersion parameter, which is the common focus of high-throughput genomic settings. Among the existing shrinkage methods for the dispersion parameter, only few methods exist that use the binomial distribution and are therefore applicable here: the BBSeq method of Zhou *et al.* (2011) and the modified extra-binomial (EB2) method of Yang *et al.* (2012), neither of which were developed for the setting of differential exon usage.

We propose a novel empirical bayes framework for estimating the dispersion parameter for data from a dispersed exponential family. In developing our framework, we model the over-dispersion for a binomial using the double exponential model of Efron (1986), also known as the exponential dispersion model (Jorgensen, 1987). Our empirical bayes method framework is based on the fact that the conditional likelihood of the dispersion parameter can be shown to be approximately Gamma distributed, which we develop below. In addition to being a tractable distribution, the estimates produced from the double exponential model have close ties to quasi-likelihood estimates which are widely used for estimation of binomial over-dispersion. Given this close connection, our method is effectively an empirical bayes method for quasi-likelihood estimation of the dispersion parameter.

Our empirical bayes framework provides two, related versions. The first is the standard empirical bayes estimator (DEB-Seq). The second (WEB-Seq) is an alternative empirical bayes estimator developed by extending the weighted likelihood method of shrinkage of Robinson and Smyth (2007) to the double exponential distribution. We show that for the double exponential, the weighted likelihood method gives a similar form for the shrinkage estimate as our empirical bayes method, and that the empirical bayes methodology provides a novel method of a data-driven estimate for the tuning parameter.

We compare the performance of our method to other methods and demonstrate that in addition to providing a fully automated method for shrinkage, our methods have superior performance on simulated data in the exon inclusion setting. We also apply these methods to mRNA-Seq data from real tumor samples generated by the Cancer Genome Atlas project (Cancer Genome Atlas Research Network, 2011) which suggests that it can similarly control the false discovery rate and find promising targets of splicing.

bioRxiv preprint doi: https://doi.org/10.1101/012823; this version posted December 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

# 2 Modeling the Dispersion

There are several different distributional choices for a dispersion model for the binomial. A very common choice is the beta-binomial model, used by the BBSeq method, which places a beta prior on the standard binomial distribution proportion parameter and results in a distribution with a separate mean and dispersion parameter. The MATS method of Shen *et al.* (2012) creates a dispersed model by placing a uniform prior on the proportion parameter of the binomial.

Another common approach is to use quasi-likelihood methods; the estimates for the proportion (mean) remain the same as that found from a binomial model, but the distribution of the estimate of the mean depends on a dispersion parameter and thus results in greater (or less if under-dispersed) estimates of variability than the standard binomial model in the case of a quasi-binomial. An existing method of analyzing proportions, the modified extra-binomial (EB2) method (Yang *et al.*, 2012), follows an alternative quasi-likelihood approach of Williams (1982).

We focus our methods on the double exponential family (Efron, 1986), also known as the exponential dispersion model (Jorgensen, 1987), which is a probability model that results in estimates closely related to the quasi-likelihood method. This class of distributions, which we will describe in detail below, adds a dispersion parameter to any member of the exponential family. This distribution has the advantage of being closely related to the quasi-likelihood approach and yet still provides a likelihood platform for shrinkage methods. Furthermore, the distribution is itself in the two-parameter exponential family, making calculations and approximations straightforward.

In what follows, the data consists of two $n \times p$ count matrices, $Y$ and $\mathcal{M}$. Each $y_{ij}$ entry is the counts for inclusion of the event *j* for sample *i* and $m_{ij}$ gives the total possible number of counts related to event $j$. For the setting of exon inclusion, $y_{ij}$ would be the number of reads including or overlapping exon $j$ and $m_{ij}$ would be the total number of reads either expressing exon $j$ or skipping exon $j$. The value $y_{ij}/m_{ij}$ is the standard estimate of the proportion of inclusion of the event. For concreteness, we will continue to refer to the events as "exons" with the understanding that the same methods could be applied to other settings. The $m_{ij}$ terms will often be referred to as the total count. In this section, we will focus on the modeling of just a single exon, and we will drop the subscript $j$ when the meaning is clear.

## 2.1 The double exponential distribution

The data for each exon is modeled using the class of double exponential distributions of Efron (1986) which generalizes any exponential distribution by including an over-dispersion parameter. Specifically, assume our initial exponential distribution is given by

$$g_{m_i}(z_i) = \exp(m_i(\eta z_i - \psi(\mu)))dG_{m_i}(z_i),$$

where $\mu = E(z_i)$ and $\eta = \eta(\mu)$ is the link function. For the case of binomial, the link function is the standard logit function, $\eta(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$ and the normalizing function $\psi$ is given by $\psi(\mu) = -\log(1-\mu)$. Note that we follow the notation of Efron (1986) so that the parameterization is such that $E(z_i) = \mu$, implying for the binomial distribution that $z_i$ is the proportion $y_i/m_i$.

Then the dispersed density, with dispersion parameter $\phi$ is given by

$$\frac{c(\mu, \phi, m_i)}{\sqrt{\phi}} \exp\left\{\frac{z\eta_\mu - \psi(\mu)}{\phi}\right\} dF_{m_i}(z_i),$$

where $c(\mu, \phi, m_i)$ is a normalizing constant. The role $\phi$ is reminiscent of the role of the variance parameter in a normal distribution where $\phi > 1$ implies over-dispersion and $\phi < 1$ implies under dispersion. The resulting variance of $Z$ is approximately $\phi\frac{V(\mu)}{m_i}$, where $V(\mu)$ is the variance function for the corresponding (non-dispersed) exponential family ($V(\mu) = \mu(1-\mu)$ in the case of the binomial).

The double exponential distribution can be reparameterized to be a member of the two-dimensional exponential family, and we work with the distribution in its canonical exponential form,

$$\exp(m_i(\lambda Z_i + \theta U_i - A_i(\theta, \lambda)))dF_{m_i}(Z_i)$$

bioRxiv preprint doi: https://doi.org/10.1101/012823; this version posted December 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where $\theta = 1/\phi$, $\lambda = \eta/\phi$, and

$$A_i(\theta, \lambda) = \psi(\mu)\theta - \frac{1}{2m_i}\log\theta - \frac{1}{m_i}\log(c(\mu, \theta, m_i)).$$

The sufficient statistics of $(\lambda, \theta)$ are given by $(Z_i, U_i)$ where $U_i = \psi(Z_i) - Z_i\eta(Z_i) = \rho(Z_i)$. The function $\rho(x) = \psi(x) - \eta_x x$ is determined by the specific exponential distribution. In the case of the binomial,

$$\rho(x) = -\log(1-x) - x\log(\frac{x}{1-x}) = -(x\log x + (1-x)\log(1-x)).$$

For the binomial, $\rho(x)$ is defined on [0,1], with $\rho(0) = \rho(1) = 0$, so it is well defined for all values of $Z_i$.

We are interested in the case where there are covariates that results in different $\mu$ for different samples $i$, $\eta_i = x_i^T\beta$ with $\beta \in R^q$. We focus on the common case in genomic studies where $x_i$ defines $K$ separate groups. Then we have a separate $\eta_k$ for each group $k$, and the joint likelihood can be written as

$$\exp\left\{M\left(\sum_k \lambda_k \frac{M_k}{M}T_k + \theta U - A(\theta, \lambda)\right)\right\} dF(Z_1, \ldots, Z_n) \tag{1}$$

where $M_k = \sum_{i\in k} m_i$, $T_k = \sum_{i\in k}\frac{m_i}{M_k}Z_i$, the standard binomial estimate of the mean $\mu_k$, and $U = \sum_i \frac{m_i}{M}\rho(Z_i)$, For convenience we can rewrite $U$ in terms of the groups, $U = \sum_k \frac{M_k}{M}U_k$, where $U_k = \sum_{i\in k}\frac{m_i}{M_k}\rho(Z_i)$.

**Approximating** $c(\mu, \theta, m)$ The normalizing constant $c(\mu, \theta, m)$ can be computationally expensive to calculate, especially in the context of genomic studies where the maximization routines will need to be calculated for every exon. Efron notes that the normalizing constant can be approximated in the case of the binomial by

$$c(\mu, \theta, m) \approx 1 + \frac{1}{12m}\frac{1-\theta}{\theta}\left(1 - \frac{1}{\mu(1-\mu)}\right) \xrightarrow{m\to\infty} 1.$$

Maximizing the approximate joint likelihood with $c(\mu, \theta, m) = 1$ gives MLE estimates for $\mu$ and $\phi$, $\hat{\mu}_k = T_k$ and

$$\hat{\phi} = \frac{2M\left(\sum_k \frac{M_k}{M}(\rho(T_k) - U_k)\right)}{n} = \frac{1}{n}\sum_k\sum_{i\in k} D(z_i, \hat{\mu}_k)$$

where $D(z_i, \hat{\mu}_k)$ is the deviance of $z_i$ from its estimated group mean $\hat{\mu}_k$. Thus the approximation $c(\mu, \theta, m) = 1$ results in the standard quasi-likelihood estimates of the dispersion based on deviance residuals, giving a likelihood based method that echoes the quasi-likelihood method (Efron, 1986).

This approximation leads to enormous computational savings as well as alignment with the standard quasi-binomial approach, and the methods we develop rely on this approximation for the likelihood.

## 2.2 Conditional Likelihood

For all of the methods of shrinkage, we rely on the conditional likelihood of $\theta$ per exon for the purpose of combining likelihoods across exons independently of the their proportion $\mu$ (Robinson and Smyth, 2007). This provides a likelihood of the data that depends solely on $\theta$ and allows us the opportunity to shrink the estimates of $\theta$ across exons independently of our estimates of $\mu$. Namely, let $\hat{\theta}$ and $\hat{\mu}$ be the joint MLEs of $\theta, \mu$; then the conditional distribution

$$P_{\theta,\lambda}(\hat{\theta}|\hat{\lambda}) = \ell(\theta)$$

defines a likelihood of $\theta$ that is independent of $\mu$ because our distribution is a member of the exponential family. The exact conditional distribution for the double exponential is not tractable; however, we can approximate the conditional distribution using the modified profile likelihood (see Pawitan (2001) for a review),

$$\ell_{AC}(\theta) = \ell(\hat{\lambda}_\theta, \theta) + \frac{1}{2}\sum_k \log\left[\frac{\partial^2}{\partial\lambda_k^2}MA(\theta, \lambda)|_{\lambda=\hat{\lambda}_\theta}\right]$$

bioRxiv preprint doi: https://doi.org/10.1101/012823; this version posted December 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where

$$\hat{\lambda}_\theta = \arg\max_\lambda \ell(\lambda, \theta).$$

Using the approximation $c(\mu, \theta, m) = 1$, the double exponential distribution implies a simple approximate form for the conditional likelihood of the sum of the deviance residuals. Let $S = \frac{1}{2}\sum_k \sum_{i \in k} D(z_i, \hat{\mu}_k)$ be half the total sum of the deviance residuals. Then, the conditional log-likelihood of $\hat{\theta}|\hat{\mu}$ is simplified as

$$\ell_{AC}(\theta) = -\theta S + \frac{n-K}{2}\log(\theta).$$

Note that maximizing the conditional distribution results in the estimate

$$\hat{\phi} = \frac{1}{n-K}\sum_k \sum_{i \in k} D(z_i, \hat{\mu}_k).$$

The change in the degrees of freedom to $n - K$, as compared to the MLE with degrees of freedom $n$, is a common result of using conditional distributions, see for example REML methods in random effects models.

Therefore, the conditional distribution of $\hat{\theta}|\hat{\mu}$ is approximately proportional to $\theta^{\frac{n-K}{2}} \exp(-\theta S)$. A change of variables to $S, T_1, \ldots, T_k$ gives us that

$$P(S|T_1, \ldots, T_k) \approx Gamma(\frac{n-K}{2}, \theta),$$

where $\theta$ is the rate parameter of the Gamma (the distribution can be equivalently expressed with $\phi$ as the scale parameter).

This result does not depend on the form of the underlying exponential family, though we will use the binomial distribution. It could similarly be used for an over-dispersed Poisson distribution for gene expression analysis, for example, though unlike the negative binomial the relationship between the mean and variance would be linear, not quadratic.

# 3 Development of Empirical Bayes Methods

Three general approaches to providing shrinkage of dispersion parameters are common in the literature: 1) modeling of the dispersion values as a function of other aspects of the data, 2) weighted likelihood shrinkage, and 3) empirical bayes shrinkage by estimating prior parameters from the marginal distribution of the data. The first two approaches are implemented in the methods DESeq (Anders and Huber, 2010) and edgeR (Robinson and Smyth, 2007), respectively, in the context of analyzing differential gene expression using a negative binomial distribution. The last approach is implemented in the methods EBSeq (Leng *et al.*, 2013) and DSS (Wu *et al.*, 2013), both for differential gene expression detection and in the case of EBSeq differential isoform expression. In addition, empirical bayes methods are used in a wide range of problems, in particular for gene expression analysis of microarrays where the widely used limma method (Smyth, 2005) provides empirical bayes shrinkage of the variance parameter of a linear model.

We developed methods for the double binomial distribution for all three of these approaches to shrinkage. In what follows, we only present the weighted likelihood and the empirical bayes approaches. The empirical bayes estimator results in our DEB-Seq method, while the fusion of the weighted likelihood and empirical bayes results in the WEB-Seq method.

The first approach, which we do not present, performs a parametric regression of the estimated dispersion per gene or event to other parameters in the model. In the example of DESeq this involves fitting a parametric curve to the estimated variance as a function of the estimated mean, as calculated across many genes. The method of BBSeq (Zhou *et al.*, 2011) use the same general principle, but adapts it to the case of the beta-binomial distribution by fitting a cubic polynomial to the independently estimated, logit transformed dispersion parameters as a function of the fitted values from the GLM. In both cases, once the parametric form is estimated, the shrinkage of the initial dispersion estimates results by replacing the individual dispersion estimates with that given by the fitted curve. This can be considered total shrinkage to the estimate of the population's pattern of variance. In our implementation of the BBSeq version of this approach under the

bioRxiv preprint doi: https://doi.org/10.1101/012823; this version posted December 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

double binomial model, we found that in the exon inclusion setting the resulting p-values grossly failed to control the false discovery rate, and so we did not consider it any further.

We also compare our methods to others existing methods appropriate for the inclusion-exclusion setting. These methods perform shrinkage of dispersion in slightly different ways. The modified extra-binomial shrinkage method (EB2) (Yang *et al.*, 2012) was developed to test for differences in allele frequencies, though it can be easily applied to the setting of differential exon usage. The method employs shrinkage by reparameterizing the variance function in terms of two global parameters that are estimated via linear regression by combining data across all SNPs. MATS (Shen *et al.*, 2012), in addition to assuming a uniform prior for the proportion parameter, adds a correlation between the two conditions being compared, a parameter that is assumed shared by all the exons and thus provides implicit shrinkage.

## 3.1 Weighted likelihood

Wang (2006) gives a general strategy for combining likelihoods across multiple datasets that relies on a likelihood equation that is a weighted combination of the likelihoods of all the experiments. Robinson and Smyth (2007) adapt this idea to the gene-expression setting to make it gene-centric. Each gene $j$ is given a separate weighted log likelihood $\ell^j_{WL}(\theta)$ which is the weighted sum of the individual gene likelihood for gene $j$, $\ell^j(\theta)$, and a common likelihood which is the sum of the individual gene likelihoods for all genes $\ell_{CM}(\theta)$,

$$\ell^j_{WL}(\theta) = \ell^j(\theta) + \delta\ell_{CM}(\theta), \qquad \ell_{CM}(\theta) = \frac{1}{p}\sum_{j=1}^{p}\ell^j(\theta)$$

Then for each gene $j$, the shrunken estimate $\hat{\theta}_j$ is given by maximizing $\ell^j_{WL}(\theta)$. In order to apply this to the dispersion parameter of a negative binomial, Robinson and Smyth (2007) use the conditional likelihood of $\theta$ given the sufficient statistic $T_j = \sum_{i=1}^{p} Y_i$ as the likelihood $\ell^j(\theta)$. We note that for the negative binomial, this requires that the total count parameter $m_i$ be equal for all samples, which is not the case in typical RNA-Seq experiments, so Robinson and Smyth (2008) provide a method of getting pseudo-totals by implementing what they call a quantile-adjusted conditional maximum likelihood procedure (qCML). Essentially, the observed data is adjusted via an iterative algorithm to simulate pseudo-data that is distributed from a negative binomial with equal library sizes.

Though we follow the same strategy to create a weighted likelihood using the approximate conditional likelihood $\ell_{AC}(\theta)$ for the double binomial setting, unlike for the negative binomial distribution, our conditional likelihood for exon $j$ does not require equal $m_{ij}$ for all samples, eliminating the need to create pseudo-data in the implementation.

Again, if we assume that $c(\mu, \theta, m) = 1$ we get an enormous simplification and can analytically solve for $\hat{\theta}^{WL}_j$. Specifically, let $S_j = M_j(R_j - U_j)$ for each exon $j$. Then for a specific exon $j*$, we define its weighted likelihood as

$$\begin{aligned}\ell^{j*}_{WL}(\theta) &= \ell^{j*}_{AC}(\theta) + \delta\frac{1}{p}\sum_{j=1}^{p}\ell^{j*}_{AC}(\theta) \\ &= -\theta(S_{j*} + \delta\frac{1}{p}\sum_{j} S_j) + \frac{n-K}{2}(1+\delta)\log(\theta)\end{aligned}$$

which gives that

$$\hat{\theta}^j_{WL} = \frac{\frac{n-K}{2}(1+\delta)}{S_j + \delta\bar{S}},$$

where $\bar{S} = \frac{1}{p}\sum_j^p S_j$.

The weight $\delta$ given to the common likelihood $\ell_{CM}(\theta)$ is a tuning parameter that must be chosen, and McCarthy *et al.* (2012) suggest that it be chosen so that it is proportional to sample size adjusted by degrees of freedom, with the `edgeR` package assigning a fixed value for $\delta$ by default equal to $\frac{20}{n-K}$.

bioRxiv preprint doi: https://doi.org/10.1101/012823; this version posted December 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

## 3.2 Empirical Bayes

Empirical bayes estimation of the dispersion parameters via an explicit likelihood formulation is a natural way to provide shrinkage estimators of the dispersion parameter. By this we mean, formulate a Bayesian model $Y|\theta \sim F$ and $\theta \sim G_\alpha$ to get estimates of $\theta$, $E_\alpha(\theta|Y)$ and then choose a parameter $\alpha$ by estimating $\alpha$ from the marginal distribution of $Y$. Many distributions, including the double binomial, do not have a prior that gives a tractable form for the marginal distribution of $Y$ to permit easy estimation of $\alpha$ from the data.

However, if we make the approximation that the normalizing constant in the distribution is equal to 1, we showed that there is a simple conditional distribution for the statistic $S$ (defined as half the sum of the deviance residuals), namely that $S|T$ is approximately Gamma distributed with known shape parameter $(n-K)/2$, and rate parameter equal to $\theta$.

Critically, the approximate distribution of $S_j$ is independent of the individual total counts, $m_{ij}$, per exon and sample. It depends only on the total sample size $n$, and therefore is comparable across exons with different total counts $m_{ij}$. This suggests a simple bayesian estimation approach for $\theta_j$ of exon $j$. With known shape parameter, a conjugate prior for the rate parameter of a gamma distribution is itself a gamma distribution, $\Gamma(\alpha_0, \beta_0)$. Then the posterior distribution of $(\theta_j|S_j, T_j)$ is $\Gamma(\frac{n-K}{2}+\alpha_0, S+\beta_0)$ and estimation of $\theta_j$ can be given by the mean of this gamma distribution.

To give an empirical bayes solution, we estimate $\alpha_0$ and $\beta_0$ from the marginal distribution of the $S_j$ across all exons by using the fact that the conjugate Gamma prior for $\theta$ results in an analytical expression for the marginal density of $S_j$ given by the generalized beta distribution (Raiffa and Schlaifer, 1961),

$$p(s_j|t_j) = \frac{s_j^{(n-K)/2-1}\beta_0^{\alpha_0}}{(s_j+\beta_0)^{(n-K)/2+\alpha_0}B((n-K)/2, \alpha_0)}.$$

This means that the $S_j$, conditional on $T_j$, are marginally identically distributed and we can use the joint likelihood of $S_j|T_j$ to find estimates $\hat{\alpha}_0$ and $\hat{\beta}_0$. We estimate $\hat{\alpha}_0$ and $\hat{\beta}_0$ by maximum likelihood estimation.

Then the empirical bayes estimate of $\theta_j$ is given by

$$\hat{\theta}^j_{EB} = E(\theta_j|S_j) = \frac{\frac{n-K}{2}+\hat{\alpha}_0}{S_j+\hat{\beta}_0}.$$

We call this Double exponential Empirical Bayes with application to Sequencing (DEB-Seq).

## 3.3 WEB-Seq

One major advantage of the empirical bayes method in estimating the dispersion is that the amount of shrinkage performed is entirely determined from the data unlike the weighted likelihood method. It is not known, especially in the context of exon usage, if a single default will perform adequately across a range of differing experiments.

It is clear from comparing the two estimators above that they take the same form, implying the weighted likelihood method can be written as an empirical bayes solution where the prior is parameterized by a single variable $\delta$ rather than the two parameters $\alpha_0$ and $\beta_0$,

$$\alpha_0 = \delta\,\frac{n-K}{2}$$
$$\beta_0 = \delta\bar{S}.$$

We are implicitly treating $\bar{S}$ as a fixed value, rather than explicitly conditioning on it, but $\bar{S}$ will normally be the average of thousands if not tens-of-thousands of exons. Note that in the weighted likelihood approach, $\delta$ is assumed to be strictly positive, and $\bar{S}$ will similarly be positive, therefore satisfying the assumptions for $\alpha_0$ and $\beta_0$ to yield a true density.

This naturally suggests an estimator based on this alternative parameterization to fuse these two methods together, which we call a Weighted Empirical Bayes shrinkage with application to Sequencing, or WEB-Seq. This results in a reparameterized marginal density of $S_j$ as,

$$p(S_j|T_j) = \frac{S_j^{(n-K)/2-1}(\delta\bar{S})^{\delta\,\frac{n-K}{2}}}{(S_j+\delta\bar{S})^{\frac{n-K}{2}(1+\delta)}B(\frac{n-K}{2}, \delta\,\frac{n-K}{2})}$$

bioRxiv preprint doi: https://doi.org/10.1101/012823; this version posted December 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Maximizing this density as described above for the empirical bayes method gives us an estimate $\hat{\delta}$ and represents the amount of shrinkage that is performed in the weighted likelihood method as determined by the data. The resulting dispersion estimates are then,

$$\hat{\theta}^j_{WEB} = \frac{\frac{n-K}{2}(1+\hat{\delta})}{S_j + \hat{\delta}\bar{S}}$$

(for details concerning optimization see Supplementary Text, Section 1).

We will see that WEB-Seq has similar performance to the original empirical bayes approach, though it is more conservative and as a result slightly less powerful. Both methods perform well, and we choose to focus on this method largely because it appears to be more robust to violations of the model due to being more conservative.

## 3.4 Estimates of under-dispersion near the boundary

We found in our initial simulations that estimated proportions lying on the boundary (i.e. either exactly 1 or 0, corresponding to a sample that displays no skipping or only skipping) have a large and adverse effect on the false discovery rate (FDR) as the sample size increases (see Supplementary Figure S1). This is because the method estimates under-dispersion for such exons, leading to a large number of false discoveries. Moreover, the effect is worse with larger sample sizes: the effect becomes noticeable around 5-10 samples per group and for increased sample sizes the FDR grows without control. The reason for the increase with larger sample sizes is that exons with proportions *all* 1 or 0 across all samples get a p-value of one, which results in an implicit filter of the data. For exons whose true proportion is near the boundary, the observed data is more likely in low sample sizes to have estimated proportions entirely on the boundary and therefore assigned a p-value of one. In larger sample sizes, there is an increased chance that a non-boundary sample will be observed, allowing the exon to remain in the analysis and have an effect on the FDR results.

There are several ad hoc approaches to this issue. One is to filter exons with mean proportion close to the boundary. While a successful filtering procedure can result in a positive impact on the power of a test (Bourgon *et al.*, 2010b), in this case it is unsatisfactory to have a test-statistic that is so sensitive to the degree of filtering. Another approach is to not allow under-dispersion by setting the dispersion in such cases to one, i.e. binomial variance (and see also the most recent version of DESeq that does not allow the dispersion estimate to be decreased via shrinkage).

We obtained better results by adjusting the degrees of freedom $n - K$ that appears in the Gamma prior to be $n_{eff} - K$, where $n_{eff}$ is the maximum of the number of non-boundary samples and $K + 1$, where K is the number of groups. Note that this means that each exon has a slightly different effective sample size. A similar difference in effective sample size can result when a sample has $m_{ij} = 0$ (see Supplementary Section 1 for details). The result is that it remains possible to estimate under-dispersion for an general exon, but this adjustment makes it more difficult to erroneously estimate under-dispersion on the boundary. With this adjustment no under-dispersion is in fact estimated at any sample size for any exon in our simulated or real data sets, while previously exons with proportion parameters near the boundary were frequently estimated to be under-dispersed. Further, a continuous range of dispersion values $\theta$ are estimated for these boundary exons which we find to be more natural than forcing them all to $\theta = 1$, i.e. no dispersion.

## 3.5 Comparison of conditions

After getting estimates of $\theta$ for each exon, we estimate $\eta$ with $\hat{\eta} = \hat{\eta}_{\hat{\theta}}$, where $\hat{\eta}_\theta$ is the maximum likelihood estimate of $\eta$ for a fixed $\theta$. We then return to our question of testing differences of inclusion between conditions. These can be formulated in the form of contrasts on the vector $\eta$ and we can test per exon the significance of the contrast. In our implementation, we focus on the common two group comparison setting though all of the methods carry through to the more general setting of contrasts of groups. In this setting, let $\beta$ be the value of the contrast defined by the difference of the two group means and $\hat{\beta}_\theta$ the maximum likelihood estimate of $\beta$ for a fixed $\theta$.

For testing the specific null hypothesis $\beta = 0$, we calculate the likelihood ratio statistic for a known value

bioRxiv preprint doi: https://doi.org/10.1101/012823; this version posted December 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of $\theta$ as

$$W_\theta \;=\; log\frac{L_\theta(\hat{\beta}_\theta)}{L_\theta(0)} = \theta(S_{H_0} - S) \;=\; \ell_{AC}^{H_1}(\theta) - \ell_{AC}^{H_0}(\theta) + \frac{K-1}{2}log(\theta),$$

where $S_{H_0}$ is the one-half the sum of the deviance residuals based on estimates calculated under the null hypothesis, and $S$ is the same quantity calculated with estimates under the alternative. In our setting, we replace $\theta$ in the likelihood with our estimate $\hat{\theta}$ in both $L(\hat{\beta}_1)$ and $L(0)$, in which case,

$$W_{\hat{\theta}} \;=\; \hat{\theta}(S_{H_0} - S) \;=\; \ell_{AC}^{H_1}(\hat{\theta}) - \ell_{AC}^{H_0}(\hat{\theta}) + \frac{K-1}{2}log(\hat{\theta}).$$

Asymptotically, $W_{\hat{\theta}}$ should follow a $F$ distribution with $(K-1)$ and $(n-K)$ degrees of freedom (Jorgensen, 1997).

Based on our simulations, we find that this approximation is poor for small sample sizes, e.g. when the size of each group is five or less. Instead we propose an alternative statistic, which re-estimates $\theta$ under the null and alternative, that has much better performance in small sample sizes,

$$W_{\hat{\theta},\hat{\theta}_{H_0}} \;=\; \hat{\theta}_{H_0}S_{H_0} - \hat{\theta}S + \frac{n}{2}\log\left(\frac{\hat{\theta}}{\hat{\theta}_{H_0}}\right) \;=\; \ell_{AC}^{H_1}(\hat{\theta}) - \ell_{AC}^{H_0}(\hat{\theta}_{H_0}) + \frac{K}{2}log(\hat{\theta}) - \frac{1}{2}log(\hat{\theta}_{H_0}).$$

We also find that the shrinkage methods result in less variability in the estimate of $\hat{\theta}$, with the result that the likelihood ratio statistic more closely follows a $\chi^2_{n-K}$ distribution than the standard $F$ distribution for unshrunken estimates.

# 4 Application to Simulated Data

## 4.1 Description of the Simulation

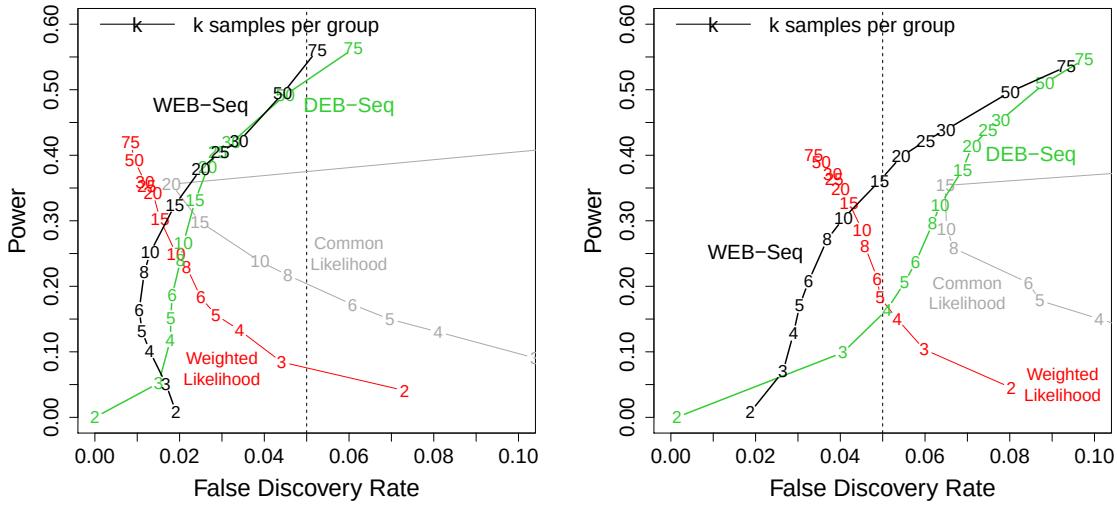
Exon count simulations were created under a two group comparison setting. For the purpose of imitating real data, we chose simulation parameters based on fitting models to 170 Acute Myeloid Leukemia samples generated by the Cancer Genome Atlas project (Cancer Genome Atlas Research Network, 2011), see Supplementary Text, Section 2 for details. We generated data under a double binomial distribution and also a beta-binomial distribution for evaluation of the robustness of our techniques, all of which were developed assuming the data come from a double binomial distribution. Ten percent of the exons were randomly selected to show differential usage between the groups; for these non-null exons, the treatment effect, $\beta_1$, was picked uniformly from the union of [-3,-0.5] and [0.5,3] , accounting for both decreased and increased exon usage. For each simulation, 85,373 exons were simulated and a basic filtering process was applied to remove exons with proportions all equal to 1 or all equal to 0.

We used the simulated data to evaluate the methods developed above: 1) the empirical bayes with a single parameter prior (WEB-Seq) 2) the general two-parameter empirical bayes method for the prior parameters (DEB-Seq), and 3) the weighted likelihood method with $\delta$ fixed to be equal to the default value implemented in edgeR ($\frac{20}{n-K}$, Robinson *et al.* (2010)). In addition to our dispersion-shrinkage methods, we implemented the shrinkage method of BBSeq and EB2. The MATS method described above does not take as input inclusion and exclusion count matrices, but rather creates it own from BAM alignment files, and thus could not be compared on the simulated data.

We also implemented three methods that fit a dispersion parameter per exon but with no shrinkage across exons: quasi-binomial GLM estimation as implemented in the `glm` function in R (R Core Team, 2013), maximum likelihood estimation based on a beta-binomial distribution, and maximum likelihood estimation based on an approximate double binomial distribution where the normalizing constant is set to 1 (see Section 2.1). The quasi-binomial GLM and the double binomial MLE are closely connected, as described in Section 2.1, and are both non-shrinkage counterparts to our methods. However, the quasi-binomial estimation by default uses Pearson residuals to estimate the dispersion, rather than deviance residuals. The beta-binomial maximum likelihood method is the non-shrinkage counterpart of the BBSeq method.

For each procedure, the estimation procedures were performed and the p-values were adjusted to control the FDR to a 0.05 level using the standard Benjamini-Hochberg FDR procedure (Benjamini and Hochberg,

bioRxiv preprint doi: https://doi.org/10.1101/012823; this version posted December 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.




(a) Double-Binomial Simulation

(b) Beta-Binomial Simulation

Figure 1: **Double Binomial Methods** Plotted is the average Power (y-axis) against FDR (x-axis) over various sample sizes across 100 double binomial simulations, based on p-values adjusted to provide a 5% FDR level. The results for a single method across different sample sizes are connected by a line. The numbers that overlay a method denote the power and FDR for that specific sample size (*per group*) in a 2 group comparison. The 5% FDR boundary is given by the dotted vertical line. The data are simulated under (a) a double binomial distribution and (b) a beta-binomial distribution. The methods shown are all based on a double binomial to account for over-dispersion: 1-parameter empirical bayes (WEB-Seq); 2-parameter empirical bayes (DEB-Seq); edgeR default weighted likelihood; and estimation of a single dispersion parameter $\theta$ for all exons (common likelihood). The double binomial MLE is not shown because it's FDR values were beyond the limits of the plot.

1995) implemented in the `p.adjust` function in R (R Core Team, 2013). The final measures of performance were the methods' ability to control false discoveries and their power to detect non-null exons over the 100 simulations.

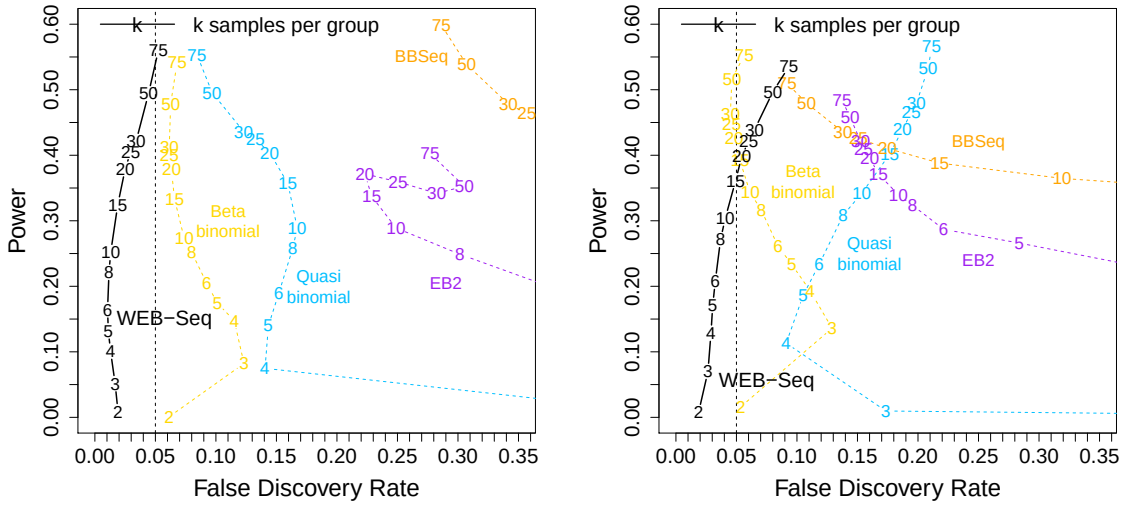
## 4.2 Results of Simulation

We show the true false discovery rate plotted against power for different sample sizes for our shrinkage methods in Figure 1. For data distributed as double binomial, WEB-Seq and DEB-Seq control the FDR at all sample sizes converging on the expected 0.05 FDR for very large sample sizes, with WEB-Seq being more conservative and with slightly less power as a result. Weighted likelihood with a pre-determined tuning parameter (based on edgeR recommendation) is slightly erratic in its control of FDR for extremely small sample sizes, but then adequately controls FDR. However, the pre-determined tuning method becomes over-conservative for large sample sizes and the result is a large drop in power for large sample sizes.

To evaluate the robustness of the methods, we consider data not following the given model, but rather the beta-binomial distribution. Here there appears to be an underlying bias due to the p-values being calculated under the wrong model, and for large sample sizes both WEB-Seq and DEB-Seq converge to around an FDR of 0.10. However for moderate sample sizes (less than 20 per group) the more conservative WEB-Seq still manages to control the FDR; DEB-Seq still has greater power, but has a slight increase of FDR to about 0.07 for moderate sample sizes.

In Figure 2 we compare to other existing methods. WEB-Seq shows great improvement in controlling the FDR compared to all of the other methods – both those that use shrinkage and those that do not. The methods that do not utilize shrinkage have large false discovery rates for small to moderate samples sizes. The beta-

bioRxiv preprint doi: https://doi.org/10.1101/012823; this version posted December 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.




(a) Double-Binomial Simulation

(b) Beta-Binomial Simulation

Figure 2: **Comparison to Alternative Methods:** Plotted is the average Power (y-axis) against FDR (x-axis) over various sample sizes across 100 double binomial simulations, based on p-values adjusted to provide a 5% FDR level, see Figure 2 for details. The alternative methods, both those that perform shrinkage and those that do not, are compared to WEB-Seq: Quasi-binomial (no shrinkage), BBSeq, EB2, and Beta-binomial MLE estimates (no shrinkage).

binomial MLE with no shrinkage performs the best of the alternative methods under both simulations, but still has an FDR larger than the target 5% for less than 10-15 per group. Quasi-binomial does not come close to controlling the FDR even with double binomial distributed data until more than 50 samples per group (with an FDR of 9.7%). In contrast, WEB-Seq controls the FDR at the desired level across the full range of sample sizes for the double binomial data, only showing increased FDR in the beta-binomial data for large sample sizes.

The EB2 and BBSeq methods both implement shrinkage and rely on beta-binomial dispersion models. Both methods do not even minimally control the FDR in our simulated setting, even for beta-binomial data with large sample sizes, with FDR values ranging from 0.73 and 0.85 ($n = 2$) to 0.092 and 0.14 ($n = 75$) for BBSeq and EB2, respectively. Similarly, the double binomial GLM (not plotted) fails to control the FDR at any sample size and converges to an FDR at around 40%.

Many biological studies focus on the top performing exons for validation and followup analysis, especially when there are large numbers of significant results. We find that the WEB-Seq method not only provides better global performance, but also gives p-values that better prioritize the truly non-null exons, i.e. the ranks of the exons based on the p-values. In Figure 3, we plot the average proportion of false discoveries in the top-ranked exons for simulations with five samples per group. We see that the alternative methods have a much higher proportion of false positives in the top-ranked exons compared to WEB-Seq. We see similar behavior for beta-binomial distributed data (Supplementary Figure S3). This demonstrates that the difference in the global FDR and power we see in Figure 2 is due to the actual choice of statistic, not merely a problem in the distributional assumptions for creating p-values.

**Filtering** In Supplementary Table S1 we compare the proportion of exons that are affected by our adjustment to the degrees of freedom (Section 3.4), for both the simulated and real data. We see that 78% of the exons are affected by these changes, with 43% having a large reduction of 5 or more. Interestingly, we see that the real data more closely follows the double binomial simulation in this respect, rather than the beta-binomial.

A major source of the large levels of false discovery in several of the methods are also boundary exons,

bioRxiv preprint doi: https://doi.org/10.1101/012823; this version posted December 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

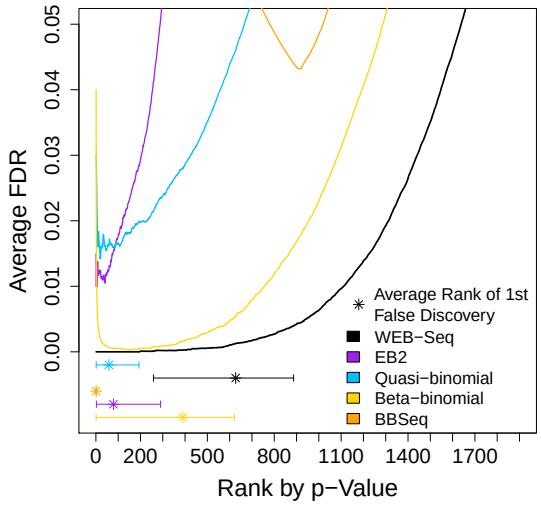



Figure 3: **False Discoveries by Rank.** Plotted is the average proportion of false discoveries (y-axis) in the top $x$ exons (x-axis) for a 5 versus 5 comparison. For each method, the inner 95% range at which the FDR first becomes positive across the 100 simulations is given by the horizontal lines with the average marked by an asterisk. The data shown here is simulated under a double binomial distribution (for beta-binomial distributed data, see Supplementary Figure S3).

where the true proportions are close to 0 or 1. In Supplementary Figure S2, we show similar performance plots for each of the methods after removing those exons whose mean proportion across all samples is different from 0 or 1 by less than 0.05. We see that EB2 and quasi-binomial with no shrinkage have much better levels of FDR when these exons are filtered and are similar to WEB-Seq. However, they have a much reduced power as a result of the filtering. The beta-binomial method, on the other hand, is not affected by the filtering.

# 5 Application to Real Data

In order to have a reasonable setting for detecting differential alternative splicing we downloaded RNA-Seq data from two different tumor types also sequenced by the TCGA: Stomach and Ovarian. For comparisons between these two sets of tumors, we expect that there should be differences in alternative splicing due to the simple fact that the tumors originated from two different tissue types, and tissue-specific alternative splicing is well documented (Pan *et al.*, 2008). The RNA-Seq data was realigned using TopHat v1.4.1 (Trapnell *et al.*, 2010), and exon inclusion and exclusion counts were calculated for exons annotated by ENSEMBL version 66 (Flicek *et al.*, 2013). See Supplementary Text, Section 3.2 for more details about the processing of the data.

We create a 'null' situation to compare the methods, where the two groups that are compared are both of the same tissue type. We note that these are tumor samples, so there may be differential alternative splicing in the different tumors even though they are the same tissue type, but since the two groups of samples are randomly chosen this is unlikely to be a significant factor. We compare the proportions of exons called significant in the null setting across methods, though we note that if we believe there are no significant exons this is not a measure of FDR directly, since *any* discoveries in an all-null setting would imply that the rate of false discoveries is 1.

We demonstrate the performance of the double binomial based shrinkage methods in Table 1. We see that all methods call almost no exons significant. We also see that like in the simulations, WEB-Seq has less 'power', i.e. the least significant calls under the real setting. This shows a possible weakness of the WEB-Seq method, where in the 2 versus 2 real comparison setting WEB-Seq lacks power and makes 0 calls. In the simulation we similarly saw that the WEB-Seq method was conservative, holding the FDR lower than

bioRxiv preprint doi: https://doi.org/10.1101/012823; this version posted December 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 1: **Comparison of Double Binomial based Methods.** Shown in the table below are the percentage of exons called significant from the Tissue Data under the null and real scenarios described above for the methods we developed based on the double binomial distribution. The total number of exons is $412,002$. The rates are percentages out of only those exons that had at least one skipping event, a number which varies with sample size but is roughly 1/4 of all exons. See Supplemental Table S2 for both the precise number of exons called and analyzed.

| Sample | WEB-Seq | | DEB-Seq | | Wt-Likelihood | |
|---|---|---|---|---|---|---|
| Size | Real | Null | Real | Null | Real | Null |
| 2 vs 2 | 0.00† | 0.00† | 0.00† | 0.00† | 0.12 | 0.00† |
| 3 vs 3 | 0.48 | 0.00† | 1.38 | 0.00† | 1.39 | 0.00† |
| 4 vs 4 | 3.67 | 0.00† | 4.45 | 0.00† | 3.55 | 0.00† |
| 5 vs 5 | 2.94 | 0.00† | 3.92 | 0.00† | 3.18 | 0.00† |
| 6 vs 6 | 5.35 | 0.00† | 6.00 | 0.00† | 4.74 | 0.00† |
| 7 vs 7 | 6.97 | 0.00† | 7.38 | 0.00† | 5.91 | 0.00† |

†Percentage is exactly zero.

necessary.

We also ran EB2, BBSeq and MATS on the TCGA data sets (Table 2). MATS could only be run in the null setting, as the stomach and ovarian samples were of different types and the MATS software could not handle this setting. BBSeq and EB2's poor control of the FDR in the simulated data appears to be echoed in the real data. As demonstrated in Table 2, EB2 finds roughly 7% of the exons significant and BBSeq finds 3-14% significant. For comparison with MATS, we applied WEB-Seq to the inclusion/exclusion count matrices produced by MATS. In the null setting, MATS appears to have a call rate between 1.8% and 3.9% (646 to 1,557 exons called significant), while WEB-Seq makes at most one call for any given sample size. These false positive rates do not directly compare with the FDR rates from the simulations, since FDR depends on the total number found significant. For comparison, if 10% of the exons were found significant and the method had 100% power, the false positive rate would have to be 0.6% to get an FDR of 5%, and in practice would need to be even lower since not all of the truly significant exons will be detected. A 3-7% false positive rate would then mean a minimum FDR of 21-38% and likely much higher. This indicates that the large rates of FDR shown in our simulation appear to be supported by implementation on the real data.

For the comparison of two different tissue types, we see many more calls made by BBSeq. The EB2 method, despite it's high false positive rate on the null setting, does not give many more calls than WEB-Seq, except in small samples sizes. Given the high false positive rate on the null set and our simulation results, it is likely that the additional calls of these methods represent a much higher level of false discoveries than reported. EB2 and quasi-binomial methods are severely impacted by data on the boundary, with a large proportion of their significant calls coming from boundary exons, particularly in small sample sizes (Supplementary Table S4); again this corresponds well to the behavior we saw in the simulated results.

## 5.1 Alternative approach of DEXSeq

We made a further comparison of the performance of our method to another popular method of finding differential alternative splicing in exons, DEXSeq (Anders *et al.*, 2012). The DEXSeq framework is quite different than the inclusion/exclusion framework. They assume knowledge of the identification of exons to genes and fit a linear model per gene to the counts per exon, allowing for an individual exon effect i.e. how different an exon is from the overall mean gene expression. Then they find alternatively spliced exons by detecting exons who have different exon effects in the two groups. In fitting this model, they use a negative binomial model for the exon counts with shrinkage of the dispersion in the same manner as DESeq for gene expression.

We emphasize that DEXSeq is not just an alternative statistical method for exon counts, but uses signifi-

bioRxiv preprint doi: https://doi.org/10.1101/012823; this version posted December 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 2: **Comparison to Alternative Methods.** Shown in the table below are the percentage of exons called significant from the Tissue Data under the null and real scenarios described above. DEXSeq was post-filtered to have the same set of exons as the inclusion/exclusion setting. For all the results shown below, except for MATS, the total number of starting exons is $412,002$ but the rates are percentages out of only those exons that had at least one skipping event, a number which varies with sample size but is roughly 1/4 of all exons. The results from MATS are based on a different set of exon data produced internally by MATS, roughly 35,000 exons; WEB-Seq results are not shown on this set of exons, but WEB-Seq makes at most one significant call on the MATS set of exons (for sample sizes 3, 5 & 7) and zero for other sample sizes. See Supplemental Table S3 for the precise number of exons called and the results from the non-shrinkage methods.

| Sample | DEXSeq | | EB2 | | BBSeq | | MATS | WEB-Seq | |
|---|---|---|---|---|---|---|---|---|---|
| Size | Real | Null | Real | Null | Real | Null | Null | Real | Null |
| 2 vs 2 | 2.62 | 0.00 | 3.99 | 6.88 | 7.18 | 3.63 | 3.45 | 0.00† | 0.00† |
| 3 vs 3 | 11.29 | 0.01 | 4.47 | 7.15 | 9.38 | 8.07 | 1.77 | 0.48 | 0.00† |
| 4 vs 4 | 20.86 | 0.00 | 4.64 | 6.68 | 10.72 | 5.60 | 2.45 | 3.67 | 0.00† |
| 5 vs 5 | 16.56 | 0.00 | 4.43 | 6.47 | 11.59 | 7.14 | 2.74 | 2.94 | 0.00† |
| 6 vs 6 | 22.11 | 0.02 | 4.79 | 6.48 | 11.98 | 14.87 | 3.93 | 5.35 | 0.00† |
| 7 vs 7 | 26.99 | 0.01 | 4.87 | 6.18 | 12.44 | 14.27 | 3.39 | 6.97 | 0.00† |

†Percentage is exactly zero.

cantly different aspects of the mRNA-Seq data, compared to the inclusion/exclusion setting, and the starting input data for the two approaches is entirely distinct (all counts overlapping an exon versus counts skipping versus overlapping). DEXSeq does not make use of the information of junctions skipping the exons, except in their contribution to reads overlapping an exon. Further, it requires a gene model and would not be applicable in a setting where the gene models are not available, unlike the inclusion/exclusion approach. However, DEXSeq can in principle find differential usage of exons that do not have inclusion/exclusion data resulting from their alternative usage, for example exons that are removed upstream from the beginning of transcripts and/or downstream of the end of transcripts. For these reasons, it is not clear that you can make a reasonable comparison between WEB-Seq and DEXSeq. However, DEXSeq is a popular method for detecting alternative splicing with just exon counts, so we attempt some basic comparisons.

When we compare our methods to DEXSeq, we note that the paradigm of inclusion/exclusion offers one possibly significant advantage regardless of the statistical method. In the inclusion/exclusion paradigm, those exons that show no reads skipping the exon in any sample of any group are naturally excluded (by getting a p-value of 1 by definition). Because of the difference in exons evaluated between the methods, we first concentrate on comparing the performance of DEXSeq for just the same set of exons that are used in WEB-Seq; namely, we run DEXSeq on all exons, as required by the algorithm, and then filter out those not found to have any skipping events. For this set of exons the performance of DEXSeq to WEB-Seq in the null setting is roughly equivalent, while DEXSeq calls many exons significant in the real setting (Table 2).

To get a sense of the value of observing reads that skip an exon, we can compare to the annotation used, here Ensembl version 66, where some exons are annotated as constitutive (i.e. should not be skipped in any of the transcripts if the annotation is completely accurate) and others as alternative. Of the constitutive exons (12.7% of the exons in the data), only 1% (529 exons) show *any* reads skipping the exon in *any* of the 30 tissue samples, while 35.0% of those annotated as alternatively spliced show skipping. This strongly suggests that the implicit removal of exons with no skipping junctions is preferentially removing null exons, which ultimately can increase the power (Bourgon *et al.*, 2010a). There is no natural way to exclude such exons in the DEXSeq model since the constitutive exons are actually important in building the gene model, though the post-analysis filtering we described above could be implemented to eliminate exons that were not skipped.

Clearly this implicit filtering can also be a disadvantage if many alternatively spliced exons are excluded because of a lack of sufficient reads to detect the skipping event. We view this as less of a practical disadvantage because we find that in practice practitioners are likely to want evidence in the form of junction

bioRxiv preprint doi: https://doi.org/10.1101/012823; this version posted December 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 3: **Percent of single-exon calls made by DEXSeq, by annotation and skipping event.**

| | % of Total Single-Exon Calls | | | | % of Total Significant Calls | | | |
|---|---|---|---|---|---|---|---|---|
| Sample Size | 2 vs 2 | 3 vs 3 | 5 vs 5 | 15 vs 15 | 2 vs 2 | 3 vs 3 | 5 vs 5 | 15 vs 15 |
| Alternatively Spliced | 81.75 | 70.74 | 67.70 | 54.26 | 85.89 | 87.20 | 87.01 | 88.11 |
| Non-skipped Constitutive | 18.00 | 29.06 | 32.12 | 45.74 | 13.88 | 12.58 | 12.79 | 11.72 |
| Skipped Constitutive | 0.25 | 0.19 | 0.18 | 0.00† | 0.23 | 0.22 | 0.20 | 0.17 |
| Total Calls | 2,356 | 2,085 | 1,681 | 916 | 16,229 | 59,871 | 87,144 | 186,570 |

†Percentage is exactly zero.

reads skipping an exon to have faith in calling an exon alternatively spliced. But more generally, because the inclusion/exclusion paradigm relies heavily on reads that span the junctions of exons, which are a small percentage of all reads, a criticism of the inclusion/exclusion paradigm is that it relies on a lower number of reads and could have lower power. It is clear in the real data comparison that DEXSeq makes more significant calls than WEB-Seq even when limited to the same set of exons. It is difficult to directly evaluate whether the additional calls made by DEXSeq are on average finding more true discoveries than false ones.

Comparing exon calls to the annotation is one way of roughly assessing the performance for calling differential exon usage: about 12% of constitutive exons are called significant by DEXSeq (Table 3). This is roughly their total representation in the data so DEXSeq does not appear to be preferentially finding exons annotated to be alternatively spliced. However, directly comparing the exons found by DEXSeq with the annotation has the problem that the method is designed to detect only differential usage as compared to the average usage of all exons in the gene as opposed to the actual exon that is alternatively spliced; these could be different, for example, if many of the exons in a gene are alternatively spliced so that relative to the mean the unusual exon are the few that are not alternatively spliced, a point the authors of DEXSeq make as well (Anders *et al.*, 2012). Using this logic, we instead compare only exons that are the sole exon called significant in their gene; when these "single-exon" significance calls are compared to the annotation, even a larger percentage are annotated as constitutive *and* furthermore have no reads skipping them in the data for any of the 30 samples (18%-46%, Table 3). In comparison, in WEB-Seq, 0.2% of the significant exons (or 76 exons) are annotated as constitutive and all of them, by definition, have reads skipping them to at least justify the call of significance (this calculation is based on all exons WEB-Seq analyzes since it is reasonable to directly compare all the calls made by WEB-Seq to the annotation, not just "single-exon" calls).

We can also evaluate the data properties of the significant exons to evaluate whether they demonstrate data characteristics that would lead us to trust the call. In Supplementary Figure S5 we compare the density of the log-Fold-Change between the groups of the odds-ratio of skipping an exon for the significant calls made by both methods. WEB-Seq clearly has a much stronger tendency to find exons with large differences in the skipping proportion, which is not surprising given that that is the basis of its test statistic, unlike DEXSeq. More striking is that for DEXSeq there are significant peaks at 0, indicating many of the exons found significant by DEXSeq do not show evidence of differential exon usage in the form of a difference in the proportion of skipping counts. The constitutive exons, in particular, are completely centered at zero. This could be because of the lack of identification of the correct exon, explained above; when we examine the "single-exon" genes which are presumed to target the appropriate exon, these exons show slightly greater propensity to be removed from zero (Figure S4c).

Ultimately, we find the inclusion/exclusion paradigm, as implemented with our methods, concentrates the analysis on those exons with tangible evidence of alternative splicing as well as directly highlighting the specific exons of interest. We suspect this will also be an effective way of preventing a large source of false discoveries as well as being robust to the behavior of the other exons in the gene.

bioRxiv preprint doi: https://doi.org/10.1101/012823; this version posted December 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

### 5.2 Computation

In an exon usage analysis, the method used needs to potentially be able to handle all exons, a number which for the human genome can be in the hundreds of thousands. Under the inclusion/exclusion paradigm there are natural filters that significantly reduce the set of exons under analysis and we have seen ranges between 40K and 200K exons in this scenario for real RNA-Seq experiments when looking at just protein coding exons. Because we have exact analytical solutions for our estimators of WEB-Seq, the only numerical optimization involves the calculation of the estimates of the prior parameters of the gamma distribution. In a 5 versus 5 setting analyzing a total of 106,208 exons WEB-Seq required only 18 seconds on a single core computer running an AMD Opteron 6272, 2.1 GHz processor. WEB-Seq can be run for any size experiment on a single core, personal laptop in under a minute.

Many other methods that we compared to require much greater computation time. In the gene-based model of DEXSeq, such filters are not appropriate and therefore the number of exons being analyzed can range between 300K and 400K exons. If we artificially set the number of exons to be 106,208 as in WEB-Seq, DEXSeq requires 3.7 hours compared to 18 seconds using WEB-Seq. However, these 106,208 exons are after implementing the inclusion/exclusion filters that are not valid for DEXSeq, which needs all exons in the gene. Without those filters, the same dataset actually contained 367,675 exons which took DEXSeq 35 hours to complete on a single core (DEXSeq allows the use of multiple cores, but for comparison purposes we restricted it to a single core). BBSeq is also time intensive requiring 4 hours to complete the analysis; EB2 only requires 2 minutes. With the MATS algorithm, it is difficult to compare computational times directly since it is only implemented in a format that requires processing of all of the raw sequences to create the counts, but in the 5 vs 5 example, it took a total of 35 hours, as well as needing approximately 150GB of available space for intermediate files.

## 6 Discussion

We have provided a novel method of providing shrinkage estimators for the dispersion parameter of a dispersed exponential family of distributions. We rely on a dispersion model that is closely connected to the common quasi-likelihood method for providing over-dispersion to a binomial, which are widely used and numerically robust. By making use of the distributional form of Efron (1986), we have shown that there is a simple formulation of the approximate conditional distribution of the dispersion parameter and that this form provides a straightforward empirical bayes method to estimate shrinkage. In effect, we provide a likelihood-based empirical bayes method for quasi-likelihood estimation of the dispersion parameter. By further relating this empirical bayes method to weighted likelihood shrinkage methods (Robinson and Smyth, 2007), we give a non-standard parameterization of the Gamma prior that leads to an alternative estimator in this class of estimators that demonstrates some areas of improved performance. Further, our distributional form and the empirical bayes method that results do not require any tuning parameters, unlike the weighted likelihood methods of edgeR.

We focus on the binomial distribution, but our entire development is completely general and can be applied to any distribution from the exponential family, though the mean-variance relationship implied by the quasi-likelihood model might not be the desired one for other distributions, such as the Poisson.

While our shrinkage method is quite general, we have focused on our motiving example, detecting differential usage of exons between conditions in order to detect group-specific alternative splicing. In particular, our data examples were drawn from mRNA-Seq data, and the simulations were based on parameters estimated from that same data. There are other settings that require the comparison of a large number of proportions between groups, for example in the setting of comparing allele frequencies, and it is possible that the performance would differ in those settings due to differences in the properties of the data.

Within the specific setting of detecting differential alternative splicing, there are other approaches of detection in mRNA-Seq data besides even the two we explored here (exon counts and inclusion/exclusion counts). Exon inclusion counts may not be the most appropriate for every setting. In particular, there are many methods for estimating the expression levels of individual isoforms (Denoeud *et al.*, 2008; Jiang and Wong, 2009; Trapnell *et al.*, 2010; Richard *et al.*, 2010; Salzman *et al.*, 2010), and comparison of the isoform levels may give more insight into alternative splicing particularly when there is a great deal of information about the transcriptome that is being sequenced.

bioRxiv preprint doi: https://doi.org/10.1101/012823; this version posted December 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

However, in our experience there are still many cases where researchers find themselves without a well constructed annotation of the transcriptome, and often rely on de-novo methods to construct genes and/or transcripts (Trapnell *et al.*, 2010; Guttman *et al.*, 2010) in an effort to understand the use of alternative splicing as a means of cell regulation. This is an extremely complicated problem, and these de-novo methods can be unreliable and unstable if used on a single, small experiment or without significant depth. In contrast, inclusion/exclusion counts rely on detection of exons and splice sites, which are much simpler problems. Such inclusion/exclusion counts still provide useful, interpretable information about the undergoing of alternative splicing within the organism and our method gives a reliable technique for the statistical analysis of such data.

# 7 Acknowledgements

We wish to thank Christopher Paciorek of the Berkeley Statistical Computing Facility for helpful input into the coding of the algorithms and parallelization of the comparisons. The results published here are in part based upon data generated by The Cancer Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at http://cancergenome.nih.gov/.
*Funding:* This research was partially funded by NIH grant U24 CA143799 and NSF grant DMS-1026441.

# References

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Nature Precedings*. doi:10.1038/npre.2010.4282.2. URL `http://precedings.nature.com/documents/4282/version/2`.

Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Research*, **22**(10), 2008–2017. ISSN 1549-5469. doi:10.1101/gr.133744.111. URL `http://dx.doi.org/10.1101/gr.133744.111`.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 289–300. URL `http://www.jstor.org/stable/2346101`.

Bourgon, R., Gentleman, R., and Huber, W. (2010a). Independent filtering increases detection power for high-throughput experiments. *Proc Natl Acad Sci USA*, **107**(21), 9546–51. doi:10.1073/pnas.0914005107.

Bourgon, R., Gentleman, R., and Huber, W. (2010b). Reply to talloen et al.: Independent filtering is a generic approach that needs domain specific adaptation — pnas. *Proceedings of the National Academy of Sciences of the United States of America*. URL `http://www.pnas.org/content/early/2010/11/05/1011698107.full.pdf+html?etoc`.

Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**(7353), 609–615.

Denoeud, F., *et al.* (2008). Annotating genomes with massive-scale RNA sequencing. *Genome Biol*, **9**(12), R175. doi:10.1186/gb-2008-9-12-r175.

Efron, B. (1986). Double Exponential Families and Their Use in Generalized Linear Regression. *Journal of the American Statistical Association*, **81**(395), 709–721.

Flicek, P., *et al.* (2013). Ensembl 2013. *Nucleic Acids Research*, **41**(D1), D48–D55. ISSN 1362-4962. doi:10.1093/nar/gks1236. URL `http://dx.doi.org/10.1093/nar/gks1236`.

Guttman, M., *et al.* (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nature Biotechnology*, **28**(5), 503–10. doi:10.1038/nbt.1633.

bioRxiv preprint doi: https://doi.org/10.1101/012823; this version posted December 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Jiang, H. and Wong, W. H. (2009). Statistical inferences for isoform expression in RNA-seq. *Bioinformatics*, **25**(8), 1026–32. doi:10.1093/bioinformatics/btp113.

Jorgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society. Series B (Methodological)*, **49**(2), pp. 127–162. ISSN 00359246. URL http://www.jstor.org/stable/2345415.

Jorgensen, B. (1997). *The Theory of Dispersion Models*. Chapman & Hall, London.

Leng, N., *et al.* (2013). EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, **29**(8), 1035–1043. ISSN 1460-2059. doi:10.1093/bioinformatics/btt087. URL http://dx.doi.org/10.1093/bioinformatics/btt087.

Marioni, J. C., *et al.* (2008). Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, **18**(9), 1509–17. doi:10.1101/gr.079558.108. URL http://genome.cshlp.org/content/18/9/1509.long.

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic acids research*, **40**(10), 4288–4297. ISSN 1362-4962. doi:10.1093/nar/gks042. URL http://dx.doi.org/10.1093/nar/gks042.

Pan, Q., *et al.* (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, **40**(12), 1413–5. doi:10.1038/ng.259.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford Univ Press.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Raiffa, H. and Schlaifer, R. (1961). *Applied statistical decision theory*. Studies in managerial economics. Division of Research, Graduate School of Business Adminitration, Harvard University, Boston. URL http://opac.inria.fr/record=b1082847.

Richard, H., *et al.* (2010). Prediction of alternative isoforms from exon expression levels in RNA-seq experiments. *Nucleic Acids Research*, **38**(10), e112. doi:10.1093/nar/gkq041. URL http://nar.oxfordjournals.org/cgi/content/full/38/10/e112.

Robinson, M. and Smyth, G. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**(21), 2881. doi:10.1093/bioinformatics/btm453. URL http://bioinformatics.oxfordjournals.org/cgi/content/full/23/21/2881.

Robinson, M. D. and Smyth, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, **9**(2), 321–332. ISSN 1468-4357. doi:10.1093/biostatistics/kxm030. URL http://dx.doi.org/10.1093/biostatistics/kxm030.

Robinson, M. D., Mccarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, **26**(1), 139–140.

Salzman, J., Jiang, H., and Wong, W. H. (2010). Statistical modeling of RNA-SEQ data. Technical Report BIO-252, Division of Biostatistics, Stanford University, Palo Alto. URL http://statistics.stanford.edu/~ckirby/techreports/BIO/BIO%20252.pdf.

Shen, S., *et al.* (2012). MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Research*, **40**(8), e61. ISSN 1362-4962. doi:10.1093/nar/gkr1291. URL http://dx.doi.org/10.1093/nar/gkr1291.

Smyth, G. K. (2005). Limma: linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, and W. H. R. Irizarry, editors, *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420. Springer, New York.

bioRxiv preprint doi: https://doi.org/10.1101/012823; this version posted December 16, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Tai, Y. and Speed, T. (2006). A multivariate empirical bayes statistic for replicated microarray time course data. *The Annals of Statistics*, **34**(5), 2387–2412. URL http://www.jstor.org/stable/25463512.

Trapnell, C., *et al.* (2010). Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, **28**(5), 511. doi:doi:10.1038/nbt.1621. URL http://www.nature.com/nbt/journal/v28/n5/abs/nbt.1621.html.

Wang, X. (2006). Approximating Bayesian inference by weighted likelihood. *Canadian Journal of Statistics*, **34**(2), 279–298.

Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **31**(2), pp. 144–148. ISSN 00359254. URL http://www.jstor.org/stable/2347977.

Wu, H., Wang, C., and Wu, Z. (2013). A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics*, **14**(2), 232–243. ISSN 1468-4357. doi:10.1093/biostatistics/kxs033. URL http://dx.doi.org/10.1093/biostatistics/kxs033.

Wu, J., *et al.* (2011). Splicetrap: a method to quantify alternative splicing under single cellular conditions. *Bioinformatics*, **27**(21), 3010–3016. doi:10.1093/bioinformatics/btr508. URL http://bioinformatics.oxfordjournals.org/content/27/21/3010.abstract.

Yang, X., *et al.* (2012). Extra-binomial variation approach for analysis of pooled DNA sequencing data. *Bioinformatics*, **28**(22), 2898–2904. ISSN 1460-2059. doi:10.1093/bioinformatics/bts553. URL http://dx.doi.org/10.1093/bioinformatics/bts553.

Yu, D., Huber, W., and Vitek, O. (2013). Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size. *Bioinformatics*, **29**(10), 1275–1282. ISSN 1367-4811. doi:10.1093/bioinformatics/btt143. URL http://dx.doi.org/10.1093/bioinformatics/btt143.

Zhou, Y. H., Xia, K., and Wright, F. A. (2011). A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics (Oxford, England)*, **27**(19), 2672–2678.