The Annals of Applied Statistics
2016, Vol. 10, No. 2, 690–725
DOI: 10.1214/15-AOAS871

SHRINKAGE OF DISPERSION PARAMETERS IN THE BINOMIAL FAMILY, WITH APPLICATION TO DIFFERENTIAL EXON SKIPPING[1]

BY SEAN RUDDY, MARLA JOHNSON AND ELIZABETH PURDOM

University of California, Berkeley

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 gene expression data, where the counts per gene are often modeled as some form of an overdispersed Poisson. Shrinkage estimates of the per-gene dispersion parameter have led to improved estimation of dispersion, particularly 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 overdispersed binomial model. We are motivated by our interest in testing for differential exon skipping in mRNA-Seq experiments. We introduce a novel shrinkage method that models the overdispersion with the double binomial distribution proposed by Efron [*J. Amer. Statist. Assoc.* **81** (1986) 709–721].

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 [*Bioinformatics* **23** (2007) 2881–2887, *Bioinformatics (Oxford, England)* **26** (2010) 139–140]. We analyze its behavior on simulated data sets as well as real data and show that our method is fast, powerful and gives accurate control of the FDR compared to alternative approaches. We provide implementation of our methods in the R package `DoubleExpSeq` available on CRAN.

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 features—for example, a t-test per feature to detect differences between two groups of samples. Gene expression studies are a well-known example of this type of marginal testing, where the features of each sample consist of measurements of the mRNA levels of tens of thousands of genes from the sample. Generally, this setting consists of few samples (sometimes on the order of 10 or less) and thousands of features or genes. In such a paradigm,

Received March 2015; revised August 2015.

[1] Supported in part by NIH Grant U24 CA143799 and NSF Grant DMS-10-26441.

Key words and phrases. Empirical Bayes, dispersion estimation, over-dispersed binomial, alternative splicing, mRNA-Seq.

previous work shows that shrinkage of the individual parameter estimates or test statistics greatly improves the results.

Due to the growth of relatively cheap sequencing technologies, sequencing has become the preferred technology for many genomic experiments. Sequencing experiments generally result in a count of the number of sequences matching a criterion, such as the number of sequences from a particular gene. However, the setting for many commonly used shrinkage routines was originally in the context of continuous, roughly log-normal intensity data from microarray experiments. As a result, there has been a growth in analytical methods that effectively use discrete distributions in settings that previously relied on normal data. For marginal testing approaches, new methods for sequencing data include appropriate use of overdispersed distributional models as well as shrinkage techniques.

A common type of analysis in sequencing experiments compares the counts of sequences measured across different conditions, such as the two-group setting described above. For example, one can search for variations in the mRNA levels of a gene in different conditions, where the data are a count of the number of sequenced mRNA for each gene. Samples differ in the total number of sequences, so a sensible metric may be whether the proportion of sequences from a given gene varies across conditions. A sample has millions of sequences spread across thousands of genes, so the proportion from a single gene is quite small. For this reason a Poisson distribution is an obvious choice for modeling the counts, with an offset parameter equal to the total number of sequences [Marioni et al. (2008)]. Most preferred gene-expression methods now use overdispersed models, typically the negative-binomial distribution [Robinson and Smyth (2007)], though some methods have incorporated overdispersed binomial distributions such as the beta-binomial. For these gene expression methods, shrinkage estimators of the dispersion parameter have greatly improved the performance of the methods in small sample sizes [Anders and Huber (2010), Leng et al. (2013), Robinson and Smyth (2007), Wu, Wang and Wu (2013), Yang et al. (2012), Yu, Huber and Vitek (2013), Zhou, Xia and Wright (2011)].

We are interested in a slightly different setting, namely, marginal testing that compares proportions that take on the full range of values from 0 to 1. Our motivation for considering such proportions is detecting differences in alternative splicing between conditions. Specifically, we are interested in an approach that measures per exon whether an exon is excluded (“spliced out”) more frequently in some conditions than others, often summarized by the proportion of sequences showing inclusion called the *proportion spliced in* (PSI or Ψ) [Barbosa-Morais et al. (2012), Brooks et al. (2014), Pan et al. (2008), Shen et al. (2012), Venables et al. (2009), Wu et al. (2011)]. Focusing on PSI per exon has the advantage that it is a relatively simple summary of the data that also makes use of the large amount of information available in those sequences that span introns. We describe and give the necessary background to this problem in Section 2 below.

Unlike gene expression settings, a Poisson approximation to the binomial distribution is not adequate in this setting because the proportion parameter of the binomial distribution cannot be assumed small, a characteristic that also excludes the many existing shrinkage and overdispersion methods that improve upon the fit of the basic Poisson model. We propose modeling the data with an overdispersed binomial distribution that is a member of a general family of dispersed exponential distributions proposed by Efron (1986), called the double exponential family of distributions.

Using this family of distributions, we develop novel shrinkage estimates of the dispersion parameter for data whose distribution is a member of the double exponential family of distributions. Our estimates are empirical Bayes estimates that are general to any distribution in the family of distributions and are based on the fact that the distribution of the dispersion parameter of this family of distributions can be shown to be approximately Gamma distributed, which we develop below. The double exponential family produces estimates that have close ties to quasi-likelihood estimates, which are widely used for estimation of binomial overdispersion. 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 also an empirical Bayes estimator with a different parameterization of the prior; it is related to the weighted likelihood method of shrinkage of Robinson and Smyth (2007) applied to the double exponential family of distributions, but, unlike that approach, our empirical Bayes methodology provides 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)], and the results suggest that our method can similarly control the false discovery rate and find promising targets of splicing when applied to actual mRNA sequencing data. Furthermore, our methods have very little computational overhead compared to many existing methods.

2. Differential exon usage.

2.1. *Alternative splicing*. The static genetic code found in the DNA of each cell must be read and converted into the molecular products that are used throughout the cell, for example, proteins. Specifically, a portion of the DNA, called a gene, encodes which specific amino acids will make up a protein. When a protein is needed, the corresponding DNA is transcribed into an independent copy of the DNA, called mRNA; the genetic code contained in the mRNA is then translated into the string of amino acids that form the protein.

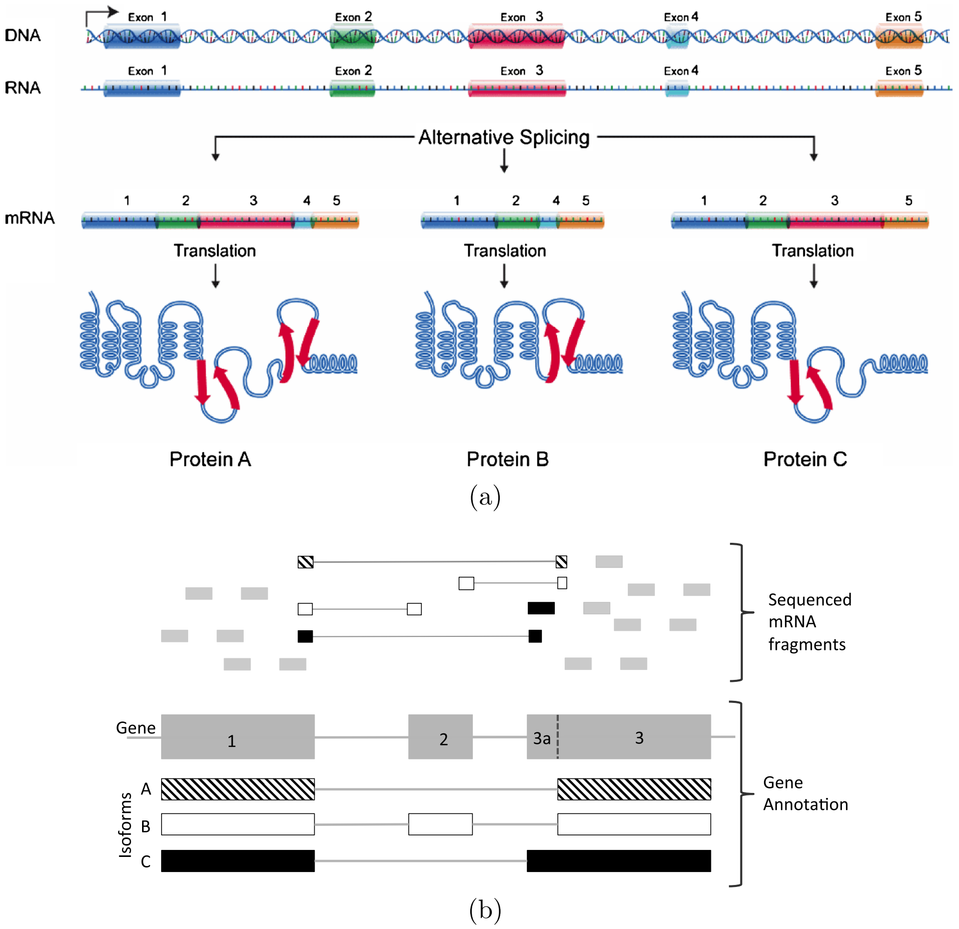
In eukaryotic cells, the process of copying the DNA into mRNA itself has stages. A direct copy of the DNA is created, called a pre-mRNA, and the pre-mRNA is then further processed into an mRNA. In many eukaryotes, the processing of pre-mRNA includes removing portions of the pre-mRNA. This means that the code for the protein contained in the final mRNA transcript is not an exact copy of the code found in the DNA of the cell. The process of cutting out portions of the mRNA product is called *splicing*. In many complex organisms, including human and many model organisms like fruitflies and mice, the splicing of a pre-mRNA product is more complex because the cell can remove different parts of the pre-mRNA depending on the environment. The end result is that a single gene or stretch of code on the DNA can result in a diversity of mRNA transcripts. This is referred to as *alternative splicing* [see Figure 1(a) for an illustration].

The different possible transcripts that can result from a single gene are often called *isoforms* of the gene. The DNA of a gene can be thought of as being divided into *introns* (the portions that will be removed) and *exons* (the portions that will remain), though because of alternative splicing some regions may be retained in one isoform and removed in another isoform. Exons that are contained in all isoforms are called *constitutive* exons, versus alternatively-spliced exons that are only in a subset of the isoforms. The number of isoforms present in a gene and the extent of overlap between isoforms vary by gene. Determining the possible isoforms based on the DNA of a gene is difficult, and the set of all possible isoforms is not completely determined even for well-characterized genomes like human.

2.2. *Sequencing of mRNA for alternative splicing*. Sequencing technologies allow researchers to determine the DNA nucleotides that make up the input DNA strand(s) (or mRNA). Advances in the efficiency of such technologies now make it practical to intensively sequence the mRNA in a cell in order to quantify the relative amounts of different mRNA in a cell (also called *mRNA expression*). Unlike earlier microarray technologies, sequencing the mRNA in a cell measures all mRNA without needing any prior information as to what genes or isoforms exist. Due to this, the sequencing of mRNA provides a great deal of information about alternative splicing in the cell.

While sequencing methods allow for direct sequencing of mRNA, the most commonly used technologies for mRNA still do not allow for an entire mRNA transcript to be sequenced. Instead, the mRNA must be cut into smaller fragments before sequencing, meaning that the sequences that are obtained are only a portion of the mRNA from which they came. To determine mRNA abundance, these partial fragments must be identified with a gene or isoform, usually by aligning or mapping the sequences to the known genome.

If the mRNA has undergone splicing, then the mRNA fragment may not directly match the genome, but it is usually possible to match a fragment back to the genome by allowing gaps in the match of the mRNA to the genome, which will correspond to introns that were removed due to splicing [see, for example,



(a)

(b)

FIG. 1. Illustration of alternative splicing and sequencing. (a) *Illustration of the process of creating multiple isoforms from a single gene* (*DNA*) *courtesy of* National Human Genome Research Institute (2014) (*www.genome.gov*). *A RNA copy of a gene is created* (*pre-mRNA*) *that contains all the gene's exons and introns. The pre-mRNA in this example can then be spliced in one of three ways, and each possibility leads to a different mRNA* (*isoform*) *and protein product.* (b) *Illustration of aligned sequenced fragments when a gene has multiple isoforms. The top set of small boxes represents the sequenced fragments, while the bottom boxes represent the annotation of a gene region* [*not the same gene as in* (a)]. *The sequenced fragments on top are shown based on their alignment to the genome below. Most sequenced fragments are represented as a single box, meaning they completely match a genome region. Some sequenced fragments have to be split into smaller boxes with connecting lines; these represent* junction fragments, *where because of alternative splicing the fragment matches two disconnected regions of the genome. The annotation below the sequenced fragments consists of all known exons* (*i.e., the gene*), *shown by the grey exons; below the gene are its three isoforms and their corresponding exons. Where possible, the sequenced fragments are also colored by the isoform to which they can be uniquely identified; most fragments are grey because it is not possible to identify the isoform from which they originate.*

Trapnell, Pachter and Salzberg (2009), Wu and Nacu (2010)]. Such fragments that do not directly map to a single exon but rather span multiple exons are often called *junction* fragments or split-reads [Figure 1(b) shows examples]. Junctions fragments are often interesting for alternative splicing because they can give direct evidence of splicing. For example, in Figure 1(b), isoforms A and B differ because B includes an exon 2 and A does not (called an exon skipping event); fragments that connect exon 1 to exon 3 without including exon 2 must come from isoform A, while those fragments that connect exon 2 to the adjacent exons 1 and 3 must come from isoform B.

Most fragments entirely match the genome and are even contained entirely within an exon. Since the bases just around the junctions between exons represent only a small portion of mRNA, junction fragments are much less common than fragments mapping within the exons, and therefore fragments matching the genome are important for accurate quantification of the overall expression of the mRNA. Within-exon fragments are less likely to provide direct evidence as to the isoforms from which they came [though we give an example in Figure 1(b) of a fragment contained entirely within an exon that must come from isoform C]. However, the quantity of fragments mapping to different regions can provide indirect evidence of alternative splicing. For example, in the case of the skipping of exon 2, described above, expression of isoform A implies less total expression of exon 2 (i.e., fewer fragments mapping to it) relative to exons 1 and 3.

2.3. *Measuring differential alternative splicing.* The information in the sequenced fragments may be summarized in different ways with respect to detecting differences in alternative splicing. Such summaries can be roughly thought of as making different use of fragments mapping to the junctions versus the interior of the exons. Depending on how the information is summarized, different statistical approaches are appropriate. Our methods here focus on analyzing data resulting from one particular approach to summarizing the information regarding alternative splicing. This method makes use of the junction fragments, and we call this an inclusion–exclusion summarization, with the summary statistic often called the *percent spliced in* or *exon inclusion percentage*.

The inclusion–exclusion approach simplifies the information in isoforms into two contrasting patterns of interest and counts how many fragments agree with one pattern compared to the other. The comparison of isoforms A and B based on their junction fragments described above is such an example. In this case, junctions that skip exon 2, joining exons 1 and 3, show evidence of isoform A, while those that include exon 2 show evidence of isoform B. This comparison ignores any other exons in the gene, only considering the three exons relevant for assessing the exon skipping event. To summarize this information, we calculate the proportion of these fragments including exon 2, which is called the percent spliced in for exon 2 abbreviated PSI or Ψ. More generally, the inclusion–exclusion approach breaks each gene into well-defined simple alternative splicing patterns (e.g., skipping exons, alternative $5'/3'$ splice starts), and evaluates whether the PSI changes between different conditions across all genes [see Barbosa-Morais et al. (2012), Brooks et al. (2014), Pan et al. (2008), Shen et al. (2012), Venables et al. (2009), Wu et al. (2011) for examples]. Programs like MATS [Shen et al. (2012)], JuncBase [Brooks et al. (2011)], DiffSplice [Hu et al. (2013)] and SpliceTrap [Wu et al. (2011)] take an annotation file of the transcriptome and create counts for these types of alternative splicing events.

In our data analysis that follows, we consider PSI values defined more broadly than just these classical definitions of splicing. These classical alternative splicing events are difficult to identify computationally for the whole genome, and they also do not include many other more complicated events. We define a PSI for each exon based on the percentage of counts—from any isoform—that include the exon out of all fragments that either include or skip the exon [see Supplementary text, Section 3.5.1, for more details Ruddy, Johnson and Purdom (2015a)]. This approach creates a measure of PSI for every exon in the annotation, which we find more satisfying than limiting to just exons that are in simple types of combinations.

We emphasize that, regardless of how we define an "event," all of these inclusion–exclusion summarizations result in the same kind of data structure from a statistical point of view: Y successes out of m trials for each event.

Alternative splicing between samples can be compared using other summaries of mRNA-Seq data. Our methods are not generally appropriate for these types of summaries (except as noted below), but we mention them here so as to make clear the distinction.

Isoform estimation. Another way to summarize the information in a gene is to estimate the individual isoform estimates directly. Most fragments do not map uniquely. This results in a deconvolution problem, where the observed expression of a fragment is a convolution of the expression of the individual isoforms from which it could originate [Denoeud et al. (2008), Jiang and Wong (2009), Katz et al. (2010), Richard et al. (2010), Salzman, Jiang and Wong (2010), Trapnell et al. (2010)]. Individual isoform expression levels can be individually compared in a similar way to gene expression estimates [see EBSeq, Leng et al. (2013)] which provides for corrections specific to isoform analysis) or specific features of the isoforms can be compared for differential isoform usage [e.g., rSeqDiff, Shi and Jiang (2013)].

Furthermore, some researchers use isoform expression measures to create PSI values per exon, where instead of using inclusion–exclusion *counts*, they use the isoform measurements to calculate, per exon, the percentage of the overall isoform expression that comes from isoforms including the exon [Shi and Jiang (2013)]. In our simple example, exon 2 is contained only in isoform B, and the isoform-based PSI for exon 2 is the isoform abundance for isoform B as a fraction of the total (sum) abundance for the gene (this estimates a similar quantity to our preferred way of calculating PSI per exon described above). Since isoform estimates are continuous, and not counts, this form of a PSI would not have the discrete statistical structure of Y successes out of m trials; however, it is common to convert isoform expressions into expected counts, in which case the expected counts would have that discrete nature.

The ability to estimate isoform expression levels depends on having complete knowledge of all possible isoforms in the gene (called the transcriptome), as well as having enough distinguishing information between the isoforms to deconvolve the isoform expression from the overall expression. However, many organisms of interest to researchers do not have well-established transcriptomes. Some computational methods attempt to construct the set of isoforms de novo, based on mRNA-Seq data [Guttman et al. (2010), Richard et al. (2010), Trapnell et al. (2010)], but this problem is extremely complicated, and these de-novo methods can be unreliable and unstable if used on a single, small experiment or without sufficient numbers of sequenced fragments. For this reason, it can be preferable to use alternative methods of summarizing the data for detecting differential alternative splicing.

Relative exon usage. Another approach to alternative splicing evaluates the relative expression of exons to detect alternative splicing, ignoring the specific information in the junction fragments. As we noted above, if isoform A is expressed, this should be apparent in relatively less total number of counts overlapping exon 2; this can be assessed without any recourse to what specific exons are joined together by junction fragments. Therefore, the input per exon is the counts of all fragments overlapping an exon, and relative exon usage does not make use of the information of how junction fragments skip the exons, except in their contribution to the count of fragments overlapping an exon.

In practice, many technical aspects of sequencing can create biases so that some exons are more likely to get sequenced, which affects the overall count of an exon relative to other exons in the gene. For this reason the goal of methods based solely on exon counts is to find *differential* changes in relative exon expression across groups. This is based on the assumption that the sequencing-related biases would be the same across samples, which may not be true.

Methods that analyze relative exon counts to detect differential alternative splicing, like DEXSeq [Anders, Reyes and Huber (2012)] and the diffSplice method of voom [Law et al. (2014)], fit a linear model per gene to the counts per exon, including an individual exon effect to quantify the relative exon expression, that is, how different an exon is from the overall mean gene expression. Then, they find differentially spliced exons by detecting exons who have different exon effects in the two groups. In fitting this model, standard count distributions like the negative binomial model, along with the corresponding shrinkage of the dispersion, can be used in the same manner as for gene expression.

Like the inclusion–exclusion approach, relative exon usage does focus on individual exons rather than global isoform estimates. However, the linear model also implies that the effect for an exon is defined relative to all other exons in the gene, which means that identification of an exon effect in the linear model does not directly translate to alternative splicing differences in that exon [Anders, Reyes and Huber (2012)]. For example, if many of the exons in a gene are alternatively spliced, then, relative to the mean, the unusual exons with large exon effects are actually the few that are *not* alternatively spliced. Similarly, introns should not be included in a relative exon usage analysis since they will overwhelm the gene model. Neither of these situations are a problem with the inclusion–exclusion framework, where the measurement of splicing given by the PSI is local and independent of other, nonadjacent, exons in the gene. On the other hand, relative exon counts can find differential usage of exons that cannot have inclusion–exclusion data, for example, exons at the beginning or end of a gene do not produce skipping fragments even if some isoforms do not use those exons.

We note that, in addition to targeting the correct exon, the paradigm of inclusion–exclusion offers one possibly significant advantage compared to an analysis of the relative usage of exon counts, regardless of the specific statistical method. In the inclusion–exclusion paradigm, those exons that show no fragments skipping the exon in *any* sample of any group are naturally excluded by getting a p-value of 1 by definition (and similarly for those exons which are always skipped, such as introns). To get a sense of the value of using information about junctions, we look at our real data example of 30 samples from 2 tissue types (described in Section 5.4) and compare the exons with skipping junctions to our preexisting annotation of the gene. Of the exons annotated as constitutive (12.7% of the exons in the data), only 1% (529 exons) show *any* junction fragments skipping the exon in *any* of the 30 tissue samples, while 35.0% of those exons annotated as alternatively spliced have such junction fragments. This strongly suggests that the implicit removal of exons with no skipping junctions is preferentially removing exons that are not alternatively spliced, which ultimately can increase the power [Bourgon, Gentleman and Huber (2010)]. Such exons are not easily excluded in the linear model analysis of exon counts; to the contrary, the constitutive exons are actually important in building the gene model—though a post-analysis filtering could be implemented to eliminate exons that were never skipped.

Clearly, this implicit filtering can also be a disadvantage if many alternatively spliced exons are excluded because of a lack of sufficient sequenced fragments 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 fragments skipping an exon in order to have faith in the call of an exon as alternatively spliced or to design an experiment to test the result. But a more general concern is that because the inclusion–exclusion paradigm relies heavily on fragments that span the junctions of exons—a small percentage of all fragments—it relies on a lower number of fragments and could have lower power.

3. Modeling the dispersion.

Several different distributional choices are possible for the dispersion model for the binomial. A common choice is the beta-binomial model which places a beta prior on the proportion parameter of a standard binomial distribution. This distribution is not a member of the exponential family, and the solution for the MLE does not have a closed form. This model was used recently in the setting of methylation data [Dolzhenko and Smith (2014), Feng, Conneely and Wu (2014), Sun et al. (2014)], but less so in that of gene expression settings [with Zhou, Xia and Wright (2011) and Hardcastle and Kelly (2013) being exceptions]. In a related fashion, the MATS method of Shen et al. (2012) creates a dispersed model by placing a uniform prior on the proportion parameter of the binomial which is a specific example of a beta prior where the beta parameters (and thus dispersion) are set in advance; unlike the other methods previously cited, MATS was actually developed for the setting of measuring PSI and detecting alternative splicing from mRNA-Seq data.

Another common approach is 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. The presence of overdispersion results in greater variability of the mean than the nondispersed model (or less if underdispersed). An existing method of analyzing proportions in a genomic setting is the modified extra-binomial (EB2) method of Yang et al. (2012), which follows an alternative quasi-likelihood approach given in Williams (1982) where the variance is aligned to match that of a beta-binomial; again, this was not developed for differential exon splicing, but for comparing allele frequencies between populations.

We focus our methods on the double exponential family [Efron (1986)], which is a set of proper probability distributions that result 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 the development of shrinkage methods. Furthermore, the distribution is itself in the two-parameter exponential family, making calculations and approximations straightforward. Because our method of shrinkage for dispersion estimates generalizes beyond just the binomial distribution, we will describe the development in general terms using notation for the entire double exponential family of distributions rather than concentrating on the binomial setting (which we will refer to as the "double binomial" distribution).

Notation. In what follows, the data from every exon consists of a pair Y_{ij} and m_{ij}. Y_{ij} refers to the count of the number of times event j was included for sample i. For the setting of exon inclusion, Y_{ij} would be the number of fragments overlapping exon j. m_{ij} gives the total possible number of counts related to event j; in the exon setting this would be the total number of fragments for the exon—the sum of the number of those expressing exon j and those skipping exon j. The value Y_{ij}/m_{ij} is the "percent spliced in" (PSI) value and is the standard binomial estimate of the probability of inclusion. For concreteness, we will assume that the PSI is per exon, with the understanding that the same methods could be applied to other ways of defining "included" and "excluded" fragments. The m_{ij} we will call the total count, meaning the total number of fragments for the exon, both including and skipping the exon.

For the rest of this section, we will focus on the modeling of the distribution of Y_{ij} for just a single exon j, and therefore we will drop the subscript j when the meaning is clear.

3.1. *The double exponential family of distributions.* The double exponential family of distributions of Efron (1986) generalizes any distribution in the exponential family of distributions by including an overdispersion parameter. Specifically, assume that the naive choice for the distribution of Y_i is a distribution in the exponential family, such as Binomial or Poisson. We also assume that each Y_i has a corresponding "sample size" m_i. For the binomial, m_i is clearly the total number of trials. For other distributions, m_i might be taken as 1 for all samples. We follow the notation of Efron (1986) so that we transform Y_i into a random variable Z_i whose mean μ under the naive distribution is independent of the sample size m_i, for example, $Z_i = Y_i/m_i$ for the binomial family. We write the p.d.f. of the naive distribution in canonical exponential family form by

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

where $\eta = \eta(\mu)$ is the link function relating the mean μ to the canonical parameter η, and $\psi(\mu)$ is the normalizing function. For the case of binomial, the link function is the standard logit function, $\eta(\mu) = \log(\frac{\mu}{1-\mu})$, and the normalizing function ψ is given by $\psi(\mu) = -\log(1-\mu)$. We can also define the standard deviance residual of Z_i from the mean μ in terms of the canonical parameters of the exponential family,

$$D(Z_i, \mu) = 2m_i\bigl\{\bigl(\eta(Z_i) - \eta(\mu)\bigr)Z_i - \bigl(\psi(Z_i) - \psi(\mu)\bigr)\bigr\}.$$

Note that $D(Z_i, \mu)$ is the deviance based on the naive (nondispersed) model.

Efron (1986) proposes adding a dispersion parameter ϕ to any distribution in the exponential family so that the dispersed distribution has a p.d.f. given by

$$\frac{c(\mu, \phi, m_i)}{\sqrt{\phi}} \exp\left\{-\frac{D(z_i, \mu)}{2\phi}\right\} dF_{m_i}(z_i),$$

where $c(\mu, \phi, m_i)$ is a normalizing constant. The role of ϕ is reminiscent of the role of the variance parameter in a normal distribution where $\phi > 1$ implies overdispersion and $\phi < 1$ implies underdispersion.

Approximating $c(\mu, \phi, m_i)$. The normalizing constant $c(\mu, \phi, m_i)$ 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 shows that $c(\mu, \phi, m_i) \to 1$, as $m_i \to \infty$. In our exon setting m_i is the total number of fragments for the exon in a particular sample—including both the fragments overlapping the exon and the junctions skipping the exon—and is *not* the number of samples.

The accuracy of this approximation for finite m_i depends on the values of μ and ϕ. For the ranges of ϕ that we see in the data, the approximation generally holds well for μ away from the boundary values of 0 or 1, and our shrinkage procedure further mitigates the negative effects seen due to the approximation in estimating $\hat{\phi}$ in these boundary settings, as we discuss further in the Supplementary text, Section 1.4 [Ruddy, Johnson and Purdom (2015a)].

Approximating $c(\mu, \phi, m_i)$ with 1 results in the mean of Z_i remaining approximately μ while the variance of Z_i from the overdispersed distribution becomes approximately $\phi\frac{V(\mu)}{m_i}$, where $V(\mu)$ is the variance function for the naive (nondispersed) distribution from the exponential family. If the naive distribution is the binomial distribution, the corresponding overdispersed distribution has variance approximately $\phi\frac{\mu(1-\mu)}{m_i}$, while for a Poisson the overdispersed distribution has variance approximately $\phi\mu$.

Comparison of double binomial to beta-binomial. By analogy, the beta-binomial distribution can be parameterized in terms of the mean μ and dispersion parameter $\rho \in (0, 1)$, which represents the correlation between the m_i Bernoulli draws. Then the variance of Z_i from a beta-binomial is given by

$$\frac{1}{m_i}\mu(1-\mu)\bigl(1+(m_i-1)\rho\bigr) = \frac{V(\mu)}{m_i}\bigl(1+(m_i-1)\rho\bigr).$$

Note that beta-binomial only allows for overdispersion, unlike the double binomial, though that is not of great concern in most genomic studies where we expect overdispersion.

We compare their density functions in Supplemental Figure S1 after aligning their first two moments [i.e., choosing ρ so that $\phi = (1 + (m_i - 1)\rho)$]. Not surprisingly, for low levels of dispersion, the double binomial and beta-binomial are quite similar since they both converge to a binomial distribution as the dispersion vanishes. They show the greatest differences for large dispersion parameters or when the mean is near the boundary of values of 0 and 1. When the dispersion is extremely large [$\log(\phi) = 3$], the double binomial puts more mass on the boundary values compared to the beta-binomial. However, dispersion values this large are rare in the data we are examining (Supplemental Figure S2). More applicable for our data is the differences between the two distributions when μ is near 0 (or 1), where the beta binomial concentrates the mass much closer to 0 while the double binomial allows for more variability; when we look at the number of zeros in real data, it matches simulations from a double-binomial more closely than that of a beta-binomial (Section 4.3). Over the rest of the range of parameters, the distributions are relatively similar, but the distribution of the beta-binomial is more difficult to work with analytically than the double binomial, particularly when we use the approximation of $c(\mu, \phi, m_i) = 1$ for the double binomial.

Multiple groups. For applications, we are interested in the case with covariates that give rise to different μ for different samples i. We focus on the common case in genomic studies where the covariates simply define K separate groups of samples, and we want to compare the mean proportion between the groups for every exon. The most common example is the two-group comparison where $K = 2$. The two groups might be disease and normal samples, samples with a gene knocked out versus those without, or any comparison of treatment samples versus control samples. Higher values of K correspond to multiple group comparisons, such as time course data, where each group might correspond to observations observed at the same time point.

In the case of multiple groups, we have a separate mean μ_k for each group k. Maximizing the joint likelihood with the approximation that $c(\mu_k, \phi, m_i) = 1$ gives simple analytical MLE estimates for μ_k and ϕ,

$$\hat{\mu}_k = \sum_{i \in k} \frac{m_i}{M_k} z_i, \qquad \hat{\phi} = \frac{1}{n} \sum_k \sum_{i \in k} D(z_i, \hat{\mu}_k),$$

where $M_k = \sum_{i \in k} m_i$ is the total coverage of all samples in group k. Thus, the approximation $c(\mu_k, \phi, m_i) = 1$ results in the standard estimates of μ_k from the nondispersed distribution and estimates of the dispersion based on deviance residuals. These estimates are the same as those in the quasi-likelihood setting, giving a likelihood-based method that closely resembles the results of 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 assume this approximation.

3.2. *Distribution of $\hat{\phi}$.* A key component in our methods of shrinkage is a statistic that relies solely on the parameter ϕ and is independent of μ. This allows us the opportunity to shrink the estimates of ϕ based on shared information across exons independently of the value of μ for a particular exon.

Distributions from the double exponential family are also members of the two-parameter exponential family. This implies that the conditional distribution of $\hat{\phi}|\hat{\mu}$ defines a conditional likelihood independent of μ and that the conditional distribution can be easily approximated using a modified profile likelihood [see Pawitan (2001) for a review]. This conditional likelihood can be used for shrinkage of the estimates $\hat{\phi}$ as in Robinson and Smyth (2007).

However, if we further make use of the approximation $c(\mu_k, \phi, m) = 1$, it can be shown that $\hat{\phi}$ is asymptotically independent of the vector of sample means, $\hat{\mu}$. Furthermore, $\hat{\phi}$ is asymptotically Gamma distributed,

$$\hat{\phi} \xrightarrow{\mathcal{L}} \text{Gamma}\left(\frac{n-K}{2}, \frac{2}{n}\phi\right),$$

where $\frac{2}{n}\phi$ is the scale parameter of the Gamma [see Supplementary text, Section 1, Ruddy, Johnson and Purdom (2015a)]. The quality of this approximation depends heavily on the parameters μ and m_i in a similar manner as the approximation of the normalization constant, and exons with boundary values of μ have a poor approximation. The quality of the approximation does not appear to be very sensitive to small samples sizes n [see discussion in the Supplementary text, Section 1.4, Ruddy, Johnson and Purdom (2015a)].

This approximation does not depend on which naive distribution formed the starting point of the analysis, though we will use the binomial distribution since we are analyzing percentages (PSI). For example, it could similarly be used for an overdispersed Poisson distribution for gene expression analysis, which we discuss in the conclusion.

4. Development of empirical Bayes methods.

We now develop the methodology underlying our two shrinkage methods for the dispersion, as well as test statistics for comparing differential skipping between groups.

Our methods rely on the empirical Bayes strategy for shrinkage of the dispersion parameter, which is widely used in genomic settings. For example, EBSeq [Leng et al. (2013)] and DSS [Wu, Wang and Wu (2013)] proposed empirical Bayes methods for differential gene expression detection in sequencing data, and in microarrays analysis the widely used limma method [Smyth (2005)] provides empirical Bayes shrinkage of the variance parameter of a normal distribution. We develop an empirical Bayes method for the family of double exponential distributions, which we call DEB-Seq (Double exponential Empirical Bayes with application to Sequencing).

We also consider the weighted likelihood approach to shrinkage implemented for the negative binomial distribution in the widely used gene expression method edgeR [Robinson and Smyth (2007)]. We show that when applied to the approximate likelihood of $\hat{\phi}$ based on the double exponential family of distributions, it gives an estimator of a similar form as our empirical Bayes estimator, except with a different parameterization of the prior distribution. To distinguish the method with this form of parameterization, we refer to this as the WEB-Seq method (Weighted Empirical Bayes shrinkage with application to Sequencing).

4.1. *Double exponential 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 the following general strategy for estimating a parameter θ: formulate a Bayesian model $Y_j|\theta_j \sim F$ and $\theta_j \sim G_\alpha$, let the estimate $\hat{\theta}_{\alpha j}$ be given by $E_\alpha(\theta_j|Y)$, and then choose the parameter α by estimating α from the joint marginal distribution of $Y_1, \dots, Y_n|\alpha$ based on the observed data, $Y_1, \dots, Y_n$. This results in an estimate $\hat{\theta}_{\hat{\alpha} j}$ which not only has the desirable shrinkage properties of Bayesian estimators but is also an explicit method for estimating the parameter α.

Many distributions, including the distributions in the double exponential family, do not have a prior that gives a tractable form for the marginal distribution of Y to permit easy estimation of α from the data. However, if we make the approximation that the normalizing constant is equal to 1, we show in Section 1 of the Supplementary text that $\hat{\phi}$ is approximately Gamma distributed [Ruddy, Johnson and Purdom (2015a)].

This suggests a simple Bayesian model of ϕ_j for each exon j. A conjugate prior for the scale parameter of a gamma distribution with a known shape parameter is the inverse gamma distribution, $\mathrm{IG}(\alpha_0, \beta_0)$[2]. Applying this conjugate prior to the scale parameter ϕ_j, we can formulate the Bayesian model

$$\hat{\phi}_j|\phi_j \sim \Gamma\left(\frac{n-K}{2}, \frac{2}{n}\phi_j\right), \qquad \phi_j \sim \mathrm{IG}(\alpha_0, \beta_0).$$

We then can define a Bayesian point estimate for the precision $(1/\phi_j)$ as the mean of the posterior distribution, and use this to give a Bayesian estimate of ϕ (see Supplemental text, Section 2.2, for details),

$$\hat{\phi}^j_{\mathrm{Bayes}} = \frac{n\hat{\phi}_j + 2\beta_0}{n - K + 2\alpha_0}. \tag{1}$$

To give an empirical Bayes solution, we estimate α_0 and β_0 from the marginal distribution of the $\hat{\phi}_j$ across all exons, making the assumption that the $\hat{\phi}_j$ are independent across exons. Critically, the approximate distribution of $\hat{\phi}_j$ is independent of the individual total counts, m_{ij}, per exon and sample. This means that under the Bayesian model above, the $\hat{\phi}_j$ are marginally identically distributed and we can use the joint marginal likelihood of $\hat{\phi}_j$ to find estimates $\hat{\alpha}_0$ and $\hat{\beta}_0$. Because the Inverse Gamma prior for ϕ is a conjugate prior, this results in an analytical expression for the marginal density of $\hat{\phi}_j$ which we use to estimate $\hat{\alpha}_0$ and $\hat{\beta}_0$ via ML estimation (see Supplemental text, Section 2.2, for details).

Substituting our estimates $\hat{\alpha}_0$ and $\hat{\beta}_0$ into our standard Bayesian point estimate in equation (1) gives the empirical Bayes estimate of ϕ_j. We call this estimation procedure Double exponential Empirical Bayes with application to Sequencing (DEB-Seq) and refer to the estimate as $\hat{\phi}^j_{\mathrm{DEB}}$.

[2] Recall that if $X \sim \mathrm{IG}(\alpha_0, \beta_0)$, this implies that $1/X \sim \Gamma(\alpha_0, \beta_0)$ where β_0 is the rate parameter.

4.2. *Empirical Bayes based on weighted likelihood.* Wang (2006) gives a general strategy for combining likelihoods by creating a (weighted) average of the log-likelihoods of all the experiments. The edgeR method of Robinson and Smyth (2007) adapts this idea to the gene-expression setting to make it gene-centric. Each gene j is given a separate weighted log-likelihood $\ell^j_{\mathrm{WL}}(\phi)$ which is the weighted sum of the individual gene's log-likelihood for gene j, $\ell^j(\phi)$ and a common log-likelihood $\ell_{\mathrm{CM}}(\phi)$, which is the average of the individual gene log-likelihoods over all genes,

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

Then for each gene j, we maximize $\ell^j_{\mathrm{WL}}(\phi)$ to obtain the shrunken estimate $\hat{\phi}_j$. The weight δ given to the common log-likelihood $\ell_{\mathrm{CM}}(\phi)$ is a tuning parameter that must be chosen. McCarthy, Chen and Smyth (2012) suggest that it be chosen so that it is proportional to sample size adjusted by degrees of freedom, with edgeR assigning a default value for δ equal to $\frac{20}{n-K}$.

In order to apply this to the dispersion parameter of a negative binomial, Robinson and Smyth (2007) use the conditional likelihood of $\hat{\phi}|\sum_{i=1}^{p} Y_i$ to define a conditional log-likelihood $\ell^j(\phi)$ which is independent of μ assuming all m_i are equal. This 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.

We can follow the same weighted likelihood strategy to create a weighted likelihood for a distribution from the double exponential family. Again, if we assume that $c(\mu_k, \phi, m) = 1$, we have that $\hat{\phi}_j$ is approximately distributed Gamma($\frac{n-K}{2}$, $\frac{2}{n}\phi$). Unlike the negative binomial distribution, the approximate likelihood of ϕ_j does not depend on the m_{ij}, eliminating the need to create pseudo-data in the implementation. Also, unlike the negative binomial distribution, the resulting estimate from maximizing the weighted log-likelihood with respect to the precision $1/\phi$ has an analytical solution given by

$$\hat{\phi}^j_{\mathrm{WL}}(\delta) = \frac{n}{n-K}\frac{\hat{\phi}_j + \delta\bar{\phi}}{1+\delta}, \tag{2}$$

where $\bar{\phi} = \frac{1}{p}\sum_j^p \hat{\phi}_j$ is the average of the $\hat{\phi}_j$ over all exons in the data (see Supplemental text, Section 2.3).

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 clear from comparing the (1) with (2) that the weighted likelihood estimate takes on the same form as that of the empirical Bayes estimate. Specifically,

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

This implies that the weighted likelihood method can be written as an empirical Bayes solution where the prior is parameterized by a single variable δ rather than the two parameters α_0 and β_0. (We are implicitly treating $\bar{\phi}$ as a fixed value, rather than explicitly conditioning on it, but $\bar{\phi}$ will normally be the average of thousands, if not tens-of-thousands, of exons.) Note that in the weighted likelihood approach, δ is assumed to be strictly positive, and $\bar{\phi}$ will similarly be positive, satisfying the assumptions for α_0 and β_0 to yield a true density.

This naturally suggests an estimator based on this alternative parameterization to fuse these two methods together. Using this parameterization for the Gamma prior, and maximizing the marginal joint likelihood of $\hat{\phi}_j|\delta$, gives us an estimate $\hat{\delta}$ and represents the amount of shrinkage that should be performed in the weighted likelihood method [for details concerning estimation of δ see Supplementary text, Section 2.3, Ruddy, Johnson and Purdom (2015a)]. We call the resulting dispersion estimates a Weighted-likelihood Empirical Bayes shrinkage with application to Sequencing (WEB-Seq), and denote them as $\hat{\phi}^j_{\mathrm{WEB}} = \hat{\phi}^j_{\mathrm{WL}}(\hat{\delta})$.

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 in simulations to violations of the model due to being more conservative (Section 5).

4.3. *Estimates of underdispersion near the boundary.* In initial simulations, estimated proportions lying on the boundary (i.e., Y_{ij}/m_{ij} either exactly 1 or 0) had a large and adverse effect on the false discovery rate (FDR) as the sample size increases [see Supplementary Figure S3, Ruddy, Johnson and Purdom (2015b)]. These exons are either never skipped or always skipped in one or more samples. This poor performance is particularly detrimental for the exon skipping application, since many exons may be rarely skipped or rarely expressed, particularly if introns are included in the analysis.

The poor behavior is due to the fact that the methods estimate under-dispersion for such boundary exons, leading to a large number of false discoveries. Ironically, 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 this is that in low sample sizes those exons whose true proportion lies near the boundary are more likely to have all observed proportions equal exactly 0 or 1; exons with proportions all 0 or 1 across *all* samples automatically get a p-value of one regardless of $\hat{\phi}$, and so they are removed from consideration by our method before the shrinkage is calculated. Larger sample sizes have an increased chance that at least one nonboundary sample will be observed, allowing the exon to remain in the analysis and have an effect on the FDR results.

One approach to address this issue could be to further filter exons to remove those with mean proportion close to the boundary. Another approach would be to not allow underdispersion by setting the dispersion in such cases to one, that is, that of a binomial distribution.

We obtained better results by defining an effective sample size, n_{eff}, the maximum of the number of nonboundary samples and $K + 1$, where K is the number of groups being compared. We use this n_{eff} to adjust the degrees of freedom, $n - K$, that appear in the Gamma prior, changing it to be $n_{\mathrm{eff}} - K$ instead. Note that this means that each exon has a slightly different effective sample size. A similar difference in effective sample size can also result when a sample has $m_{ij} = 0$, in which case the sample cannot be included in the analysis of that exon [see Supplementary text, Section 2.4, for details Ruddy, Johnson and Purdom (2015a)].

Using these effective degrees of freedom, our methods are less likely to erroneously estimate underdispersion on the boundary, though it remains technically possible to estimate underdispersion. With this adjustment, we report no underdispersion 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 underdispersed and thus given inflated significance. Furthermore, we estimate a continuous range of dispersion values ϕ for these boundary exons. We find this to be more natural than forcing them all to have no dispersion, which would be the case if we arbitrarily set those with underdispersion estimates to have $\hat{\phi} = 1$.

In Supplementary Table S1 [Ruddy, Johnson and Purdom (2015c)] we compare the proportion of exons that are affected by our adjustment for data simulated under either a double binomial or beta-binomial distribution, as well as the real data. We see that 78% of the exons are affected by these changes, with 43% having a large reduction of 5 or more. This highlights the importance of boundary control in the exon skipping problem. Interestingly, we see that the real data more closely follows the double binomial simulations in this respect, rather than the beta-binomial.

4.4. *Comparison of conditions.* After obtaining estimates of ϕ for each exon based on shrinkage across the exons, 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 $\eta(\mu)$, and we can test the significance of the contrast per exon. In our implementation, we focus on the common two-group comparison, though all of the methods carry through to the more general setting of contrasts of groups. In the setting of comparing two groups, we reparameterize so that $\beta_1 = \eta_2 - \eta_1$ is the value of the contrast defined by the difference of the two groups (the log-odds ratio for a double binomial distribution). We test the null hypothesis $H_0 : \beta_1 = 0$.

Let $\hat{\phi}^*$ be an estimator of ϕ under the full model with no constraints on β_1 (e.g., DEB-Seq or WEB-Seq, or the MLE $\hat{\phi}$). When testing the null hypothesis $\beta_1 = 0$, the standard GLM approach defines the likelihood ratio statistic as

$$W_{\hat{\phi}^*} = \log \frac{L(\hat{\beta}_0^{H_0}, \beta_1 = 0; \hat{\phi}^*)}{L(\hat{\beta}_0, \hat{\beta}_1; \hat{\phi}^*)} = \frac{S_{H_0} - S}{\hat{\phi}^*},$$

where S and S_{H_0} are one-half of the sum of the deviance residuals of the full or null model, respectively [i.e., $1/2 \sum_k \sum_i D(Z_i, \hat{\mu}_k)$ for $\hat{\mu}_k$ estimated under the full or null model, respectively].

Because our shrinkage methods give analytical solutions for the estimates of ϕ, we can easily compare the effect of using the shrinkage estimator $\hat{\phi}^j_{\mathrm{WEB}}$ instead of the original MLE $\hat{\phi}_j$ on the statistic W. $W_{\hat{\phi}^j_{\mathrm{WEB}}}$ will be smaller than $W_{\hat{\phi}_j}$ if $\hat{\phi}_j$ is less than $\bar{\phi}$, the mean of the $\hat{\phi}_j$ across all exons. Therefore, those exons with small estimates of variability will become less significant after shrinkage. If the distribution of $\hat{\phi}_j$ *is* roughly gamma, this implies that the majority of test statistics are reduced in significance since the gamma distribution is skewed right and therefore the median is less than the mean.

Asymptotically, $W_{\hat{\phi}}$ should follow a F distribution with $(K-1)$ and $(n-K)$ degrees of freedom [Jørgensen (1997)]. We find that the shrinkage methods result in less variability in the estimate of $\hat{\phi}$, with the result that the likelihood ratio statistic more closely follows a χ^2_{n-K} distribution than the standard F distribution for unshrunken estimates.

Based on our simulations, we also find that this approximation is poor for small sample sizes, for example, when the size of each group is five or less, leading to poor control of Type I errors. Instead we propose an alternative statistic, which re-estimates ϕ under the null and alternative, that has much better performance in small sample sizes,

$$W_{\hat{\phi}^*, \hat{\phi}^*_{H_0}} = \log \frac{L(\hat{\beta}_0^{H_0}, \beta_1 = 0; \hat{\phi}^*_{H_0})}{L(\hat{\beta}_0, \hat{\beta}_1; \hat{\phi}^*)} = \frac{S_{H_0}}{\hat{\phi}^*_{H_0}} - \frac{S}{\hat{\phi}^*} + \frac{n}{2} \log\left(\frac{\hat{\phi}^*_{H_0}}{\hat{\phi}^*}\right).$$

This means the likelihoods are not strictly nested, so that $W_{\hat{\phi}, \hat{\phi}_{H_0}}$ can technically take on negative values. Here, $\hat{\phi}^*_{H_0}$ and $\hat{\phi}^*$ are estimates of ϕ based on the null $(K = 1)$ and the full $(K = 2)$ models, respectively. For our shrinkage estimates of ϕ, this means that the value of K and the unshrunken ML estimate $\hat{\phi}_j$ in equations (1) and (2) change. The estimates $\hat{\alpha}_0$, $\hat{\beta}_0$, $\hat{\delta}$ and $\bar{\phi}$ are all based on the distribution of the unshrunken estimates of ϕ, and so we accordingly also re-estimate their values under either the assumption that $K = 1$ or $K = 2$.

5. Evaluation of methods.

We evaluate WEB-Seq and DEB-Seq on simulated and real data sets. Our evaluations consider the two-group comparison in everything that follows, and we will denote n_G to be the number of samples *per group*, so that $n = 2n_G$.

We also compare our methods to other existing methods that provide shrinkage of the dispersion parameter when comparing proportions across groups. 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)], mentioned above, was developed for detecting differences in PSI and assumes a uniform prior for the proportion parameter of a binomial and further adds a correlation between the two conditions being compared; this parameter is assumed shared by all the exons and thus provides implicit shrinkage. BBSeq [Zhou, Xia and Wright (2011)] is a method developed for gene expression studies that shrinks the dispersion parameter of a beta-binomial model. Their shrinkage method fits a cubic polynomial to the independently estimated, logit-transformed dispersion parameters as a function of the fitted values of the observed data. The method of Feng, Conneely and Wu (2014) provided in the DSS package was developed for DNA methylation data and also provides shrinkage of the dispersion parameter of a beta-binomial distribution. They do so by fitting an empirical Bayes model that assumes the dispersion parameter has a log-normal prior distribution and their computations are based on method of moments estimators for the beta-binomial. For convenience, we will refer to this as the DSS method, though DSS actually refers to the corresponding gene expression technique that the same authors developed earlier in Wu, Wang and Wu (2013).

5.1. *Description of the simulation.* We simulated exon counts 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 3.4, for details [Ruddy, Johnson and Purdom (2015a)]. We generated data under a double binomial distribution and also a beta-binomial distribution for evaluation of the robustness of our techniques which were developed assuming the data come from a double binomial distribution. We randomly selected either 1% or 10% of the exons to show differential usage between the groups; for these non-null exons, we chose the treatment effect, β_1, uniformly from the union of $[-3, -0.5]$ and $[0.5, 3]$, to account for both decreased and increased exon usage. Otherwise, we gave exons a treatment effect of 0, comprising our null set of exons. For each simulation, we simulated 85,373 exons and applied a basic filtering process to remove exons with proportions all equal to 1 or all equal to 0 across the samples (these result in p-values of 1).

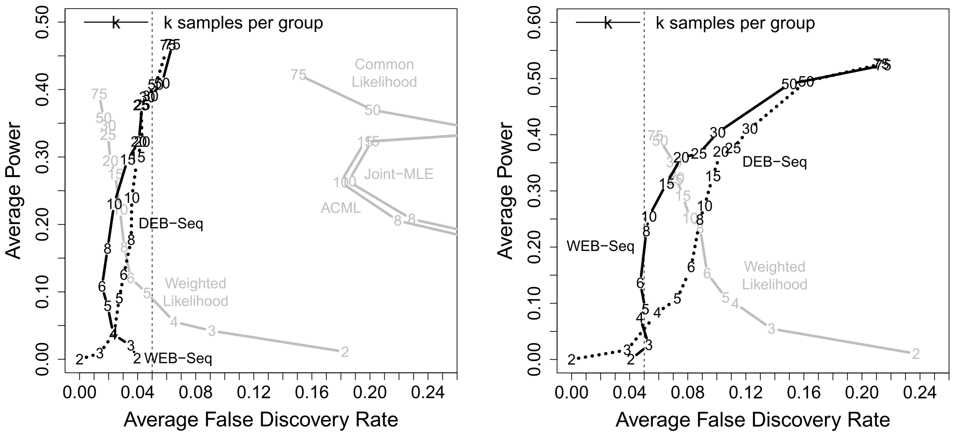
We used the simulated data to evaluate the methods developed above: (1) the empirical Bayes method with a single parameter prior (WEB-Seq), (2) the general two-parameter empirical Bayes method for the prior parameters (DEB-Seq), and (3) the weighted likelihood method with δ fixed to be equal to the default value implemented in edgeR ($\delta = \frac{20}{n-K}$). In addition to our dispersion-shrinkage methods, we implemented the shrinkage methods of BBSeq, EB2 and DSS, briefly described above. The MATS method does not take as input inclusion and exclusion count matrices, but rather creates its 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 implemented using the `betabin` function from the `aod` package in R, and maximum likelihood estimation based on an approximate double binomial distribution where the normalizing constant is set to 1. The quasi-binomial GLM and the double binomial MLE are closely connected, as described in Section 3.1, and are both nonshrinkage 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 nonshrinkage counterpart of the BBSeq and DSS methods that rely on the beta-binomial distribution.

For each method, we implemented the estimation routines and then adjusted the p-values to control the FDR to a 0.05 level using the standard Benjamini–Hochberg FDR procedure [Benjamini and Hochberg (1995)] as 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.

5.2. *Comparison of double binomial methods on simulated data.* We first compare the performance of WEB-Seq and DEB-Seq to other methods that also rely on the double binomial distribution, particularly those without shrinkage. As hoped, the shrinkage methods improve upon the estimation of the dispersion parameter, reducing the mean squared error (MSE) significantly from the unshrunken versions of the double binomial [Supplementary Figure S5, Ruddy, Johnson and Purdom (2015b)]. DEB-Seq had the smallest MSE, followed by WEB-Seq, and both were a significant reduction compared to double binomial estimation methods with no shrinkage.

We compare their ability to control the false discovery rate across a range of sample sizes for a fixed FDR cutoff (Figure 2) and see that WEB-Seq and DEB-Seq were also superior in control of FDR across a range of samples sizes. For data simulated as double binomial, they both control the FDR well, while unshrunken methods have high rates of FDR compared to the target 5%. Weighted likelihood shrinkage for the double binomial with a predetermined tuning parameter (based on edgeR default) is erratic in its control of FDR for small sample sizes ($n_G \leq 5$), but then adequately controls FDR. However, the predetermined tuning method becomes over-conservative for large sample sizes, and the result is a large drop in power for weighted likelihood for large sample sizes.



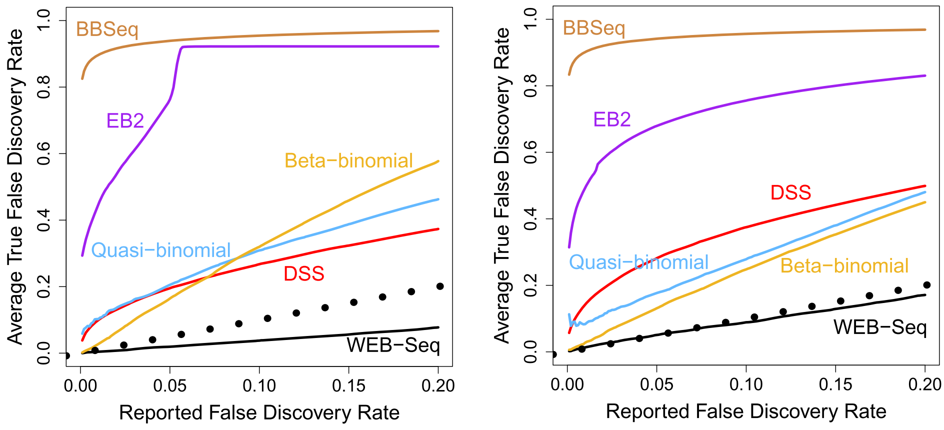
(a) Double Binomial Simulation (b) Beta-binomial Simulation

FIG. 2. Double binomial methods compared across different sample sizes for a fixed 5% FDR cutoff. *We plot the average power* (y*-axis*) *against the FDR* (x*-axis*) *on simulated data for various sample sizes based on* p*-values adjusted to provide a* 5% *FDR level. The calculated values of average power and FDR are plotted with the sample size* n_G *as the plotting character. The results for a single method across different sample sizes are connected by a line. The* 5% *FDR boundary is given by the dotted vertical line. The results are based on* 100 *simulations under* (a) *a double binomial distribution and* (b) *a beta-binomial distribution*; *the simulations set* 1% *of exons as differentially used. The plot compares only methods that assume the data follow a double binomial distribution, with different methods being distinguished by different line types and grey scales. The joint MLE, ACML and common likelihood are not shown in the beta-binomial simulation because their FDR values were beyond the limits of the plot.*

To evaluate the robustness of the methods, we consider data that do not follow the double binomial distribution, but rather the beta-binomial distribution. Again, in small sample sizes ($n_G \leq 4$) both WEB-Seq and DEB-Seq control the FDR accurately. For moderate sample sizes ($n_G \leq 10$), the more conservative WEB-Seq maintains control of the FDR; DEB-Seq still has greater power, but has an increase of FDR to about 10% for these moderate sample sizes. Large sample sizes ($n_G \geq 10$) show an underlying bias, probably because the p-values were calculated under the wrong model and, as a result, the FDR of both WEB-Seq and DEB-Seq starts growing well beyond the 5% target.

5.3. *Comparison to other methods on simulated data.*

FDR control. In Figure 3, we compare how WEB-Seq controls the FDR in small sample sizes ($n_G = 5$) compared to other existing methods. WEB-Seq shows superb control of the FDR, while all of the other methods—both those that use shrinkage and those that do not—have FDR rates far beyond what their adjusted p-values would indicate. This pattern holds for different distributions of the data



(a) Double Binomial Simulation

(b) Beta-binomial Simulation

FIG. 3. Control of FDR. *We plot the true FDR level against the reported FDR for different methods, with methods indicated by different colors. The reported FDR refers to the cutoff used for calling exons significant based on FDR-adjusted p-values. The true FDR is calculated for those exons called significant by knowing which exons are truly false discoveries in our simulations. The dotted line is the $y = x$ line. We simulate the data under* (a) *a double binomial distribution and* (b) *a beta-binomial distribution. Here the percent of non-null exons was* 1% *and $n_G = 5$. For* 10% *non-null and other sample sizes, see Supplementary Figures* S7–S14 [*Ruddy, Johnson and Purdom* (2015b)].

(double binomial or beta-binomial), for different choices of sample sizes, for unbalanced sample sizes, and for different choices of the percent of non-null exons [Supplementary Figures S7–S12, Ruddy, Johnson and Purdom (2015b)].

Among the alternative methods, the beta-binomial MLE with no shrinkage performs the best in controlling the FDR under both distributions, but still has an FDR much larger in small sample sizes than that indicated by their adjusted p-values, even when the data is distributed according to a beta-binomial distribution (e.g., 10–15% FDR instead of the target rate of 5%). The performance is even worse for double binomial distributed data, indicating a lack of robustness to the modeling assumptions. Quasi-binomial (also with no shrinkage) appears to perform similarly to the beta-binomial MLE in Figure 3, but evaluations of other sample sizes [Supplementary Figures S6–S10, Ruddy, Johnson and Purdom (2015b)] shows that its FDR performance is erratic and can be much larger than its target value; similarly, the double binomial GLM (not plotted) fails to control the FDR at any sample size we explored and converges to an FDR at around 40%.

The alternative methods we consider that do perform shrinkage (EB2, BBSeq and DSS) also rely on beta-binomial dispersion models, but do quite badly in FDR control even for beta-binomial data. The true FDR rates for BBSeq and EB2 start at 0.96 and 0.98 for small sample sizes ($n_G = 3$) and remain poor even for $n_G = 75$ with FDR of 0.27 and 0.41, respectively. DSS is slightly better since it eventually controls the FDR with larger n, but it only does so starting at sample sizes that are quite large for genomic studies ($n_G > 20$); for $n_G = 3$ and $n_G = 5$ per group, itself true FDR rates are 0.53 and 0.28, respectively.

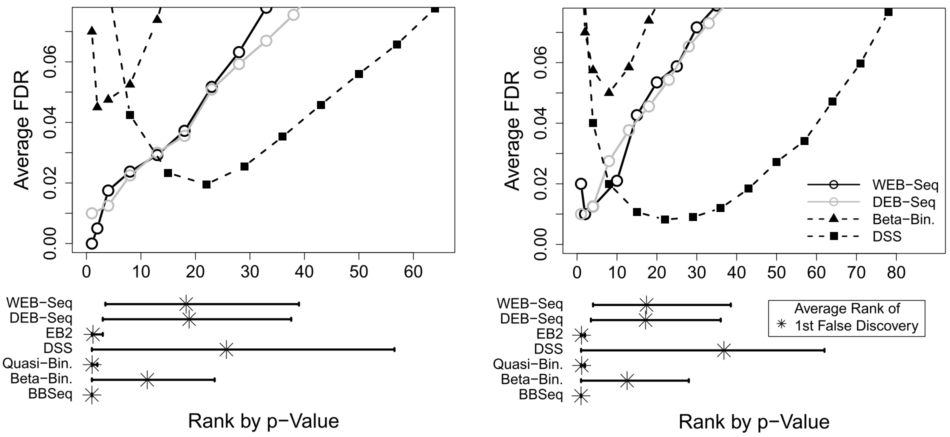
In contrast, WEB-Seq controls the FDR at the desired level for the double binomial data and beta-binomial data in small sample sizes, and only shows poor control of the FDR for fairly large sample sizes ($n_G > 10$), and even then only for beta-binomial distributed data.

Due to multiple testing corrections, the comparative performance of the various methods hinges on the distributional behavior of their test statistics far in the tails of the distribution. At standard levels of individual (per exon) Type I error control (e.g., 0.01–0.05), the p-values of all of the methods do a reasonable job of controlling the Type I error. However, after controlling for multiple testing, the effective raw p-value cutoff for significance at a target FDR rate of 5% is around 0.0001 if only 1% of the exons are non-null (and 0.001 if 10% are non-null). In this tail of the distribution, the test statistics can perform quite differently with respect to Type I error control and their resulting power. In our simulations, WEB-Seq controls Type I error even far into the tails of the distribution, unlike any other method (Supplemental Figures S25–S28), indicating that the performance of the methods in the tail of the distribution (and not the p-value adjustment method) leads to this discrepancy.

All of the specific numbers given above assume 1% of exons are differentially used; the same lack of control holds for 10%, though with different specific values. We would also note that for exon analysis, the number of exons evaluated can be quite high compared to gene expression analysis (easily 50,000–100,000 exons), and in many studies we would not expect the percent of exons that are skipped *differentially* between groups to be very large: even 1% translates to hundreds of differentially spliced exons. With an even smaller percentage of true non-nulls, which could be common in studies with subtle effects, control of the FDR will become even worse for these other methods.

Power. For comparison of the power of the methods, we must similarly focus on the power exhibited for small levels of Type I error control to be relevant for multiple testing corrections. Traditional ROC curves in this range [Supplementary Figures S15–S18, Ruddy, Johnson and Purdom (2015b)] show that WEB-Seq results in greater power than all the other methods except DSS when the data comes from a double binomial distribution; for data distributed under a beta-binomial distribution, it still has slightly greater or equivalent power to all the methods except DSS. This again illustrates the robustness of WEB-Seq to the modeling assumptions. DSS is the only method that has clearly improved power, mainly in small samples sizes (e.g., power of 3% for WEB-Seq versus 8% for DSS when $n_G = 3$); however, as we have seen, DSS also shows very poor control of the FDR for those sample sizes.

Many biological studies focus on the top performing exons for validation and followup analysis, especially when large numbers of significant results are found.



(a) Double Binomial Simulation

(b) Beta-binomial Simulation

FIG. 4. False discoveries by rank. *Plotted is the average proportion of false discoveries (y-axis) in the top x exons (x-axis) for a* 3 *versus* 3 *simulation with* 1% *of exons alternatively spliced. All methods but WEB-Seq, DEB-Seq, DSS and beta-binomial had false discovery rates too large to be shown on the range of values used for the y-axis in this plot. The horizontal lines at the bottom give an indication of the rank of just the* first *false discovery for each method (where larger values mean better performance); the horizontal lines indicate the* 2.5% *to the* 97.5% *percentiles of the rank of the first false discovery across the* 100 *simulations. The average rank of the first false discovery across simulations is marked by an asterisk.*

We find that the WEB-Seq method provides p-values that do well at prioritizing the truly non-null exons in small sample sizes. In Figure 4, we plot the average proportion of false discoveries in the top-ranked exons for simulations with $n_G = 3$ samples per group. We see that for this small sample size, the alternative methods have a much higher proportion of false positives in the top-ranked exons compared to WEB-Seq, regardless of whether the data is distributed beta-binomial or double binomial. Beta-binomial MLE performs clearly worse in rankings for small sample sizes, even when the data is distributed according to a beta-binomial. For slightly larger sample sizes (e.g., $n_G = 5$), DSS outperforms our method in ranking the exons, but again has poor control of the FDR. For much larger sample sizes (e.g., $n_G = 10$), the difference between the DSS, beta-binomial and WEB-Seq methods is minimal and they all perform better than any of the other competing methods [Supplementary Figures S19–S22, Ruddy, Johnson and Purdom (2015b)].

Because we have exact analytical solutions for our estimators of WEB-Seq, the calculation of our shrinkage parameters is also quite fast, so that WEB-Seq can be run for any sized experiment on a single core, personal laptop in under a minute. Several of the other methods that we compared require hours of computation time [see Supplementary text, Section 3.2, Ruddy, Johnson and Purdom (2015a)].

In summary, WEB-Seq is much superior in giving accurate FDR adjusted p-values across the ranges of sample sizes that correspond to those frequently seen in practice, while every other method performs poorly, and often dramatically so. WEB-Seq also shows high power compared to the other methods and performs well at prioritizing exons. For some ranges of sample sizes the DSS method has more power than WEB-Seq, but at those same sample sizes, DSS has highly inaccurate p-values for assessing significance. The main competitor in small sample sizes when considering both FDR control and power is the beta-binomial MLE with no shrinkage, but it still has poor FDR control, less power and worse rankings of the exons in small sample sizes—even when the data is actually simulated from the beta-binomial distribution.

5.4. *Comparison of methods on TCGA 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 large 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)]. In fact, the differences in tissue types is a rather extreme example since in this case we do expect a large number of significant exons. See Supplementary text, Section 3.5, for details about the processing of the raw BAM files [Ruddy, Johnson and Purdom (2015a)].

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 differential alternative splicing in the different tumors may exist even though they are the same tissue type, but since these samples are randomly assigned to the two groups, this is unlikely to be a significant factor. We ran the double binomial methods, as well as the other methods on the TCGA data sets (Table 1). MATS could only be run in the null setting, as the stomach and ovarian samples were of different read lengths and the MATS software did not support this.

We compare the proportions of exons called significant in the null setting across methods based on FDR adjusted p-values. Note that this is not a measure of FDR or traditional Type I error; in fact, if no significant exons exist, *any* discoveries in an all-null setting would technically imply that the FDR is 1. For comparison, if 10% of the exons were truly non-null and the method had 100% power, the percentage of false positives would have to be 0.6% to get an FDR at the target level 5%; 1% of exons non-null would require the percentage of false positives to be 0.05%. A 3–7% false positive rate would then mean a minimum FDR of 21–38%. In practice, the false positive rate would need to be *much* lower since no method has anywhere near 100% power (in fact, the power is quite low for these small sample sizes).

Focusing on just our double binomial-based shrinkage methods (WEB-Seq and DEB-Seq), both call essentially 0% of exons significant in the null case [see Supplementary Table S2, Ruddy, Johnson and Purdom (2015c)]. Examining individual simulations rather than the average of 100 simulations shows that in the majority

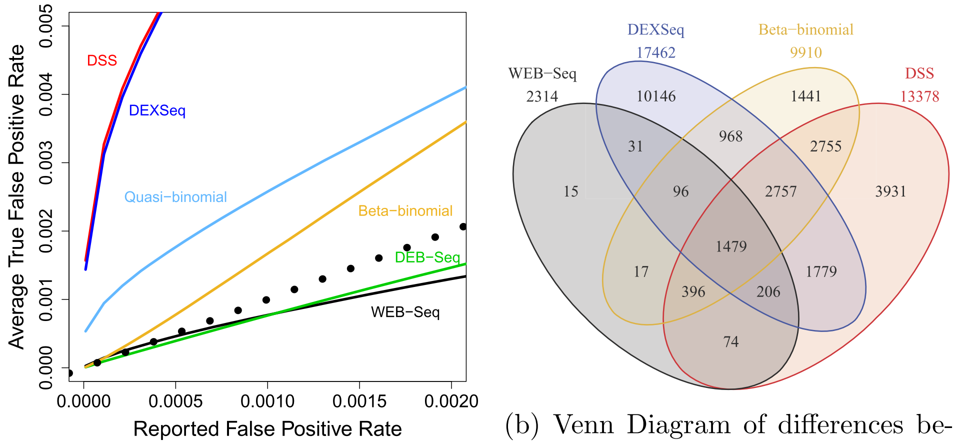
TABLE 1

Comparison to alternative methods. *Shown in the table below are the average percentage of exons called significant based on FDR corrected p-values from the Tissue Data under* 100 *simulations of 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, *k 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 any sample size. Furthermore, MATS was run on a single random sample because of the time involved in processing a single run of the data (see Supplemental text, Section* 3.2). *See Supplemental Table* S4 *for the precise number of exons called and the results from methods not shown here*

Sample size	**DEXSeq**		**EB2**		**BBSeq**		**MATS**	**DSS**		**Beta-bin.**		**WEB-Seq**	
	Real	**Null**	**Real**	**Null**	**Real**	**Null**	**Null**	**Real**	**Null**	**Real**	**Null**	**Real**	**Null**
2 vs 2	13.5	1.46	11.08	7.47	18.78	6.56	3.45	10.43	1.31	5.06	0.00	0.89	0.00
3 vs 3	17.17	0.62	11.56	7.20	24.95	7.37	1.77	12.19	0.58	9.70	0.00	2.82	0.00
4 vs 4	21.77	0.12	11.94	6.85	28.90	6.20	2.45	14.50	0.19	13.73	0.00	6.73	0.00
5 vs 5	26.63	0.08	12.67	6.54	31.28	6.02	2.74	17.37	0.16	17.73	0.00	12.16	0.00
6 vs 6	30.00	0.11	13.27	6.26	32.62	5.73	3.93	19.44	0.18	20.41	0.01	15.94	0.00
7 vs 7	33.86	0.06	14.08	6.00	33.82	4.98	3.39	22.03	0.14	23.42	0.00	20.26	0.00

of the simulation in every sample size, exactly zero exons are called significant. As in the simulations, WEB-Seq has slightly less "power" than DEB-Seq in that it makes fewer significant calls under the real setting where we compare the different tissue types.

Turning to the alternative methods, the average false positive rates (based on FDR corrected p-values) given the null analysis in Table 1 suggest BBSeq and EB2's poor control of the FDR in the simulated data is echoed in the real data. Across the samples sizes, EB2 finds roughly 7% of the exons significant and BBSeq finds 4–6% significant in the null setting. DSS shows better performance in the null analysis with less than 1% false calls (except for $n_G = 2$); but in the exon setting with many exons to evaluate, this still ranges from around 700 exons ($n_G = 3$) to 170 exons ($n_G = 7$) incorrectly called significant compared to only 4 and 1 incorrect exon calls for WEB-Seq in those same sample sizes, respectively [see Supplementary Tables S4 and S3, Ruddy, Johnson and Purdom (2015c)]. 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 1557 exons called significant), while WEB-Seq makes at most one call for any given sample size. These results, when roughly translated to FDR rates assuming 1% or 10% non-null exons, indicate that the lack of control of FDR shown in our simulations appears to be supported by implementation on the real data.



(a) Type I error control in null setting

(b) Venn Diagram of differences between tissues

FIG. 5. Evaluation on TCGA data, $n_G = 3$. (a) *Type* I *error for TCGA "null" setting where we plot the percent of exons found significant in the null setting as a function of the* p*-value cutoff based on uncorrected* p*-values*. (b) *We show Venn diagrams of the analysis of differences between tissues for WEB-Seq* (*black*), *DEXSeq* (*blue*), *Beta-binomial MLE* (*yellow*) *and DSS* (*red*). *DEXSeq is limited to only those exons also found to have skipping events and so also considered by the other methods. Numbers in the Venn diagrams are based on overlapping counts averaged over the* 100 *simulations. See Supplemental Figure S*29 *for* $n_G = 6$.

The only other method that gives comparable results to WEB-Seq is the beta-binomial MLE method. However, since the needed false positive rate based on adjusted p-values must be so small in many real settings, it is difficult to compare these two on just these evaluations. Instead, in the null setting we directly compare the Type I error rate based on the uncorrected p-values as we vary the cutoff (see Figure 5(a) and Supplementary Figure S29); we see that the performance directly echoes that seen in the simulations (Supplementary Figures S25–S28) [Ruddy, Johnson and Purdom (2015b)]. Only WEB-Seq (and DEB-Seq) controls Type I error in the tails of their distributions for small sample sizes. As in the simulations, beta-binomial is the next best, but still has an increase in its Type I error comparable to that of the simulations, which in simulations resulted in FDR rates of around 15% rather than the target 5%. All of these analyses strongly suggest that the problems we see in controlling the FDR in simulations are real and will appear in analyzing real data sets.

For the comparison of two different tissue types, we see many more calls made by the other methods compared to WEB-Seq. Given the problems these methods have in obtaining accurate FDR control in both the null setting and our simulations, we expect that some portion of these additional calls in the real setting are due to a much higher level of false discoveries than reported. We compare the overlap in these calls [Figures 5(b), 6 and Supplementary Figure S29, Ruddy, Johnson and Purdom (2015b)], and we see that in low sample sizes almost all of the calls made by WEB-Seq are also found significant by another method. If we look at the overlap of WEB-Seq with the six alternative methods we considered here, we see that 85% of WEB-Seq's calls are supported by at least *four* other methods for $n_G = 3$. Beta-binomial MLE is the next best candidate, yet only has 20% of its calls so strongly supported, in spite of the fact that two of the other methods (DSS and BBSeq) are also based on the beta-binomial distribution.

5.5. *Comparison to relative exon usage.* We made a further comparison of the performance of our method to another popular method of finding differential alternative splicing in exons, DEXSeq [Anders, Reyes and Huber (2012)]. The DEXSeq method relies on the relative exon usage framework described above in Section 2.3 and *only* uses exon counts, without using information about how the junctions skip exons. Therefore, DEXSeq is not just an alternative statistical method, but also uses a fundamentally different summarization of the mRNA-Seq data as compared to the inclusion–exclusion setting for which our methods are developed. For these reasons, we must be cautious in comparing between the two.

As we discussed in Section 2.3, relative exon usage evaluates all the exons, not just those showing skipping events. We concentrate on comparing the performance of DEXSeq for just those 125,398 exons that show skipping in at least one of the 30 samples; to do this, we ran DEXSeq on all 412,002 expressed exons, as required by the algorithm, then filtered down to the 125,398 exons with skipping, and calculated adjusted p-values based only on these filtered exons. For this set of exons the false positive rate of DEXSeq is similar to that of DSS, particularly for the smaller sample sizes (Table 1), as is its control of the Type I error rate on the "null" setting [Figure 5(a) and Supplementary Figure S29, Ruddy, Johnson and Purdom (2015b)]. This suggests that DEXSeq will have similar problems as DSS in having much higher rates of false discoveries than indicated by the adjusted p-values.

DEXSeq clearly calls more exons significant in the real setting than any of the methods based on junction counts, even when limited to the same set of exons. This could be a sign of increased power when using the exon counts, as mentioned above. When we look at the overlap of DEXSeq with other methods in Figures 5(b) and 6, we see that this increase constitutes tens of thousands of additional exons called significant, and 30–40% of the calls of DEXSeq are not supported by any of the junction count methods, regardless of sample size (Figure 6). Furthermore, as we explained in Section 2.3 (and has been noted by the authors of DEXSeq), relative exon usage does not always reliably indicate the right exon within the gene that has differential usage. The additional calls may also be a reflection of this problem, not only increased power.

We consider DEXSeq's performance more broadly and consider their calls on all exons, even those without skipping fragments. About 12% of constitutive exons are called significant by DEXSeq (Supplemental Table S6). This is roughly their

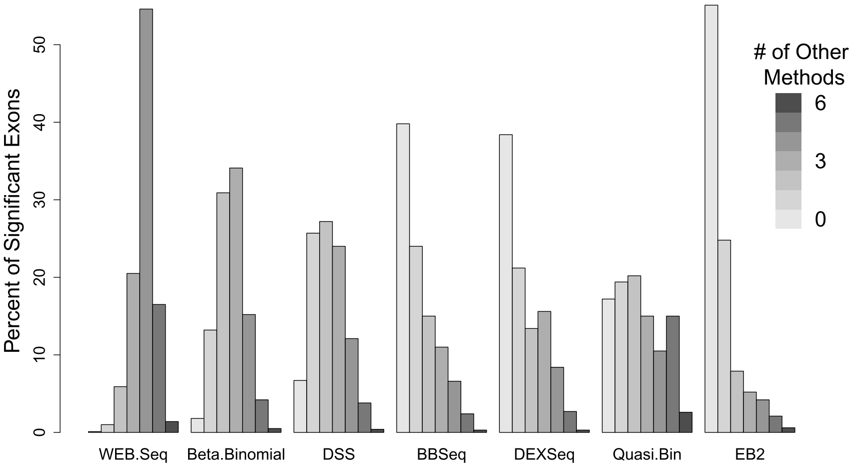


FIG. 6. Percentage overlap with calls by other methods, $n_G = 3$. *For each method, we determine which significant exons were also called significant by other methods. For each method, the above barplot gives the percentage of significant exons for that method that were called significant by other methods, broken down by how many other methods called the exons significant. The gray scale of each bar indicates how many* other *methods called the exons significant (ranging from* 0–6). *For more detailed plots of how the methods overlapped or for results on* $n_G = 6$, *see Supplementary Figures* S30–S31 [*Ruddy, Johnson and Purdom* (2015b)].

total representation in the data so DEXSeq does not appear to be preferentially finding exons annotated to be alternatively spliced. If we instead compare only exons that are the sole exon called significant in their gene, following the theory they are more likely to be targeting the correct exon, even a larger percentage are annotated as constitutive *and*, furthermore, have no fragments skipping them in the data for *any* of the 30 samples (18–46%, Supplemental Table S6). In comparison, in WEB-Seq, 0.2% of its significant exons (or 76 exons) are annotated as constitutive and all of them, by definition, have fragments skipping them to at least justify the call of significance.

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. We compare the density of the log-fold-change of the odds-ratio of skipping an exon for the significant calls made by both methods [Supplementary Figure S24, Ruddy, Johnson and Purdom (2015b)]. WEB-Seq clearly has a much stronger tendency to find exons with large differences in the skipping proportion, which is not surprising given that is the basis of its test statistic, unlike DEXSeq. More striking is that DEXSeq has 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 found by DEXSeq, 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 [Supplementary Figure S23c, Ruddy, Johnson and Purdom (2015b)].

Ultimately, we find the inclusion–exclusion paradigm 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.

6. Discussion.

We have developed a novel method for 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 overdispersion to a binomial, which is widely used and numerically robust. By making use of the distributional form of Efron (1986), we have shown a simple formulation of the approximate 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 nonstandard 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.

In comparison to other methods for analyzing exon-skipping events, we showed in both simulated and real data that our method is the only one that can accurately control the FDR in the sample sizes that are commonly seen in genomic studies (often less than 5 samples per treatment group), with the other methods having very large false discovery rates compared to their reported rate. We also show that our method has good power and the ability to prioritize truly significant genes.

For detecting differential alternative splicing, we have discussed that alternative summaries of the data require different statistical techniques than those presented here. The PSI statistic may not be the most appropriate for every setting. We compared directly to one such alternative approach—using only exon counts without using the information in the junction fragments—and we illustrated that reliance on junction fragments naturally filters the problem to those exons most likely to be differentially used. Another alternative approach relies on estimating the expression levels of individual isoforms, and this may give more insight into alternative splicing, particularly when a great deal of information about the transcriptome is known. However, in our experience many situations arise where researchers find themselves without a well-constructed annotation of the transcriptome and would have to rely on de novo methods to construct genes and/or transcripts. 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 [see also Anders, Reyes and Huber (2012) for a discussion of isoform versus exon analysis]. In contrast, inclusion–exclusion counts rely on detection of exons and splice sites, which are much simpler problems. In short, inclusion–exclusion counts 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.

While our shrinkage method is quite general, we have focused on our motivating 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. Other genomic settings also require the comparison of a large number of proportions between groups, for example, in the setting of comparing allele frequencies or differential methylation, and it is possible that the performance would differ in those settings due to differences in the properties of the data.

Because of our interest in exon inclusion probabilities, our evaluation of the shrinkage method was based on the binomial distribution (our initial “naive” distribution), but our entire methodological development is general and can be applied to any distribution from the exponential family. While every type of data should have careful development for its unique properties, it is useful to have a single framework that can be the starting point for so many settings. Possible examples could be that of analyzing differential gene expression data (based on the Poisson distribution) or differential proportions of isoforms (based on multinomial distribution). Indeed, we tested a Poisson-based version of our WEB-Seq on gene expression data along with fourteen other gene expression techniques from the literature and found that on simulated data our method performed well compared to the other methods across the range of distributions we tried. WEB-Seq ranked the significant genes better than most other methods (Supplemental Figures S32, S33). Its control of the FDR was also better than many common methods, particularly in small sample sizes (Supplemental Figure S34), though not giving accurate FDR control like we presented in the exon setting (with true FDR for $n_G = 5$ of around 7% rather than target of 5%). Furthermore, these conclusions hold true with data simulated under the negative binomial distribution which is the distribution for which most of the other methods (except ours) were developed. The only methods that did equivalent or better in *both* FDR control and power were voom [Law et al. (2014)] and baySeq [Hardcastle and Kelly (2010)]; notably, the popular edgeR [Robinson, Mccarthy and Smyth (2010)] and DESeq [Anders and Huber (2010)] methods did not do as well in either power or control of FDR. We find this particularly encouraging for the general use of our shrinkage method in other settings, since we did not change or adjust the method in any way for the gene expression setting, other than switching the choice of distribution within the exponential family.

In summary, our method gives reliable and robust improvement to the analysis of exon splicing, which is a straightforward but important approach to analyzing the complicated structure of alternative splicing. Furthermore, because the shrinkage ideas apply generally to an exponential family of distributions and have close links to the common GLM approach for analyzing data, it has the potential to be relevant for other applied problems.

Acknowledgments. We wish to thank the anonymous Associate Editor and reviewers that offered suggestions that substantially improved the paper. 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/.

SUPPLEMENTARY MATERIAL

Supplement A: Supplemental text (DOI: 10.1214/15-AOAS871SUPPA; .pdf). Supplemental text to accompany manuscript.

Supplement B: Supplemental tables (DOI: 10.1214/15-AOAS871SUPPB; .pdf). Supplemental tables to accompany manuscript.

Supplement C: Supplemental figures (DOI: 10.1214/15-AOAS871SUPPC; .pdf). Supplemental figures to accompany manuscript.

REFERENCES

ANDERS, S. and HUBER, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* **11** 106.

ANDERS, S., REYES, A. and HUBER, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22** 2008–2017.

BARBOSA-MORAIS, N. L., IRIMIA, M., PAN, Q., XIONG, H. Y., GUEROUSSOV, S., LEE, L. J., SLOBODENIUC, V., KUTTER, C., WATT, S., COLAK, R., KIM, T., MISQUITTA-ALI, C. M., WILSON, M. D., KIM, P. M., ODOM, D. T., FREY, B. J. and BLENCOWE, B. J. (2012). The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338** 1587–1593.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BOURGON, R., GENTLEMAN, R. and HUBER, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proc. Natl. Acad. Sci. USA* **107** 9546–9551.

BROOKS, A. N., YANG, L., DUFF, M. O., HANSEN, K. D., PARK, J. W., DUDOIT, S., BRENNER, S. E. and GRAVELEY, B. R. (2011). Conservation of an RNA regulatory map between drosophila and mammals. *Genome Res.* **21** 193–202.

BROOKS, A. N., CHOI, P. S., DE WAAL, L., SHARIFNIA, T., IMIELINSKI, M., SAKSENA, G., PEDAMALLU, C. S., SIVACHENKO, A., ROSENBERG, M., CHMIELECKI, J., LAWRENCE, M. S., DELUCA, D. S., GETZ, G. and MEYERSON, M. (2014). A pan-cancer analysis of transcriptome changes associated with somatic mutations in U2AF1 reveals commonly altered splicing events. *PLoS ONE* **9** e87361.

CANCER GENOME ATLAS RESEARCH NETWORK (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* **474** 609–615.

DENOEUD, F., AURY, J.-M., SILVA, C. D., NOEL, B., ROGIER, O., DELLEDONNE, M., MORGANTE, M., VALLE, G., WINCKER, P., SCARPELLI, C., JAILLON, O. and ARTIGUENAVE, F. (2008). Annotating genomes with massive-scale RNA sequencing. *Genome Biol.* **9** R175.

DOLZHENKO, E. and SMITH, A. D. (2014). Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC Bioinformatics* **15** 215.

EFRON, B. (1986). Double exponential families and their use in generalized linear regression. *J. Amer. Statist. Assoc.* **81** 709–721. MR0860505

FENG, H., CONNEELY, K. N. and WU, H. (2014). A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res.* **42** e69–e69.

GUTTMAN, M., GARBER, M., LEVIN, J. Z., DONAGHEY, J., ROBINSON, J., ADICONIS, X., FAN, L., KOZIOL, M. J., GNIRKE, A., NUSBAUM, C., RINN, J. L., LANDER, E. S. and REGEV, A. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* **28** 503–510.

HARDCASTLE, T. J. and KELLY, K. A. (2010). BaySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* **11** 422.

HARDCASTLE, T. J. and KELLY, K. A. (2013). Empirical Bayesian analysis of paired high-throughput sequencing data with a beta-binomial distribution. *BMC Bioinformatics* **14** 135.

HU, Y., HUANG, Y., DU, Y., ORELLANA, C. F., SINGH, D., JOHNSON, A. R., MONROY, A., KUAN, P. F., HAMMOND, S. M., MAKOWSKI, L., RANDELL, S. H., CHIANG, D. Y., HAYES, D. N., JONES, C., LIU, Y., PRINS, J. F. and LIU, J. (2013). DiffSplice: The genome-wide detection of differential splicing events with RNA-seq. *Nucleic Acids Res.* **41** e39.

JIANG, H. and WONG, W. H. (2009). Statistical inferences for isoform expression in RNA-seq. *Bioinformatics* **25** 1026–1032.

JØRGENSEN, B. (1997). *The Theory of Dispersion Models*. *Monographs on Statistics and Applied Probability* **76**. Chapman & Hall, London. MR1462891

KATZ, Y., WANG, E. T., AIROLDI, E. M. and BURGE, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7** 1009–1015.

LAW, C. W., CHEN, Y., SHI, W. and SMYTH, G. K. (2014). Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15** R29.

LENG, N., DAWSON, J. A., THOMSON, J. A., RUOTTI, V., RISSMAN, A. I., SMITS, B. M. G., HAAG, J. D., GOULD, M. N., STEWART, R. M. and KENDZIORSKI, C. (2013). EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* **29** 1035–1043.

MARIONI, J. C., MASON, C. E., MANE, S. M., STEPHENS, M. and GILAD, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18** 1509–1517.

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 Res.* **40** 4288–4297.

NATIONAL HUMAN GENOME RESEARCH INSTITUTE (2014). Alternative splicing. Available at www.genome.gov.

PAN, Q., SHAI, O., LEE, L. J., FREY, B. J. and BLENCOWE, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40** 1413–1415.

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

R CORE TEAM (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

RICHARD, H., SCHULZ, M. H., SULTAN, M., NÜRNBERGER, A., SCHRINNER, S., BALZEREIT, D., DAGAND, E., RASCHE, A., LEHRACH, H., VINGRON, M., HAAS, S. A. and YASPO, M.-L. (2010). Prediction of alternative isoforms from exon expression levels in RNA-seq experiments. *Nucleic Acids Res*. **38** e112.

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** 139–140.

ROBINSON, M. D. and SMYTH, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23** 2881–2887.

ROBINSON, M. D. and SMYTH, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* **9** 321–332.

RUDDY, S., JOHNSON, M. and PURDOM, E. (2015a). Supplement A to "Shrinkage of dispersion parameters in the binomial family, with application to differential exon skipping." DOI:10.1214/15-AOAS871SUPPA.

RUDDY, S., JOHNSON, M. and PURDOM, E. (2015b). Supplement B to "Shrinkage of dispersion parameters in the binomial family, with application to differential exon skipping." DOI:10.1214/15-AOAS871SUPPB.

RUDDY, S., JOHNSON, M. and PURDOM, E. (2015c). Supplement C to "Shrinkage of dispersion parameters in the binomial family, with application to differential exon skipping." DOI:10.1214/15-AOAS871SUPPC.

SALZMAN, J., JIANG, H. and WONG, W. H. (2010). Statistical modeling of RNA-Seq data. Technical Report No. BIO-252, Division of Biostatistics, Stanford Univ., Palo Alto.

SHEN, S., PARK, J. W., HUANG, J., DITTMAR, K. A., LU, Z.-X., ZHOU, Q., CARSTENS, R. P. and XING, Y. (2012). MATS: A Bayesian framework for flexible detection of differential alternative splicing from RNA-seq data. *Nucleic Acids Res*. **40** e61.

SHI, Y. and JIANG, H. (2013). rSeqDiff: Detecting differential isoform expression from RNA-seq data using hierarchical likelihood ratio test. *PloS One* **8** e79448.

SMYTH, G. K. (2005). Limma: Linear models for microarray data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry and S. Dudoit, eds.) 397–420. Springer, New York.

SUN, D., XI, Y., RODRIGUEZ, B., PARK, H. J., TONG, P., MEONG, M., GOODELL, M. A. and LI, W. (2014). MOABS: Model based analysis of bisulfite sequencing data. *Genome Biol*. **15** R38.

TRAPNELL, C., PACHTER, L. and SALZBERG, S. L. (2009). TopHat: Discovering splice junctions with RNA-seq. *Bioinformatics* **25** 1105–1111.

TRAPNELL, C., WILLIAMS, B. A., PERTEA, G., MORTAZAVI, A., KWAN, G., VAN BAREN, M. J., SALZBERG, S. L., WOLD, B. J. and PACHTER, L. (2010). Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat*. *Biotechnol*. **28** 511.

VENABLES, J. P., KLINCK, R., KOH, C., GERVAIS-BIRD, J., BRAMARD, A., INKEL, L., DURAND, M., COUTURE, S., FROEHLICH, U., LAPOINTE, E., LUCIER, J.-F., THIBAULT, P., RANCOURT, C., TREMBLAY, K., PRINOS, P., CHABOT, B. and ELELA, S. A. (2009). Cancer-associated regulation of alternative splicing. *Nature Publishing Group* **16** 670–676.

WANG, X. (2006). Approximating Bayesian inference by weighted likelihood. *Canad*. *J*. *Statist*. **34** 279–298. MR2323997

WILLIAMS, D. A. (1982). Extrabinomial variation in logistic linear models. *J*. *Roy*. *Statist*. *Soc*. *Ser*. *C* **31** 144–148. MR0673714

WU, T. D. and NACU, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* (*Oxford*, *England*) **26** 873–881.

WU, H., WANG, C. and WU, Z. (2013). A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* **14** 232–243.

WU, J., AKERMAN, M., SUN, S., MCCOMBIE, W. R., KRAINER, A. R. and ZHANG, M. Q. (2011). SpliceTrap: A method to quantify alternative splicing under single cellular conditions. *Bioinformatics* **27** 3010–3016.

YANG, X., TODD, J. A., CLAYTON, D. and WALLACE, C. (2012). Extra-binomial variation approach for analysis of pooled DNA sequencing data. *Bioinformatics* **28** 2898–2904.

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** 1275–1282.

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** 2672–2678.

S. RUDDY
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
367 EVANS HALL
BERKELEY, CALIFORNIA 94720-3860
USA
E-MAIL: sruddy17@gmail.com

M. JOHNSON
GROUP IN BIOSTATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
101 HAVILAND HALL
BERKELEY, CALIFORNIA 94720-7358
USA
E-MAIL: johnsonmk@stat.berkeley.edu

E. PURDOM
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
367 EVANS HALL
BERKELEY, CALIFORNIA 94720-3860
USA
E-MAIL: epurdom@stat.berkeley.edu