Detecting differential usage of exons from RNA-Seq data

Simon Anders*‡, Alejandro Reyes*, Wolfgang Huber
European Molecular Biology Laboratory, 69111 Heidelberg, Germany
* contributed equally ‡ sanders@fs.tum.de

21 October 2011

Abstract

RNA-Seq is a powerful tool for the study of alternative splicing and other forms of alternative isoform expression. Understanding the regulation of these processes requires comparisons between treatments, tissues or conditions. For the analysis of such experiments, we present *DEXSeq*, a statistical method to test for differential exon usage in RNA-Seq data. *DEXSeq* employs generalized linear models and offers good detection power and reliable control of false discoveries by taking biological variation into account. An implementation is available as an R/Bioconductor package.

1 Background

In higher eukaryotes, a single gene can give rise to a multitude of different transcripts (isoforms) by means of varying the usage of splice sites, transcription start sites and polyadenylation sites. We are only beginning to understand which part of this diversity is functional (recently reviewed, e.g., by Nilsen and Graveley (2010) and by Grabowski (2011)). High-throughput sequencing of mRNA (RNA-Seq) promises to become an important technique for the study of alternative isoform regulation, especially in comparisons between samples of different tissues types or of cells in different environmental conditions. This paper presents a method and software tool to analyse data from such experiments.

The regulation of a gene's expression can be separated into two aspects: (i) how many transcripts are produced (regulation of expression strength) and (ii) which of the gene's possible transcripts are produced and, if more than one isoform is present, which proportion of the gene's total output fall onto each transcript (alternative isoform regulation, AIR). Various methods have been published for the analysis of RNA-Seq data with respect to aspect (i), including *edgeR* (Robinson *et al.*, 2010*b*), *DESeq* (Anders and Huber, 2010) and *BaySeq* (Hardcastle and Kelly, 2010). Our method provides statistical inference with respect to aspect (ii).

Statistical inference relies on the comparison of observed differences with the experimentally and biologically expected variability. If the aim is to make a statement about a biological condition with some generality, rather than about one particular biological sample, then biological replicates are needed. While this may be obvious to a reader unfamiliar with the field, it is noteworthy that many experiments reported in the literature have evaded this point. Wang *et al.* (2008) used only one sample per tissue type for inference of AIR. Griffith *et al.* (2010) compared a cell line derived from a single colorectal tumour resistant to a drug with a cell line derived from a single tumour sensitive to the drug. Trapnell *et al.*

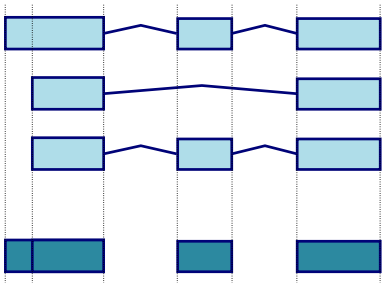

Figure 1: Flattening of gene models: This (fictional) gene has three annotated transcripts involving three exons (light blue), one of which has alternative boundaries. We form *counting bins* (dark blue boxes) from the exons as depicted; the exon of variable length gets split into two bins.

(2010), when presenting their *cufflinks*/*cuffdiff* method, compared consecutive time points, using data from one sample for each time point. Brooks *et al.* (2010) had replicates but did not use them to assess biological variability. Twine *et al.* (2011) compared brain samples from subjects with Alzheimer's disease with samples from normal controls but pooled the material before sequencing and so were unable to assess sample-to-sample variation. A notable instance where biological variation was accounted for in the statistical analysis is the work of Blekhman *et al.* (2010). However, their method relies on the availability of a moderate to large number of samples, and no software implementation was provided.

The importance of accounting for biological variation has been pointed out by Baggerly *et al.* (2003) and recently by Hansen *et al.* (2011). Methods to do so when inferring differential expression were suggested by Baggerly *et al.* (2003) and Lu *et al.* (2005). Subsequently, Robinson and coworkers presented the *edgeR* method (Robinson and Smyth, 2007, 2008; Robinson *et al.*, 2010*b*), which introduced the use of the negative binomal (NB) distribution to RNA-Seq analysis. Robinson *et al.* (2010*a*)) extended *edgeR* with generalized linear models (GLMs) and the Cox-Reid dispersion estimator, discussed later. Their work provides a crucial starting point for the method presented here. In addition, our approach is similar to that of Blekhman *et al.* (2010) and uses ideas from *DESeq* (Anders and Huber, 2010).

In this article, we will first explain the statistical inference procedure (Section 2) and then demonstrate its implementation, the Bioconductor package *DEXSeq*, using a published data set by Brooks *et al.* (2010) (Section 3). In the Discussion (Section 4), we compare with competing methods, especially with the analysis provided by Brooks *et al.* (2010) for their data (which is based on the method of Wang *et al.* (2008)), and with the *cuffdiff* tool provided with the *cufflinks* software by Trapnell *et al.* (2010).

2 Description of the method

2.1 Preparation: Flattening gene models and counting reads

The initial step of an analysis is the alignment of the sequencing reads against the target genome. Here it is important to use a tool capable of properly handling reads that straddle introns. Then, transcriptome annotation with coordinates of exon boundaries is required. For model organisms, reference gene model databases, as provided, e.g., by Ensembl (Flicek *et al.*, 2011), may be used. In addition, such a reference may be augmented by information retrieved from the RNA-Seq dataset that is being studied. Garber *et al.* (2011) review tools for the above tasks.

The central data structure for our method is a table that, in the simplest case, contains for each exon of each gene the number of reads in each sample that overlap with the exon. Special attention is needed, however, if an exon's boundary is not the same in all transcripts. In such cases, we cut the exon in two or more parts (Figure 1). We use the term *counting bin* to refer to exons or parts of exons derived in this manner. Note that a read that overlaps with several counting bins of the same gene is counted for each of these.

2.2 Model and Inference

We denote by k_{ijl} the number of reads overlapping counting bin l of gene i in sample j. We interpret k_{ijl} as a realization of a random variable K_{ijl}. The number of samples is denoted by m, i.e., $j = 1, \ldots, m$.

We write q_{ijl} for the expected value of the concentration of cDNA fragments contributing to counting bin l of gene i, and relate the expected read count, $E(K_{ijl})$ to q_{ijl} via the *size factor* s_j, which describes how deep sample j was sequenced: $E(k_{ijl}) = s_j q_{ijl}$. Note that s_j depends only on j, i.e., the differences in sequencing depth are assumed to cause a linear scaling of the read counts. As the units of s_j and q_{ijl} can be chosen arbitrarily, we use the convention that the geometric mean of the size factors s_j be 1. We estimate the size factors with the same method as in *DESeq* (Anders and Huber, 2010):

$$s_j = \operatorname*{median}_{i,l} \frac{k_{ijl}}{\left(\prod_{j'} k_{ij'l}\right)^{1/m}}.$$

2.2.1 A generalized linear model

We employ generalized linear models (GLMs) (McCullagh and Nelder, 1989) to model read counts. Specifically, we assume K_{ijl} to follow a negative binomial (NB) distribution

$$K_{ijl} \sim NB(\text{mean} = s_j q_{ijl}, \text{dispersion} = \alpha_{il}), \tag{1}$$

where α_{il} is the dispersion parameter (a measure of the distribution's spread) for counting bin (i, l), and the mean is predicted via a logarithmic link by a linear model as

$$\log q_{ijl} = \beta_i^{\mathrm{G}} + \beta_{il}^{\mathrm{E}} + \beta_{i\rho_j}^{\mathrm{C}} + \beta_{i\rho_j l}^{\mathrm{EC}}. \tag{2}$$

The negative binomial distribution in Equation (1) has been useful in many applications of count data regression (Cameron and Trivedi, 1998). Lu *et al.* (2005) and Robinson and Smyth (2007) motivated its use for SAGE or RNA-Seq data; we briefly summarise their argument in Supplementary Note A.1.

We fit one model for each gene i, i.e., the index i in Equation (2) is fixed. The linear predictor q_{ijl} is decomposed into four factors as follows: β_i^{G} represents the expression strength of gene i. β_{il}^{E} is (up to an additive constant) the logarithm of the expected fraction of the reads mapped to gene i that overlap with counting bin l. $\beta_{i\rho_j}^{\mathrm{C}}$ is the logarithm of the fold change in overall expression of gene i under condition ρ_j (the experimental condition of sample j). Finally, $\beta_{i\rho_j l}^{\mathrm{EC}}$ is the effect that condition ρ_j has on the fraction of reads falling into bin l.

To make the model identifiable, constraints on the coefficients are needed; see Supplementary Note A.2.

Of interest in this model are the effects $\beta_{i\rho}^{\mathrm{C}}$ and $\beta_{i\rho l}^{\mathrm{EC}}$. If one of the $\beta_{i\rho l}^{\mathrm{EC}}$ is different from zero, that indicates that the exon it refers to is differentially used. A value of $\beta_{i\rho}^{\mathrm{C}}$ different from zero indicates an overall differential abundance that equally affects all exons, i.e., differential expression. Before we describe the analysis-of-deviance (ANODEV) procedure to test for these effects, we need to discuss the aspect of dispersion.

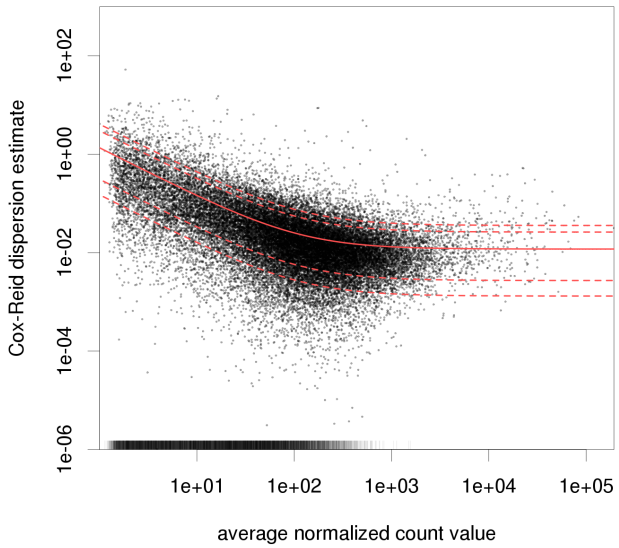

Figure 2: Dependence of dispersion on the mean. Each dot corresponds to one counting bin in the data of Brooks *et al.* (2010), the x axis denotes the normalized count, averaged over all samples, and the y axis is the estimate of the dispersion. The bars at the top and bottom denote dispersion values outside the plotting range (especially including cases where the sample dispersion is essentially zero). The solid red line is the fit used, the dashed lines mark the 1-, 5-, 95- and 99-percentiles of the χ^2 distribution with 5 degrees of freedom scaled to have the fitted mean.

2.2.2 Parameter fitting

For a fixed choice of the dispersion parameter, the NB distribution is a member of the exponential family with respect to the mean. Hence, the iteratively reweighted least square (IRLS) algorithm, which is commonly employed to fit GLMs (McCullagh and Nelder, 1989), allows fitting of the model (1,2) if the dispersion α_{il} is given.

Ordinary maximum likelihood estimation of the dispersion is not advisable, as it has a strong negative bias for small numbers m of samples. This bias is caused by the lack of accounting for the loss of degrees of freedom that arises when estimating the coefficients. Robinson and Smyth (2008) reviewed alternatives and derived an estimator based on the work of Cox and Reid (1987) and Smyth and Verbyla (1996). Cox and Reid suggested to modify the profile log likelihood for the parameter of interest (here: the dispersion) by subtracting a term containing the Fisher information for the other parameters as an approximation to conditioning on the profiled-out parameters. This helps if the parameter of interest is approximately independent from the other parameters with respect to Fisher information, which is the case for the NB likelihood with respect to its parameters mean and dispersion.

However, calculating the Cox-Reid correction term for dispersion estimation in GLMs is not straightforward. The (to our knowledge) best method has been proposed by Smyth et al., who have been using it in the *edgeR* package since version 1.7.18 of September 2010. This method, that has so far only been published

as the software[1] (Robinson *et al.*, 2010*a*), uses information from the last iteration of the preceding IRLS fit, with which the estimates for the coefficients were obtained. Specifically, they propose to subtract from the log profile likelihood for the dispersion the term $\sum_i \log |R_{ii}|$, where R is the upper triangular matrix from the QR decomposition of the reweighted design matrix. We make use of this approach to estimate the dispersion for each counting bin. See Supplementary Note A.3 for details on our implementation.

2.2.3 Two noise components

It is helpful to decompose the extra-Poisson variation of K_{ijl} into two components: variability in gene expression and variability in exon usage. If the expression of a gene i (i.e., the total number of transcripts) in sample j differs from the expected value for experimental condition ρ_j, the values q_{ijl} for all the exons l of gene i will deviate from the values expected for condition ρ_j by the same factor. We denote this the variability in gene expression. By variability in exon usage, we refer to variability in the usage of particular exons. The dispersion parameter α_{il} in Equation (1) with respect to the model of Equation (2) contains both of these parts. However, if we replace Equation (2) with

$$\log \mu_{ijl} = \beta_i^{\mathrm{G}} + \beta_{il}^{\mathrm{E}} + \beta_{ij}^{\mathrm{S}} + \beta_{i\rho_j l}^{\mathrm{EC}}, \tag{3}$$

i.e., instead of fitting one parameter $\beta_{\rho_j}^{\mathrm{C}}$ for the effect of each condition ρ on the expression, we fit one parameter β_{ij}^{S} for each *sample* j, the gene expression variability is absorbed by the model parameters and we are only left with the exon usage variability. Hence, we use model (3) to increase power in our test for differential exon usage. If we wish to test for differential expression, however, we go back to model (2).

We fit model (1,3) for each gene i separately and use Smyth's Cox-Reid dispersion estimation, as described above, to fit a dispersion value $\hat{\alpha}_{il}$ for each counting bin l in the gene.

2.2.4 Information sharing across genes.

If only few replicates are available, as is often the case in high-throughput sequencing experiments, we need to be able to deal with the fact that the dispersion estimator for a single exon has a large sampling variance. A commonly used solution is to share information across exons or genes (Tusher *et al.*, 2001; Lönnstedt and Speed, 2002). We assume that exons with similar count value have similar dispersion and choose the model

$$\alpha(\mu) = \frac{a_1}{\mu} + a_0, \tag{4}$$

to parametrize this relation. This relation appears to fit many data sets we have encountered in practice. (See also Di *et al.* (2011) for a comparison of approaches to model mean-variance relations in RNA-Seq data.)

We regress the dispersion estimates $\hat{\alpha}_{il}$ for all counting bins from all genes on their average normalized count values $\hat{\mu}_{il}$ with a gamma-family GLM to obtain the coefficients α_0 and α_1. As we expect that not all exons follow the regressed trend, we robustify the fit by iteratively leaving out exons with large residuals.

Figure 2 shows a scatter plot of dispersion estimates $\hat{\alpha}_{il}$ against average normalized count values $\hat{\mu}_{il}$, together with the fit $\alpha(\mu)$. For many exons, the difference between per-exon estimate $\hat{\alpha}_{il}$ and fitted value $\alpha(\hat{\mu}_{il})$ appears compatible with a

[1] Note to reviewers: we are aware of a manuscript by Smyth and coworkers that will explain this method and aim to cite it here when available.

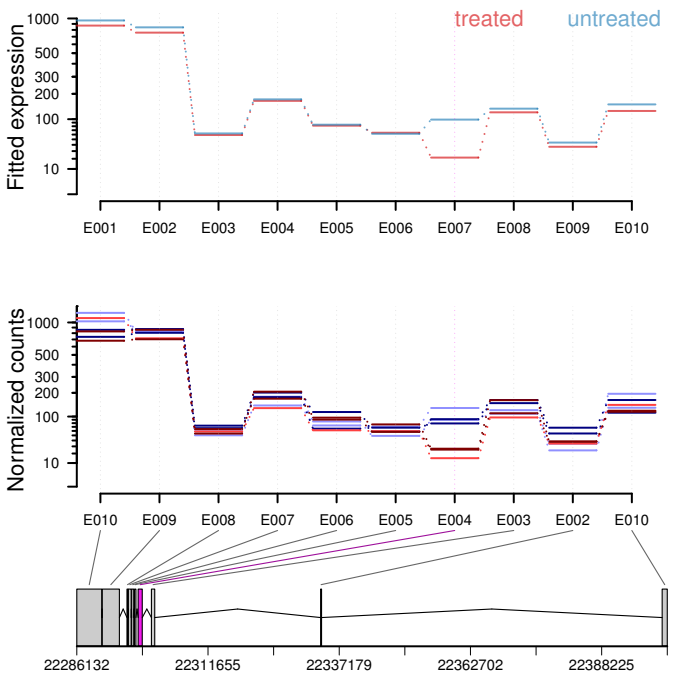

Figure 3: The treatment of knocking down the splicing factor *pasilla* affects the fourth exon (counting bin E004) of the gene *Ten-m* (CG5723). The top panel shows the fitted values according to the linear model, the middle panel shows the normalized counts for each sample, and the bottom panel shows the flattened gene model. Data for knock-down samples are shown in red and for control in blue. In the counts panel, light-coloured lines indicate data from single-end samples and dark colours data from paired-end data.

χ^2 sampling distribution (indicated by the dashed lines). Nevertheless, there are sufficiently many exons with an estimate $\hat{\alpha}_{il}$ so much larger than the fitted value $\alpha(\hat{\mu}_{il})$ that it would not be justified to only rely on the fitted value. Hence, we use as dispersion value α_{il} for the ANODEV (see below) the maximum of the per-exon estimate $\hat{\alpha}_{il}$ and fitted value $\alpha(\hat{\mu}_{il})$. This may cause an overestimation of dispersion, which costs power, but is preferable to using the fitted values only, which carries the risk of losing type-I error control.

2.2.5 Analysis of deviance

We test for each exon whether it is differentially used between different conditions. More precisely, we test against the null hypothesis that the fraction of reads overlapping with a counting bin l, of all the reads overlapping with the gene, does not change between conditions. To this end, we fit for each gene a reduced model with no counting-bin–condition interaction

$$\log \mu_{ijl} = \beta_i^{\mathrm{G}} + \beta_{il}^{\mathrm{E}} + \beta_{ij}^{\mathrm{S}}, \tag{5}$$

and, for each bin l', a model with an interaction coefficient for *only* this bin (indicated by the Kronecker delta $\delta_{ll'}$),

$$\log \mu_{ijl} = \beta_i^{\mathrm{G}} + \beta_{il}^{\mathrm{E}} + \beta_{ij}^{\mathrm{S}} + \beta_{ijl}^{\mathrm{EC}} \delta_{ll'}. \tag{6}$$

We compute the likelihood of these models using the dispersion values α_{il} as estimated from model (3), with the information-sharing scheme of Section 2.2.4. To get an analysis-of-deviance p value, we use a χ^2 likelihood-ratio test.

To test for overall differential expression of the gene, we compare the models

$$\log \mu_{ijl} = \beta_i^{\mathrm{G}} + \beta_{il}^{\mathrm{E}} \tag{7}$$

and

$$\log \mu_{ijl} = \beta_i^{\mathrm{G}} + \beta_{il}^{\mathrm{E}} + \beta_{i\rho_j}^{\mathrm{C}}, \tag{8}$$

using the dispersion estimates obtained for model (2).

Note that differential expression and differential exon usage is aliased if a gene has only one counting bin with non-zero counts. Hence, we mark all counting bins with zero counts in all samples, and all bins in genes with less than two non-zero bins, as *not testable*. Furthermore, we skip counting bins with a count sum across all samples below a threshold chosen low enough that a significant result would be unlikely, to improve power by independent filtering (Bourgon *et al.*, 2010).

2.2.6 Additional covariates

The flexibility of GLMs makes it easy to account for further covariates. For example, if we wish, in addition to the experimental condition ρ_j, for a further covariate τ_j, we extend model (3) as follows:

$$\log \mu_{ijl} = \beta_i^{\mathrm{G}} + \beta_{il}^{\mathrm{E}} + \beta_{ij}^{\mathrm{S}} + \beta_{i\tau_j l}^{\mathrm{EB}} + \beta_{i\rho_j l}^{\mathrm{EC}},$$

When testing for differential exon usage, the extra term $\beta_{i\tau_j l}^{\mathrm{EB}}$ is added to both the reduced model (5) and the full model (6).

When testing for differential expression, we instead add a factor $\beta_{i\tau_j}^{\mathrm{B}}$ to Equations (7) and (8).

An example is provided in Section 3.1 with Equation (10).

2.3 Visualization

The *DEXSeq* package offers facilities to visualize data and fits. An example is shown in Figure 3, using the data discussed in Section 3. Data and results for a gene are presented in three panels. The top panel depicts the fitted values from the GLM fit. For this plot, the data is fitted according to model (2), with the y coordinates showing the exponentiated sums

$$\mu_{ijl} = \exp\left(\tilde{\beta}_i^{\mathrm{G}} + \tilde{\beta}_{il}^{\mathrm{E}} + \tilde{\beta}_{i\rho_j}^{\mathrm{C}} + \tilde{\beta}_{i\rho_j l}^{\mathrm{EC}}\right). \tag{9}$$

The tildes indicate that a decomposition of the linear predictors has been used that separates the effects of expression and isoform regulation, as described in Supplementary Note A.2.

For genes with differential overall expression, it can be difficult to see the evidence for differential exon usage in a plot based on Equation 9. For these cases, the software offers the option to average over the expression effects. Supplementary Figure S6 shows this for the pasilla gene.

Variance stabilizing transformation In Figure 3, a special axis scaling is used, as neither a linear nor logarithmic scale seems appropriate. Instead, the software "warps" the axis scale such that, for data that follows the fitted mean-dispersion relation, the standard deviation corresponds to approximately the same scatter in the y direction throughout the dynamic range. See Supplementary Note A.4 for details.

3 Application

3.1 Analysis of the dataset by Brooks et al.

We considered the data by Brooks *et al.* (2010), who used *Drosophila melanogaster* cell lines and studied the effect of knocking down *pasilla* with RNA-Seq. *Pasilla* and its mammalian homologues NOVA1 and NOVA2 are well-studied splicing factors.

Brooks *et al.* (2010) prepared libraries from RNA extracted from seven biologically independent samples, three control samples and four knock-down samples. They sequenced the libraries on an Illumina Genome Analyzer II, partly using single-end and partly paired-end sequencing and using various read lengths. We obtained the read sequences from the NCBI Gene Expression Omnibus (accession numbers GSM461176 to GSM461181), trimmed them to a common length of 37 nt and aligned them against the *D. melanogaster* reference genome (assembly BDGP5/dm3, without heterochromatic sequences; Hoskins *et al.* (2007)) with TopHat 1.2 (Trapnell *et al.*, 2009). We defined counting bins, as described in Section 2.1, based on the annotation from FlyBase 5.25 (Tweedie *et al.*, 2009) as provided by Ensembl 62 (Flicek *et al.*, 2011).

After counting read coverage for the counting bins, we estimated dispersion values for each bin by fitting, for each gene, a model based on Equations (2,3). Here, since we have a mixture of single-end and paired-end libraries, we extended Equation (3) to

$$\log \mu_{ijl} = \beta_i^{\mathrm{G}} + \beta_{il}^{\mathrm{E}} + \beta_{ij}^{\mathrm{S}} + \beta_{i\rho_j l}^{\mathrm{EC}} + \beta_{i\tau_j l}^{\mathrm{ET}}, \tag{10}$$

where $\tau_j = 1, 2$ is the library type of sample j, single-end or paired-end (see also Section 2.2.6).

The estimated dispersions are shown in Figure 2. The fitted line is given by $\alpha(\mu) = 1.3/\mu + 0.012$, which has the form of Equation (4). The parameter $a_0 = 0.012$ represents the amount of biological variation: Taking the square root, we can see that the exon usage typically differs with a coefficient of variation of around 11% between biological replicates for strongly expressed exons.

Here, we can also see the advantage of absorbing expression variability in a sample coefficient (Section 2.2.3). Had we used Equation (2) instead of Equation (3), we would have had to work with a higher dispersion, namely $\alpha'(\mu) = 1.6/\mu + 0.018$, and so would have lost power.

We performed the test for differential exon usage described in Section 2.2.5 for all counting bins that had at least 10 counts summed over all 7 samples. We controlled the false discovery rate (FDR) with the Benjamini-Hochberg method and found, at 10% FDR, significant differential exon usage for 259 counting bins, affecting 159 genes.

Figure 3 shows gene *Ten-m*, which exhibited a clear signal for differential usage of counting bin E004 ($p = 2.1 \cdot 10^{-11}$, after Benjamini-Hochberg adjustment $p_{\mathrm{adj}} = 1.2 \cdot 10^{-8}$). Similar plots can be found, for all genes in this study, at `http://www-huber.embl.de/pub/DEXSeq/psfb/testForDEU.html`.

Figure 4 gives an overview of the test results and shows how the detection power depends on the mean. Remember from Equations (11) and (4) that the coefficient of variation (CV) of the normalized counts, K_{ijl}/s_j, is modeled as

$$\mathrm{CV}\left(\frac{K_{ijl}}{s_j}\right) \geq \frac{1+\alpha_1}{s_j}\frac{1}{\mu_{ijl}} + \alpha_0,$$

and hence detection power increases with expression strength: For strongly expressed exons, $\log_2$ fold changes around 0.5 (corresponding to fold changes around 40%) can be significant, while for weakly expressed with around 30 counts, fold changes above 2-fold are required.

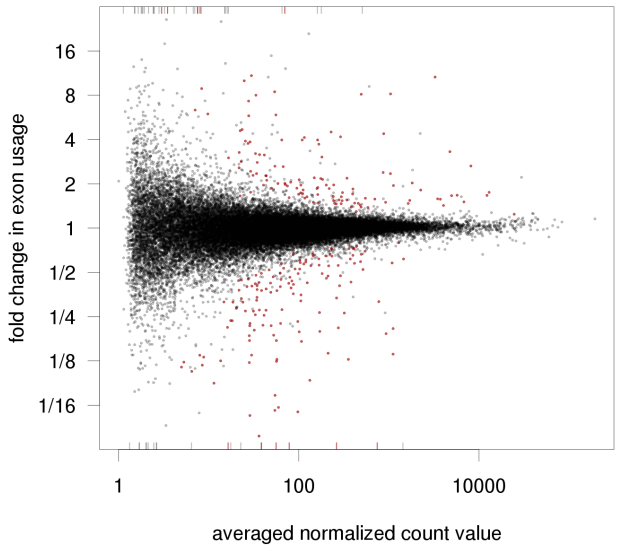

Figure 4: Fold changes of exon usage versus averaged normalized count value for all tested counting bins for the Brooks et al. data. Red colour indicates significance at 10% FDR. Bars at the margin point to bins with fold changes outside the plotting range.

4 Discussion

4.1 Importance of modelling overdispersion

The method presented here differs from previous work by accounting for sample-to-sample variation in excess of Poisson variation. In the following, we investigate whether this extra variation is large enough to influence results in practice.

To address this question for our inference procedure, we re-computed the tests for differential exon usage after setting the dispersion values α_{il} in Equations (1, 5, 6) to zero. This corresponds to assuming that the variation in the data follows a Poisson distribution. Cutting again the Benjamini-Hochberg–adjusted p values at 10%, we obtained 36 times as many hits: significant differential exon usage was reported for 9,432 counting bins in 3,610 genes. (See Supplementary Figure S7 and compare with Figure 4.) For these extra hits, however, the treatment effect was not large compared to the variation seen between replicates, and thus there is no evidence for them being true positives.

The assumption that variability is limited to Poisson noise is also implicit in analysis methods based on Fisher's test, which we discuss next.

4.1.1 Analyses based on Fisher's test

To test for for differential isoform regulation, of Wang *et al.* (2008) and Brooks *et al.* (2010) employed 2×2 contingency tables and Fisher's exact test. In this approach, the contingency table's rows corresponded to control and treatment, The cells in one column contained the numbers of reads supporting inclusion of an exon (i.e., reads overlapping the exon) and the cells in the other column gave the numbers of reads supporting exclusion (e.g., in the case of casette exons, reads straddling the exon). In the study of Wang *et al.* (2008), each row corresponded

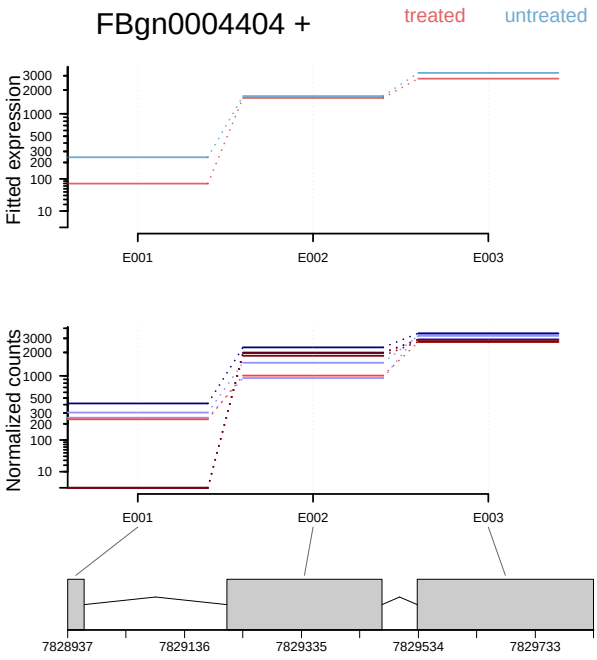

Figure 5: Ribosomal protein gene RpS14b is shown here as an example for a gene with very heterogeneous dispersion. The first exon has zero count in the paired-end samples *untreated 2*, in the single-end sample *treated 2* and in the paired-end sample *treated 3*. Colours as in Figure 3.

to a single sample, while Brooks *et al.* (2010) summed up the number of reads from their replicates. Hence, in both cases, the contingency tables did not contain information on sample-to-sample variability (Baggerly *et al.*, 2003) and so, the results may be expected to contain an inflated number of false positives.

As an example, Supplementary Figure S8 shows gene *Lk6*, for which Brooks et al. reported differential use of its alternative first exons. Our analysis, too, indicated that the average expression strength of exon E002 was different between the conditions. However, examining the counts from the individual biological replicates revealed that the variance within treatment group was large compared to this difference, and hence, it should not be considered significant.

4.2 Heterogeneity of dispersions

In our model, we allow the counting bins of a gene to have different dispersion values. Gene *RpS14b* (Figure 5) exhibits very different variability for its three exons and so illustrates the need for this modelling choice.

The first exon also illustrates the value of replicates, and the importance of making use of their information. This exon had between 252 and 416 (normalized) counts in four of the samples and no counts in three. However, this difference cannot be attributed to the treatment because both the control and the treatment group contained samples with zero counts as well as samples with several hundreds of counts. Hence, the reason for the difference in read counts for this exon cannot be the knock-down of pasilla and is likely some other difference between the samples' treatment that was not under the experimenters' control.

If one just adds up or averages the samples in a treatment group, as done in the contingency table method, one would only see a sizeable difference, as in the upper panel of the figure, and might call a significant effect. It is also crucial that

the test for differential exon usage does not rely on the fitted dispersion (solid line in Fig. 2) only, as the effect size would seem significant if one did not take note that the actual observed within-group variance is so much larger that the fitted value is implausible. The maximum rule of Section 2.2.4 assures this.

4.3 Comparison with cuffdiff

Cufflinks (Trapnell *et al.*, 2010) is a tool to infer gene models from RNA-Seq data and to quantify the abundance of transcript isoforms in an RNA-Seq sample. In addition to this, the *cuffdiff* module allows testing for differences in isoform abundance. *Cuffdiff*, as described in Trapnell *et al.* (2010), compares a single sample with another one and does not attempt to account for sample-to-sample variability. The latter is also true for the version described by Roberts *et al.* (2011), which allows processing of replicate samples, but uses this for the assessment only of bias, not of variability. Hence, the same drawbacks may be expected as discussed earlier for the Fisher-test-based methods. More recently, starting with version 1.0.0, cufflinks attempts to assess overdispersion and account for it.

We compared the three knock-down samples of the Brooks et al. data set against the four control samples with version 1.1.0 of *cuffdiff*. With nominal FDR control at 10%, *cuffdiff* reported differential splicing for 108 genes.

To test the control of false-positive rate, we made use of the fact that we had four replicates for the control condition. We formed one group *C1* from samples 1 and 3 and another group *C2* from samples 2 and 4. We tasked both *DEXSeq* and *cuffdiff* with comparing group C1 versus group C2 at a nominal FDR of 10%. As this is a comparison between replicates, we ideally should not get any significant calls. Note that each group contained one single-end and one paired-end sample, i.e., the blocking caused by the library type was balanced between the groups.

In this comparison, *DEXSeq* found 8 genes to be significantly affected by differential exon usage, compared to 159 in the comparison of treatment versus control. *Cufflinks* found 207 genes with significant differential splicing, which –contrary to what is expected– was more than the 108 genes found in the treatment-control comparison.

4.4 Comparing exon or isoform usage

The interpretation of the results of our method is straightforward when a single exon of a gene with many exons is called differentially used. However, if many exons within a gene are affected, the interpretation is more complex. For instance, consider a gene with two isoforms, a long one with several exons, and a short one consisting of only the first exon. If an experimental condition causes more transcripts to be truncated after the first exon, but does not affect the total number of transcripts, one might expect an analysis to indicate differential usage for all but the first exon. However, our method cannot distinguish this situation from one where the gene is overall down-regulated, while the first exon is more strongly used.

Hence, if differential exon usage is detected within a gene, we can safely conclude that this gene is affected by alternative isoform regulation. However, the test's output with regard to *which* of the counting bins are affected can be unreliable if the isoform regulation affects a large fraction of the exons. In practice, the assignment to counting bins is reliable as long as only one or a few counting bins in a gene are called significant.

Methods that attempt to estimate not just the abundance of exons but of isoforms, such as the method of Jiang and Wong (2009), *cufflinks* (Trapnell *et al.*, 2010) and *MMSeq* (Turro *et al.*, 2011), may be able to circumvent this issue.

Of these, only *cufflinks*/*cuffdiff* offers the functionality of comparing between samples. We discussed *cuffdiff* in Section 4.3.

4.5 Implementation

We implemented *DEXSeq* in the statistical pogramming language *R* (R Development Core Team, 2009) and have made it available as open source software via the *Bioconductor* project (Gentleman *et al.*, 2004). See the *Bioconductor* web page for downloading instructions. *DEXSeq* can be used on MacOS, Linux and Windows.

For the preparation steps, namely the "flattening" of the transcriptome annotation to counting bins and the counting of the reads overlapping each counting bin, two Python scripts are provided, which are built on the *HTSeq* framework (Anders, 2011). The first script takes a GTF file with gene models and transforms it into a GFF file listing counting bins, the second takes such a GFF file and an alignment file in the SAM format and produces a list of counts. The R package is used to read these counts, estimate the size factors and dispersions, fit the dispersion-mean relation and test for differential exon usage. Other R or Bioconductor functionality can be used for downstream analyses. Furthermore, *DEXSeq* can create a set of HTML pages that contain the results of the tests, and, for each gene, plots like Figures 3 and 5. The HTML output allows browsing of the results with a web browser and facilitates sharing of the results with colleagues by uploading the files to a web server.

The *DEXSeq* package provides functions on different levels. In the simplest case, a single function is called that runs all the steps of a standard analysis. To give experienced users the possibilty to interfere with the workflow, functions are also provided to run each step seperately, to run some steps only for single genes, and to inspect intermediate and final results.

The use of the package is explained in the vignette (a manual with a worked example) and documentation pages for all functions.

As the *DEXSeq* method relies on fitting GLMs of the NB family, a performant IRLS fitting function is required. We use the function *nbglm.fit* (McCarthy *et al.*, 2011) from the *statmod* package, which offers better performance and convergence than alternative, older implementations.

Fitting GLMs for many genes and counting bins is a computationally expensive process. When running on a single core of a current desktop computer, the analysis of the Brooks et al. data presented here takes a couple of hours. However, the method lends itself for straight-forward parallelization: we use the *multicore* package (Urbanek, 2011) to offer spreading the computation on several CPU cores.

5 Conclusion

We have presented a method, called *DEXSeq*, to test for evidence of differential usage of exons and hence of isoforms in RNA-Seq samples from different experimental conditions using generalized linear models. *DEXSeq* achieves reliable control of false discovery rate by estimating variability (dispersion) for each exon or counting bin and good power by sharing dispersion estimation across features. The method is implemented as an open-source *Bioconductor* package, which also facilitates data visualization and exploration.

References

Anders, S. 2011 HTSeq: Analysing high-throughput sequencing data with Python. http://www-huber.embl.de/users/anders/HTSeq/.

Anders, S. and Huber, W. 2010 Differential expression analysis for sequence count data. *Genome Biology*, **11**(10), R106. (doi:10.1186/gb-2010-11-10-r106)

Baggerly, K. A., Deng, L., Morris, J. S. and Aldaz, C. M. 2003 Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics*, **19**(12), 1477–1483. (doi:10.1093/bioinformatics/btg173)

Blekhman, R., Marioni, J. C., Zumbo, P., Stephens, M. and Gilad, Y. 2010 Sex-specific and lineage-specific alternative splicing in primates. *Genome Research*, **20**(2), 180–9. (doi:10.1101/gr.099226.109)

Bourgon, R., Gentleman, R. and Huber, W. 2010 Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences*, **107**(21), 9546–51. (doi:10.1073/pnas.0914005107)

Brooks, A. N., Yang, L., Duff, M. O., Hansen, K. D., Park, J. W., Dudoit, S., Brenner, S. E. and Graveley, B. R. 2010 Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Research*, **21**(2), 193–202. (doi:10.1101/gr.108662.110)

Cameron, A. C. and Trivedi, P. K. 1998 *Regression analysis of count data*. Cambridge University Press.

Cox, D. R. and Reid, N. 1987 Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society, Series B*, **49**(1), 1–39.

Di, Y., Schafer, D. W., Cumbie, J. S. and Chang, J. H. 2011 The NBP negative binomial model for assessing differential gene expression from RNA-Seq. *Statistical Applications in Genetics and Molecular Biology*, **10**(1). (doi:10.2202/1544-6115.1637)

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S. *et al.* 2011 Ensembl 2011. *Nucleic Acids Research*, **39**(Database issue), D800–6. (doi:10.1093/nar/gkq1064)

Garber, M., Grabherr, M. G., Guttman, M. and Trapnell, C. 2011 Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, **8**(6), 469–77. (doi:10.1038/nmeth.1613)

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y. *et al.* 2004 Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, **5**, R80. Project homepage: `http://www.bioconductor.org`.

Grabowski, P. 2011 Alternative splicing takes shape during neuronal development. *Current Opinion in Genetics & Development*. (doi:10.1016/j.gde.2011.03.005)

Griffith, M., Griffith, O. L., Mwenifumbo, J., Goya, R., Morrissy, A. S., Morin, R. D., Corbett, R., Tang, M. J., Hou, Y.-C. *et al.* 2010 Alternative expression analysis by RNA sequencing. *Nature Methods*, **7**(10), 843–7. (doi:10.1038/nmeth.1503)

Hansen, K. D., Wu, Z., Irizarry, R. A. and Leek, J. T. 2011 Sequencing technology does not eliminate biological variability. *Nature Biotechnology*, **29**(7), 572–573. (doi:10.1038/nbt.1910)

Hardcastle, T. J. and Kelly, K. A. 2010 baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, **11**(1), 422. (doi:10.1186/1471-2105-11-422)

Hoskins, R. A., Carlson, J. W., Kennedy, C., Acevedo, D., Evans-Holm, M., Frise, E., Wan, K. H., Park, S., Mendez-Lago, M. *et al.* 2007 Sequence finishing and mapping of Drosophila melanogaster heterochromatin. *Science*, **316**(5831), 1625–8. (doi:10.1126/science.1139816)

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

Lönnstedt, I. and Speed, T. 2002 Replicated microarray data. *Statistica Sinica*, **12**, 31–46.

Lu, J., Tomfohr, J. K. and Kepler, T. B. 2005 Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach. *BMC Bioinformatics*, **6**, 165. (doi:10.1186/1471-2105-6-165)

McCarthy, D. J., Chen, Y. and Smyth, G. K. 2011 Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Submitted.

McCullagh, P. and Nelder, J. A. 1989 *Generalized Linear Models*. Chapman & Hall/CRC, 2nd edn.

Nilsen, T. W. and Graveley, B. R. 2010 Expansion of the eukaryotic proteome by alternative splicing. *Nature*, **463**(7280), 457–63. (doi:10.1038/nature08909)

R Development Core Team 2009 *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0; `http://www.R-project.org`.

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L. and Pachter, L. 2011 Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*, **12**(3), R22. (doi:10.1186/gb-2011-12-3-r22)

Robinson, M., McCarthy, D., Chen, Y. and Smyth, G. 2010*a* edgeR: Empirical analysis of digital gene expression data in R. Bioconductor package, available from http://www.bioconductor.org.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. 2010*b* edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140. (doi:10.1093/bioinformatics/btp616)

Robinson, M. D. and Smyth, G. K. 2007 Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**(21), 2881–2887. (doi:10.1093/bioinformatics/btm453)

Robinson, M. D. and Smyth, G. K. 2008 Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, **9**(2), 321–32. (doi:10.1093/biostatistics/kxm030)

Smyth, G. K. and Verbyla, A. P. 1996 A conditional likelihood approach to residual maximum likelihood estimation in generalized linear models. *Journal of the Royal Statistical Society, Series B*, **58**, 565–572.

Trapnell, C., Pachter, L. and Salzberg, S. L. 2009 TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**(9), 1105–11. (doi:10.1093/bioinformatics/btp120)

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. *Nature Biotechnology*, **28**(5), 511–515. (doi:10.1038/nbt.1621)

Turro, E., Su, S.-Y., Goncalves, A., Coin, L. J. M., Richardson, S. and Lewin, A. 2011 Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biology*, **12**(2), R13. (doi:10.1186/gb-2011-12-2-r13)

Tusher, V., Tibshirani, R. and Chu, C. 2001 Significance analysis of microarrays applied to ionizing radiation response. *Proceedings of the National Academy of Sciences*, **98**, 5116–5121. (doi:10.1073/pnas.091062498)

Tweedie, S., Ashburner, M., Falls, K., Leyland, P., McQuilton, P., Marygold, S., Millburn, G., Osumi-Sutherland, D., Schroeder, A. *et al.* 2009 FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Research*, **37**(Database issue), D555–9. (doi:10.1093/nar/gkn788)

Twine, N. A., Janitz, K., Wilkins, M. R. and Janitz, M. 2011 Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimer's disease. *PloS one*, **6**(1), e16 266. (doi:10.1371/journal.pone.0016266)

Urbanek, S. 2011 *multicore: Parallel processing of R code on machines with multiple cores or CPUs*. R package version 0.1-7, available from `http://cran.r-project.org`.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P. and Burge, C. B. 2008 Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**(7221), 470–6. (doi:10.1038/nature07509)

Supplement

A Supplementary Notes

A.1 The negative binomial distribution from a gamma-Poisson hierarchical model

The negative binomial (NB) distribution (Equation (1)) has been useful in many applications of count data regression (Cameron and Trivedi, 1998). A motivation for its use with SAGE or RNA-Seq data has been given by Lu *et al.* (2005) and Robinson and Smyth (2007), and we briefly summarise their argumentation. If we denote by Q_{ijl} the concentration of cDNA fragments mapping to counting bin l of gene i in sample j, then the number of counts K_{ijl}, conditioned on $Q_{ijl} = q_{ijl}$, is Poisson-distributed with mean $s_j q_{ijl}$. This follows from the fact that each of the cDNA molecules (whose number is proportional to q_{ijl}) has the same small probability of getting sequenced, i.e., sequencing can be seen as a Bernoulli process with small success probability. As variance and mean are equal in a Poisson distribution, we have

$$\operatorname{Var}(K_{ijl} \,|\, Q_{ijl} = q_{ijl}) = \operatorname{E}(K_{ijl} \,|\, Q_{ijl} = q_{ijl}) = s_j q_{ijl}$$

and, by the law of total variance,

$$\operatorname{Var}(K_{ijl}) = s_j E(Q_{ijl}) + s_j^2 \operatorname{Var}(Q_{ijl}). \tag{11}$$

This fixes the first two moments of the distribution of K_{ijl} by the first two moments of Q_{ijl}. In order to fix the higher order moments one commonly assumes Q_{ijl} to follow a gamma distribution, because then, the distribution of K_{ijl} becomes the negative binomial, which is easy to handle.

The relationship between variance v and mean μ of a NB distribution is commonly parametrized as $v = \mu + \alpha\mu^2$, where the constant α is known as the *dispersion parameter*. Comparing with Eq. (11) shows that the dispersion parameter can be readily interpreted as the squared coefficient of variation (SCV) of Q_{ijl}.

A.2 Balancing

When setting up a design matrix for a linear models with categorical variables, one needs to chose a contrast encoding that constrains the coefficients for the different levels of each factor. When fitting our models, we follow the standard approach of setting the coefficients concerning the control condition $\rho = 1$ and those concerning counting bin $l = 1$ to zero. However, the latter is a problem in interpreting the estimated coefficient and when using them for visualization, as it lets counting bin 1 appear differently and will not show any differential usage of it. (Note that this issue does not affect testing, as in the tests (Equation (6)), we have interaction terms for only one counting bin at a time.)

To treat all counting bins equally in Equation (9), we "spread" the gene effect over all counting bins by setting

$$\tilde{\beta}^{\mathrm{E}}_{il} = \beta^{\mathrm{E}}_{il} - \overline{\beta}^{E}_{i}, \qquad \tilde{\beta}^{\mathrm{EC}}_{i\rho l} = \beta^{\mathrm{EC}}_{i\rho l} - \overline{\beta}^{EC}_{i\rho},$$
$$\tilde{\beta}^{\mathrm{G}}_{i} = \beta^{\mathrm{G}}_{i} + \overline{\beta}^{E}_{i}, \qquad \tilde{\beta}^{\mathrm{C}}_{i\rho} = \beta^{\mathrm{C}}_{i\rho} + \overline{\beta}^{EC}_{i\rho},$$

where the shifts $\overline{\beta}^{E}_{i}$ and $\overline{\beta}^{EC}_{i\rho_j}$ are weighted averages of the original exon and exon-interaction coefficients:

$$\overline{\beta}^{E}_{i} = \frac{\sum_l w_{il}\beta^{E}_{il}}{\sum_l w_{il}}, \qquad \overline{\beta}^{EC}_{i\rho} = \frac{\sum_l w_{il}\beta^{EC}_{i\rho l}}{\sum_l w_{il}}.$$

This is similar to the use of "sum contrasts" offered by statistical software packages. The difference is that we weight the contributions to the average by the reciprocal of an estimate of their sampling variance, as these can differ strongly. (An exon with low count could otherwise get undue influence on the average.) As proxy for this, we use the expected variance (as given by the dispersion values used in the fit) of the logarithm of the normalized counts for exon l, i.e., we set

$$\frac{1}{w_{il}} = \frac{1}{\overline{\mu}_{il}} + \alpha_{il},$$

where $\overline{\mu}_{il}$ is the fitted expression of exon l, averaged over all conditions,

$$\overline{\mu}_{il} = \exp\left[\beta_i^{\mathrm{G}} + \beta_{il}^{\mathrm{E}} + \frac{1}{n_C}\sum_{\rho}^{n_C}\left(\beta_{i\rho}^{\mathrm{C}} + \beta_{i\rho l}^{\mathrm{EC}}\right)\right]$$

(with n_C the number of conditions). These "balanced" coefficients are reported as estimates for the strengths of differential exon usage and used in plotting (Sec. 2.3).

A.3 Details on the Cox-Reid dispersion estimation

When maximizing a profile likelihood one needs to find a maximum-likelihood estimate of the nuisance parameters each time the optimizer evaluates the objective function, i.e., the profile log likelihood. This can lead to long computation times. In the case of NB GLMs, the coefficients found by IRLS depend only weakly on the value one has used for the dispersion. Hence, we use the following short-cut, which gives nearly the same results as a full profile likelihood maximization: For each gene, we first perform an IRLS fit, using some initial value for the dispersion, then, we insert these fitted values in the log likelihood function with Smyth's Cox-Reid term and find its maximum using Brent's line search. One might iterate this, i.e., obtain new fitted values with the maximizing dispersion and redo the maximization, but for typical data, this changes the dispersion estimate only negligibly, and hence, we go without iterating the procedure.

Furthermore, as the coefficients hardly change when the dispersion is varied, it is sufficient to perform the IRLS only once at the beginning of the optimization. In each optimization step, the only computationally expensive part left is the QR decomposition of the weighted design matrix, which needs to be redone because the weights depend on the dispersion.

A.4 Variance stabilizing transformation

To achieve the axis warping described at the end of Section 2.3, a variance stabilizing transformation (VST) is derived from Equation (4):

$$\tau(x) = \int^x \frac{d\mu}{\sqrt{v(\mu)}} = \int^x \frac{d\mu}{\sqrt{\mu + \alpha(\mu)\mu^2}}$$
$$= \frac{2}{\sqrt{\alpha_0}} \log\left(2\alpha_0\sqrt{x} + 2\sqrt{\alpha_0(\alpha_0 x + \alpha_1 + 1)}\right)$$

To the extent that the counts k_{ijk} follow the dispersion relation (4), the transformed data $\tau(k_{ijl}/s_j)$ is approximately homoscedastic, and hence, transforming the y coordinates in the plots with the function τ achieves the desired effect.

Another use of the VST is in ranking a list of counting bins with significant differential use. Ranking by logarithmic fold change estimates $\beta_{i,2,l}^{\mathrm{EC}} - \beta_{i,1,l}^{\mathrm{EC}}$ is typically unsatisfactory, as this will bring to the top many bins with few counts due to the large sampling variance of their logarithmic fold change estimates. Ranking by $\tau\left(\exp\beta_{i,2,l}^{\mathrm{EC}}\right) - \tau\left(\exp\beta_{i,1,l}^{\mathrm{EC}}\right)$ gives more informative results.

B Supplementary Figures

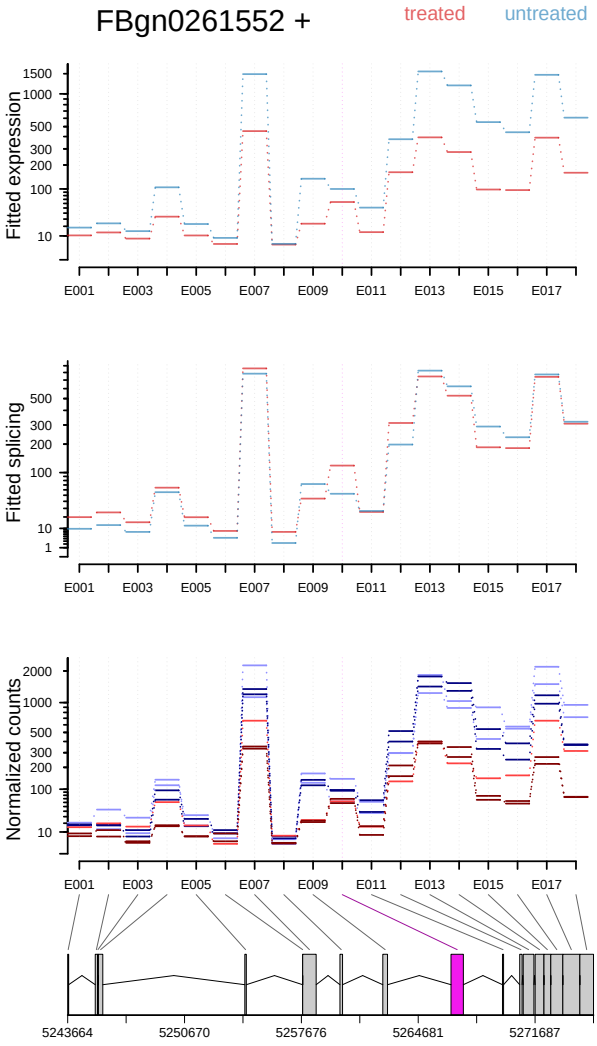

Figure S6: As pasilla is knocked down, its counts are much lower in treatment then in control samples (first and third panel). This makes it difficult to see why exon E010 is called significant (bottom panel), while the plot in the second panel, which replaces the overall expression coefficient $\beta^C_{i\rho_j}$ by its mean (see text for details). Colours as in Figure 3. There are two possible biological interpretations of this data: either, pasilla influences its own splicing, or the RNAi knockdown has different efficiency for the gene's different isoforms.

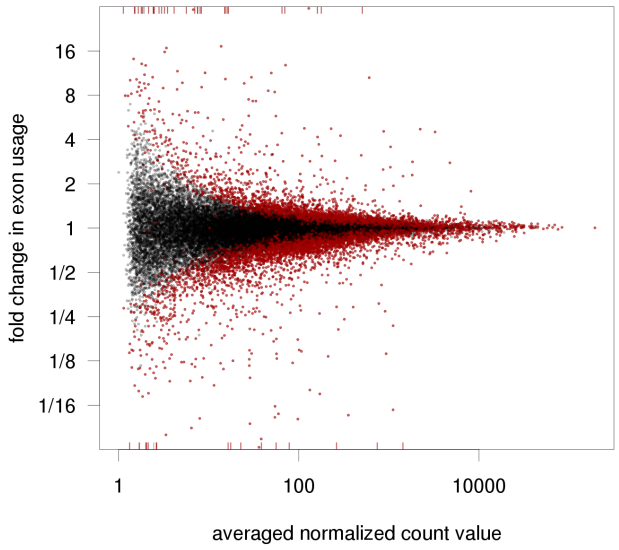

Figure S7: The same plot as in Figure 4, but with the red colour now indicating counting bins which appear to show significant differential exon usage when neglecting to account for biological variation in the test.

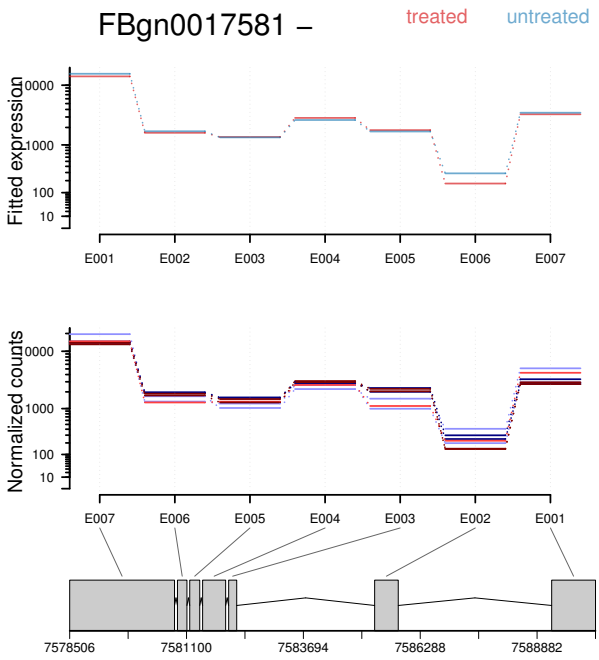

Figure S8: For this gene, *Lk6* (CG17342, in Brooks et al.'s annotation SG11207) Brooks et al. report a significant change in category *alternative first exon*. In fact, the usage of the two isoforms seems to change from sample to sample. Due to the high within-group variation this difference should not be attributed to the treatment. Colours are as in Figure 3.