Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

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

23 March 2012

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 sensitive and specific detection of differential isoform abundance in comparisons between conditions, cell types or tissues. We present *DEXSeq*, a statistical method to test for differential exon usage in RNA-Seq data. *DEXSeq* employs generalized linear models and offers reliable control of false discoveries by taking biological variation into account. *DEXSeq* detect genes, and in many cases specific exons, that are subject to differential exon usage with high sensitivity. We demonstrate the versatility of *DEXSeq* by applying it to several data sets. The method facilitates the study of regulation and function of alternative exon usage on a genome-wide scale. An implementation of *DEXSeq* is available as an R/Bioconductor package.

Introduction

In higher eukaryotes, a single gene can give rise to a multitude of different transcripts (isoforms) by 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 different tissues or cell types, or between cells in different environmental conditions or with different genetic backgrounds.

Shotgun sequencing The median length of human transcripts is 2186 nucleotides (nt), with the longest transcripts having sizes of up to 101206 nt (these numbers are based on UCSC hg19). An ideal RNA-Seq technology would produce sequence reads that directly correspond to full length transcripts. Current implementations of RNA-Seq, however, employ shorter reads and use a shotgun sequencing approach. For instance, Illumina's HiSeq 2000 produces reads of length 101 nt, which are typically paired so that they cover the two ends of shotgun fragments of lengths between 200 and 500 nt.

Approaches to the analysis of such data may be grouped into three main categories. First, in an approach that is reminescent of microarray expression profiling, one simply counts the fragments from each gene locus, irrespective of transcript isoform, to measure each gene's overall expression strength in each of the experimental samples. Several methods have been published for the detection of statistically significant differences in such count values across conditions, including *edgeR* (Robinson *et al.*, 2010*b*), *DESeq* (Anders and Huber, 2010) and *BaySeq* (Hardcastle and Kelly, 2010).

Second, one tries to assemble the fragments into full-length transcripts, using the fragment coverage to estimate each transcript's expression strength in each of the experimental samples. This approach has been pursued by Jiang and Wong (2009), Trapnell *et al.* (2010) and Turro *et al.* (2011). Of these, only Trapnell *et al.* (2010) attempt inference of differential expression by comparing between these estimates. Such inference is challenging, due to uncertainties from the first step. In addition, the accumulation of uncertainties might lead to less inferential power for certain types of questions than the third category of approaches, as is shown in the following.

Third, one avoids the assembly step and looks for differences across conditions between quantities that are directly observable from the shotgun data, such as the (relative) usage of each exon. That is the approach which is described in this article.

Transcript catalogisation versus differential expression Shotgun RNA-Seq data can be used both for transcript catalogisation and differential expression analysis. In catalogisation, one annotates the regions of the genome than can be expressed, i. e. the exons, and how the pre-RNAs are spliced into transcripts. In differential expression analysis, one aims to study the regulation of these processes across different conditions. In the method described here, we assume that catalogisation has been done, and focus on differential expression.

Biological variability If our aim is to make a statement about the regulation of a biological process across different conditions with some generality, rather than only making statements about singular biological samples, then a suitable level of replication in the data is needed. While that may be obvious to a reader unfamiliar with the field, it is noteworthy that most methods suggested so far for the study of alternative isoform regulation (AIR) have evaded this point. Wang *et al.* (2008) presented a method for inference of differential exon usage based on 2 × 2 contingency tables of read counts and Fisher's exact test. As we show in the Discussion, this method cannot account for biological variability, and in fact, the data used to demonstrate the method comprised only one sample each per tissue type. In follow-up work, Katz *et al.* (2010) refined this method (now termed *MISO*); however, they still compared only one knock-down sample with a single control sample and made no attempt of addressing biological variability. Griffith *et al.* (2010) demonstrate their *AlexaSeq* analysis method by comparing 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. This method, too, is not applicable to replicated samples. Trapnell *et al.* (2010), when presenting the *cufflinks*/*cuffdiff* tool chain, compared consecutive time points, using data from one sample for each time point. The *cuffdiff* software tool, in the version described in the paper, can only process pairs of samples without replicates. Brooks *et al.* (2010) used replicates, but did not use them to assess biological variability, because they used a modified version of Wang *et al.* (2008)'s method. 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

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

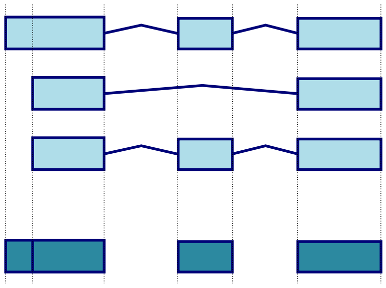

Figure 1: Flattening of gene models: This (fictional) gene has three annotated transcripts involving three exons (light blue), one of which has alternative boundaries. We form *counting bins* (dark blue boxes) from the exons as depicted; the exon of variable length gets split into two bins.

(Robinson and Smyth, 2007, 2008; Robinson *et al.*, 2010*b*), which introduced the use of the negative binomial 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. The basic approach of using exon–condition interactions in linear or generalized linear models to detect differential exon usage has been explored before by Cline *et al.* (2005) and Purdom *et al.* (2008) for exon microarrays and by Blekhman *et al.* (2010) for RNA-Seq data. Our approach can be seen as a further development of these approaches that also incorporated ideas from *DESeq* (Anders and Huber, 2010).

In this article, we will first explain the statistical inference procedure and then use it to reanalyse published data sets by Brooks *et al.* (2010), by Brawand *et al.* (2011) and by the ENCODE Project Consortium (2011). In the Discussion, we elaborate on the observation that most published methods are unable to account for biological variation, focusing on the analysis provided by Brooks *et al.* (2010) for their data (which is based on the method of Wang *et al.* (2008)), and illustrate how this leads to unreliable results. Finally, we compare *DEXSeq* with the one competing tool that claims to account for biological variation, namely the new versions of *cuffdiff*.

Results

Description of the method

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 data set 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.

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

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., $j = 1, \dots, m$.

We write μ_{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 μ_{ijl} via the *size factor* s_j, which describes how deep sample j was sequenced: $E(k_{ijl}) = s_j \mu_{ijl}$. 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. We estimate the size factors with the same method as in *DESeq* (Anders and Huber, 2010); for details, please see Supplementary Note S.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

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

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

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

The negative binomial distribution in Equation (1) has been useful in many applications of count data regression (Cameron and Trivedi, 1998). It can be seen as a generalization of the Poisson distribution: for a Poisson distribution, the variance v is equal to the mean μ, while for the negative binomial, the variance is $v = \mu + \alpha\mu^2$, with the dispersion α describing the squared coefficient of variation in excess of the Poisson case. Lu *et al.* (2005) and Robinson and Smyth (2007) motivated the use of the NB distribution for SAGE or RNA-Seq data; we briefly summarise their argument in Supplementary Note S.2.

We fit one model for each gene i, i.e., the index i in Equation (2) is fixed. The linear predictor μ_{ijl} is decomposed into four factors as follows: β_i^{G} represents the baseline 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 S.3.

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 counting bin 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 counting bins, i.e., overall differential expression of the gene. Before we describe the analysis-of-deviance (ANODEV) procedure to test for these effects, we need to discuss the aspect of dispersion.

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 suitable, as it has a strong negative bias when the number of samples is small. The bias is caused

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

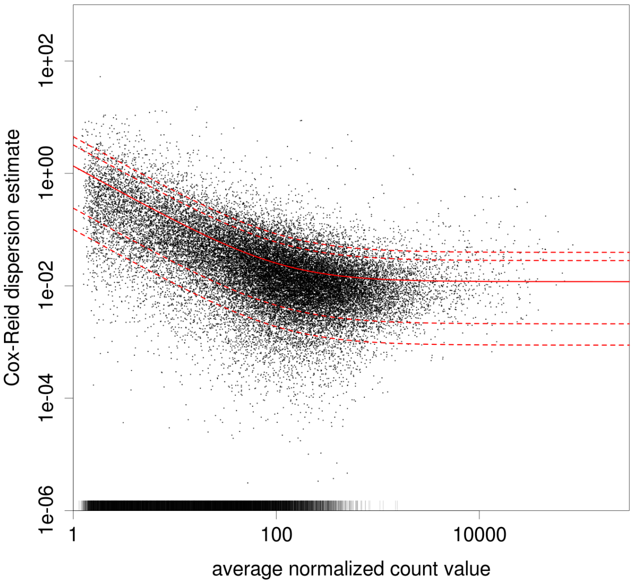

Figure 2: Dependence of dispersion on the mean. Each dot corresponds to one counting bin in the data of Brooks *et al.* (2010), the x axis denotes the normalized count, averaged over all samples, and the y axis shows the estimate of the dispersion. The bars at the bottom denote dispersion values outside the plotting range (in particular, those cases where the sample dispersion is close to zero). The solid red line shows the regression line, the dashed lines mark the 1-, 5-, 95- and 99-percentiles of the χ^2 distribution with 4 degrees of freedom scaled to have the fitted mean.

by not 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 likelihood for the parameter of interest (here: the dispersion) by dividing out a term containing the Fisher information for the other parameters as an approximation to conditioning on the profiled-out parameters. This works 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 McCarthy *et al.* (2012). The authors have been using it in their *edgeR* package (Robinson *et al.*, 2010*a*) since September 2010 (version 1.7.18). We make use of this approach to estimate the dispersion for each counting bin; details are provided in Supplementary Note S.4.

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 μ_{ijl} for all the counting bins 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

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

or counting bins. 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. This is possible because we test for an interaction effect. If the aim were to test for a main effect such as differential expression, dispersion estimation would need to be based on model (2).

We fit the model (3) for each gene i separately and use the Cox-Reid dispersion estimator of McCarthy *et al.* (2012), as described above, to obtain a dispersion value $\hat{\alpha}_{il}$ for each counting bin l in the gene.

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 counting bin has a large sampling variance. A commonly used solution is to share information across estimators (Tusher *et al.*, 2001; Lönnstedt and Speed, 2002). We noted that there is a systematic trend of dispersions as a function of the mean, and consider the relationship

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

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.) To obtain the coefficients a_0 and a_1, 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 robustify the fit, we iteratively leave out bins with large residuals until convergence is achieved (Huber, 1981).

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 counting bins, the difference between the sample estimate $\hat{\alpha}_{il}$ and the fitted value $\alpha(\hat{\mu}_{il})$ is compatible with a χ^2 sampling distribution (indicated by the dashed lines). Nevertheless, there are sufficiently many bins with a sample 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 values. Hence, for the ANODEV (see below) we use as dispersion value α_{il} the maximum of the per-bin estimate $\hat{\alpha}_{il}$ and the fitted value $\alpha(\hat{\mu}_{il})$. On average, this overestimates the true dispersion, and costs power, but we consider this preferable to using either only the fitted values or the sample estimates, both of which carry the risk of producing many undesirable false positives. More sophisticated alternatives for this step, which usefully interpolate between the two extremes, and perhaps incorporate further covariates besides μ, might become available in the future.

Analysis of deviance We test for each counting bin whether it is differentially used between 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 i a reduced model with no counting-bin–condition interaction

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

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

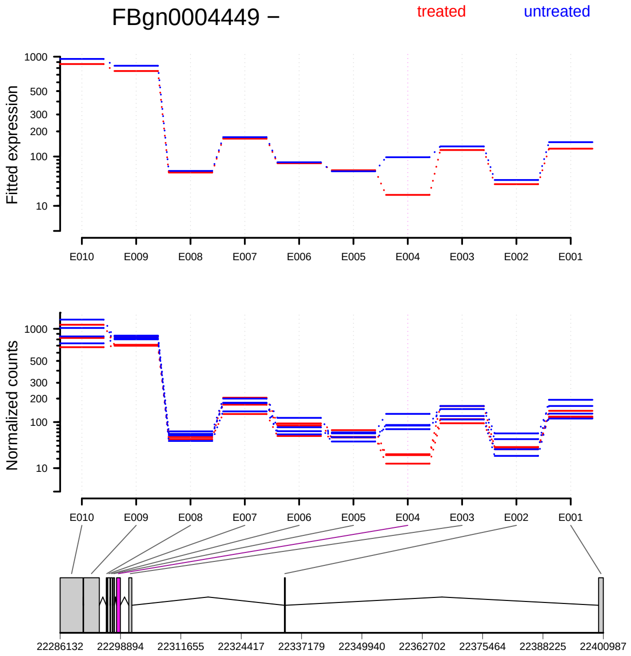

Figure 3: The treatment of knocking down the splicing factor *pasilla* affects the fourth exon (counting bin E004) of the gene *Ten-m* (CG5723). The top panel shows the fitted values according to the linear model, the middle panel shows the normalized counts for each sample, and the bottom panel shows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

and, separately for each bin l' of gene i, a model with an interaction coefficient for *only* this bin, but as in Equation (5), main effects for all bins l,

$$\log \mu_{ijl} = \beta^{\mathrm{G}}_i + \beta^{\mathrm{E}}_{il} + \beta^{\mathrm{S}}_{ij} + \beta^{\mathrm{EC}}_{i\rho_j l}\delta_{ll'}. \tag{6}$$

Here, $\delta_{ll'}$ is the Kronecker delta symbol, which is 1 if $l = l'$ and 0 otherwise. We compute the likelihood of these models using the dispersion values α_{il} as estimated from model (3), with the information-sharing scheme of presented earlier. Comparing the fit (6) for counting bin l' of gene i with the fit (5) for gene i, we get an analysis-of-deviance p value $p_{il'}$ for each counting bin by means of a χ^2 likelihood-ratio test. Note that we test against the null hypothesis that *none* of the conditions influences exon usage, and hence, if there are more than two different conditions ρ, we aim to reject the null hypothesis already if any one of the conditions causes differential exon usage.

Differential exon usage, as treated here, cannot be distinguished from overall differential expression of a gene if the gene only consists of a single counting bin or if all but one of its counting bins have 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 speed up computation. Such filtering can also improve power (see Bourgon *et al.* (2010)).

Additional covariates The flexibility of GLMs makes it easy to account for further covariates. For example, if in addition to the experimental condition ρ_j

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

we wish to account for a further covariate τ_j, we extend model (3) as follows:

$$\log \mu_{ijl} = \beta^{\mathrm{G}}_i + \beta^{\mathrm{E}}_{il} + \beta^{\mathrm{S}}_{ij} + \beta^{\mathrm{EB}}_{i\tau_j l} + \beta^{\mathrm{EC}}_{i\rho_j l}, \tag{7}$$

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

An example is provided in the next section with Equation (9).

Visualization

The *DEXSeq* package offers facilities to visualize data and fits. An example is shown in Figure 3, using the data discussed in the next section. 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

$$\mu_{ijl} = \exp\left(\tilde\beta^{\mathrm{G}}_i + \tilde\beta^{\mathrm{E}}_{il} + \tilde\beta^{\mathrm{C}}_{i\rho_j} + \tilde\beta^{\mathrm{EC}}_{i\rho_j l}\right). \tag{8}$$

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 S.3.

For genes with differential overall expression, it can be difficult to see the evidence for differential exon usage in a plot based on Equation (8). For these cases, the software offers the option to average over the expression effects. Supplementary Figure S1 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 seem 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 S.5 for details.

Applications of the method

Analysis of the data set 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 above, 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).

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

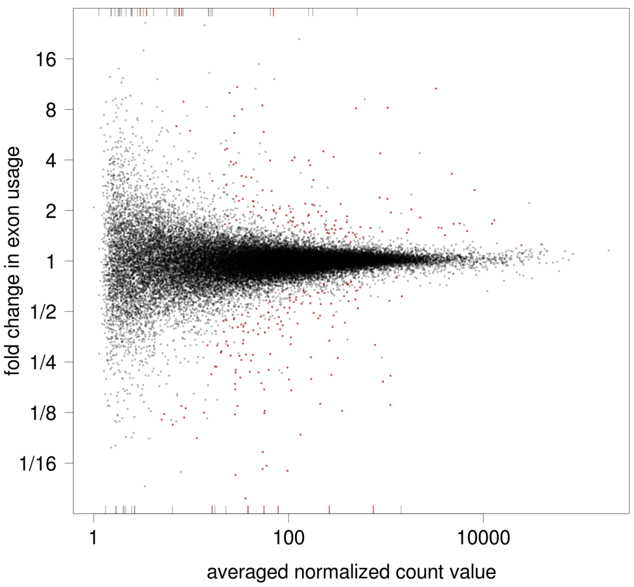

Figure 4: Fold changes of exon usage versus averaged normalized count value for all tested counting bins for the Brooks et al. data. Red colour indicates significance at 10% FDR. Bars at the margin represent bins with fold changes outside the plotting range.

Here, since we have a mixture of single-end and paired-end libraries, we extended Equation (3) to account for this additional covariate

$$\log \mu_{ijl} = \beta_i^{\mathrm{G}} + \beta_{il}^{\mathrm{E}} + \beta_{ij}^{\mathrm{S}} + \beta_{i\rho_j l}^{\mathrm{EC}} + \beta_{i\tau_j l}^{\mathrm{ET}}, \tag{9}$$

where $\tau_j = 1, 2$ is the library type of sample j, single-end or paired-end.

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. 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 the context of Equations (5) and (6) 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 the 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: 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. This is

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

a consequence of the fact that the coefficient of variation (CV) of the count values decreases with their mean, as explained in more detail in Supplementary Note S.2.

Analysis of the chimpanzee data by Brawand et al.

While the preceding application was on a controlled experiment with a cell culture under sharp treatment, in this section we analyse data from an observational study with complex subject-to-subject variation (Brawand *et al.*, 2011). This data set includes RNA-Seq from prefrontal cortex samples from 5 chimpanzees and cerebellum samples from 2 further chimpanzees. We used DEXSeq to test for exon usage differences between these two brain tissue types.

We aligned the RNA-Seq reads (GEO accessions GSM752664–GSM752671) from these samples to the chimpanzee genome (CHIMP2.1.4 from Ensembl 64) using GSNAP 2012-01-11 (Wu and Nacu, 2010). Prior to alignment, we trimmed all reads to a common length of 76 nt, single-ended. The trimming was necessary to make the data comparable across samples; *DEXSeq* itself has no length limitation and can deal with any read length.

At 10% FDR, *DEXSeq* found significant differential exon usage for 866 counting bins in 650 genes. The result table, with plots for all genes with significant differential exon usage, can be found at at `http://www-huber.embl.de/pub/DEXSeq/chimp/testForDEU.html`. Exploration of this hit list reveals interesting differences between the tissues. For example, one of the top hits, the gene *PRKCZ* (protein kinase C zeta; ENSPTRG00000000042) expresses its first four exons only in cerebellum but not in the prefrontal cortex (Supplementary Figure S4). Inspecting the *Pfam* (Finn *et al.*, 2010) and *SMART* (Letunic *et al.*, 2012) databases of protein domains reveals that these four exons encode the heterodimerization domain *PB1*. This suggests the hypothesis that the gene product loses its ability to bind to its partner protein in the prefrontal cortex. Indeed, a literature search revealed that these two isoforms are well studied (reviewed by Hirai and Chida (2003)). The long isoforms protein product, *PKZζ*, is widely expressed and is activated by a second messenger, *PARD6A*, which removes the protein's autoinhibition by binding to the PB1 domain. The truncated protein, denoted *PKMζ*, is specific to the brain and, due to the lack of the *PB1* domain, constitutively active. It plays a major role in long-term potentiation and memory formation. In this context, it is noteworthy that, as our analysis shows, its expression is confined to certain brain regions.

Another example is provided by gene *PLCH2* (phospholipase C eta 2; ENSPTRG00000000051), for which DEXSeq indicated differential usage of counting bin E011 (fourth exon). According to *SMART* and *Pfam*, this exon contains an *EF hand*, a calcium binding helix-loop-helix motif. Here, we are not aware of prior work on the isoform(s) lacking this exon. We can speculate that the shorter isoform's activity might no longer depend on calcium concentration, on which *PLCH2*'s enzymatic activity normally depends strongly (Nakahara *et al.*, 2005). Furthermore, Zhou *et al.* (2008) studied the activation of *PLCH2* by G$\beta\gamma$ complexes and found that the *EF hand* domain of *PLCH2* is required for this interaction. Another hypothesis might hence be that a functional consequence of the observed tissue-specific usage of fourth exon is a modulation of the regulation of *PLCH2* by G proteins.

For gene ENSPTRG00000000130, *DEXSeq* reports increased usage of the second exon in the cerebellum and of the second-to-last exon in the prefrontal cortex. This gene codes for precortistatin, a protein that gets cleaved to give rise to the neuropeptid cortistatin, which (in human) comprises the last 17 aa of the full protein's C-terminus (de Lecea *et al.*, 1997), which are contained in the last exon. While the overall expression differences seen in the data agree with the known main location of cortistatin, the cortex, the observed differential exon usage is more difficult to interpret: the affected parts of the protein are considered non-

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

functional. Nevertheless, presence or absence of parts could affect the efficiency of the cleavage process or the stability of the mRNA, to coregulate the tissue-specific expression.

These three examples illustrate how a *DEXSeq* analysis can serve as a starting point for hypothesis formation. We picked these three genes by inspecting the first ten genes with significant differential exon usage, as sorted by numerical Ensembl gene ID (not by p value); that is, in essence we inspected a random subset of 10 hits. The richness of the biology seen indicates that many novel insights into gene function and regulation may be expected from the analysis of tissue specific isoform usage patterns.

Comparison of human cell lines

As a third application, we present a comparison between two human cell lines. The ENCODE Project Consortium (2011) performed RNA-Seq experiments for a number of human cell lines, of which we chose H1 human embryonic stem cells (h1-hESC) and human umbilical vein endothelial cells (HUVEC) (Laboratory of B. Wold; sequenced with 76 nt paired-end reads; GEO accessions GSM758573 and GSM767856), because they were performed in biological duplicates. Such a comparison offers high detection power because of the typically small within-group variability that one may expect for untreated cells and the many differences between these two cell lines. In fact, we find 7,795 genes to be affected by differential exon usage, which can be seen in the report generated by *DEXSeq*, available at `http://www-huber.embl.de/pub/DEXSeq/encode/testForDEU.html`. For a plot of exon usage fold change, see Supplementary Figure S5 and for an example of a differentially spliced gene, see Supplementary Figure S6. Since the cell lines were derived from different subjects, the many observed differences could be due both to the difference in cell type and to differences in their genetic background.

Discussion

Importance of modelling overdispersion

The method presented here differs from previous work by using an error model that accounts for sample-to-sample variation in excess of Poisson variation. In the following, we investigate whether this extra variation is important enough to influence results in practice.

To address this question for our inference procedure, we re-computed the tests for differential exon usage for the Brooks et al. data 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 S2 and compare with Figure 4.) For these extra hits, however, the treatment effect was not large compared to the variation seen between replicates, and the data not provide 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.

Analyses based on Fisher's test

To test 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.,

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

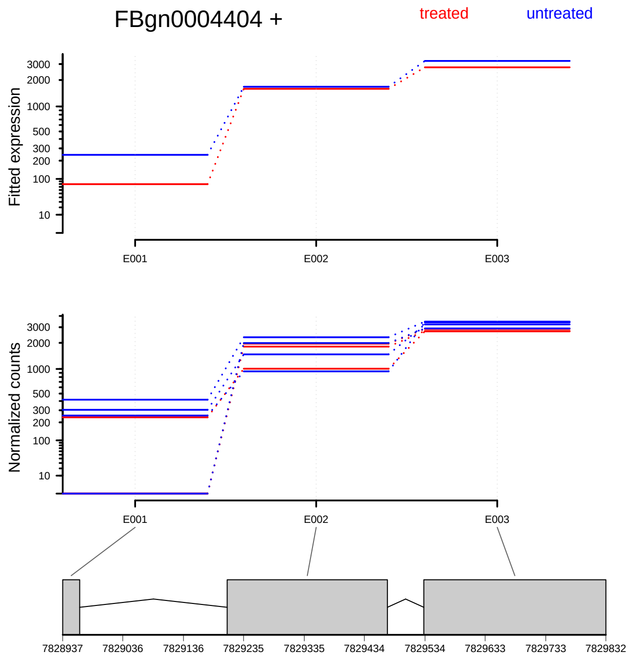

Figure 5: Ribosomal protein gene RpS14b (from the Brooks et al. data) is shown here as an example for a gene with heterogeneous dispersion. The first exon has zero count in the paired-end samples *untreated 2*, in the single-end sample *treated 2* and in the paired-end sample *treated 3*, and large non-zero counts in the four other samples. Colours are as in Figure 3.

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 to a single sample, while Brooks *et al.* (2010) summed up the number of reads from their replicates. The MISO method (Katz *et al.*, 2010) proposed a different way of setting up the contingency table. In all cases, the contingency tables did not contain information on sample-to-sample variability (Baggerly *et al.*, 2003) and so, the results are expected to contain an inflated number of false positives.

As an example, Supplementary Figure S3 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, the data do not support a significant effect of the treatment.

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 discussed in the Section on information sharing assures this.

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 1.3.0 of *cuffdiff*. With nominal FDR control at 10%, *cuffdiff* reported differential splicing for only 37 genes, and so showed less power than our approach.

To test the control of false-positive rates, we made use of the fact that there were four replicates for the untreated condition. We formed one group from samples 1 and 3 and another group from samples 2 and 4. We tasked both *DEXSeq* and *cuffdiff* with comparing between the two groups at a nominal FDR of 10%. As this is a comparison between replicates, ideally no significant calls should be made. 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 mock comparison, *DEXSeq* found only a few, namely 8 genes significant, as expected, compared to 159 in the comparison of treatment versus control. Surprisingly, *cufflinks* found many more genes in the mock comparison than in the proper between-groups comparison, namely 455 genes. Supplementary Note S.6 describes further tests, which confirmed *cufflinks*'s difficulty with providing type-I error control in this data set.

We also performed the same type of comparison on a data set with quite different characteristics and experimental design, the chimpanzee data of Brawand et al. In a comparison of the 6 chimpanzee prefrontal cortex (PFC) samples with the 2 cerebellum samples, *cuffdiff* 1.3.0 reported 108 genes at 10% FDR, again showing less power then *DEXSeq* (700 genes, see above).

We then used the 5 PFC samples from male chimpanzees to assess type-I error rates. Both tools were tasked to compare any combination of two samples versus two other samples. *DEXSeq* in each case found substantially fewer genes in these mock comparisons than in the proper comparison (with one exception, always less than 1/65). *Cuffdiff*, however, each time found at least twice as many genes in the mock comparisons than in the proper one. For details, see Supplementary Note S.6.

Also note Supplement II, which contains the exact commands used for all

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

computations performed for this paper.

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 n exons, and a short one consisting of only the first $n/2$ exons. If an experimental condition increases the number long transcripts on the expense of the short ones, without changing the total number, one might expect an analysis to indicate differential usage for the last $n/2$ exons. However, our method cannot distinguish this situation from one where the gene is overall down-regulated, while the first $n/2$ exons are 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 a small fraction of counting bins in the 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 commented on *cuffdiff* in the preceding section.

Apart from the lack of tools for inferring differential expression at the transcript level, there can be concrete advantages in per-exon analysis. If, for example, several transcripts have most exons in common and differ by only a few exons, their abundance estimates will contain substantial correlated uncertainties that reduce the power for inference of differential expression. The remedy would be to disregard the reads which inform about the shared parts of the transcripts and to focus on those reads in which they differ. Hence, an exon-centric analysis might be a crucial component even of a transcript-level method.

In addition, it is not clear that inference about transcripts is always more useful for biological interpretation than inference at the per-exon level. After all, we have knowledge about the functional differences of multiple translated isoforms of a gene for only a small number of proteins. If currently a researcher finds that a gene of interest expresses different transcripts in different conditions, her further analysis will typically start with assessing the difference between the two transcripts, seeing, for example, that they differ in the presence of certain exons and asking which regulatory signals or functional domains these exons may contain. Therefore, we expect that a method such as ours that pinpoints the location of the differences by focusing on specific exons will be valuable for biological interpretation, and sometimes perhaps more valuable than a transcript-centric approach. This expectation is supported by the three example hits discussed in the analysis of the chimpanzee data. A next step will be to leverage in a systematic and automated way databases with annotation for parts of gene products, e. g. information on protein domains provided by resources such as *Pfam* (Finn *et al.*, 2010), *SMART* (Letunic *et al.*, 2012) and *Prosite* (Sigrist *et al.*, 2010), or predicted miRNA target sites.

Junction reads

Junction reads are reads whose genomic alignment contains a gap because they start in one exon, end in another exon, and “jump” over the intron in between and possibly over skipped exons. In *DEXSeq*, such reads are counted for each counting bin with which they overlap, i.e., they appear multiple times in the count table. However, as we test for each exon separately, this does not affect the validity of the test.

Junction reads contain additional information that is especially valuable when inferring gene models and the positions of splice junction. Unless one works with a very well annotated model system, this information should be used when defining the counting bins, by parsing the spliced alignments with appropriate tools.

Furthermore, junction reads give evidence for connections between counting bins and so are crucial for isoform deconvolution tools such as *cufflinks* and *MMSeq*. For our exon-by-exon test, however, leveraging this information is not essential, and also not straight-forward. In the presented method, we essentially consider for each sample the ratio of the number of reads overlapping with an exon to the number of read falling onto the whole gene. Alternatively, one could consider the ratio of the number of reads skipping over the exon under consideration to the total count. We anticipate that the latter would offer a moderate increase in power in cases where the counting bin is much shorter than the typical read length. It may be an interesting future extension to the *DEXSeq* method to switch to this scheme for bins that are short compared to the read length.

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

Implementation

We implemented *DEXSeq* as a package for 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. After the analysis has been performed all the results are available, together with the input data, in a object of derived from the *ExpressionSet* class, *Bioconductor*'s standard container type for data from high-throughput assays. The results provided include for each counting bin the following data: the conditional-maximum-likelihood estimate for the dispersion, the dispersion value actually used in the test (which may be different, due to the information sharing across genes), the p value from the test for differential exon usage, the Benjamini-Hochberg-adjusted p value, and the fit coefficients describing the fitted $\log_2$ fold change between treatment control (or, if there are more than two conditions, for pairs of conditions as chosen by the user). Other *R* or *Bioconductor* functionality can be used for downstream analyses of these results. If required, the other coefficients as described in Supplementary Note S.3 are also available.

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 and Supplementary Figures S1 and S3. The HTML output allows interactive browsing of the results 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

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

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.*, 2012) from the *statmod* package, which offers better performance and convergence than 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 several hours. However, the method lends itself easily to parallelization: we use the *multicore* package (Urbanek, 2011) to distribute the computation on several CPU cores.

The complete workflow used to perform all calculations for this paper are documented in Supplement II.

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. We have demonstrated *DEXSeq* on three data sets of different type and illustrated how the results of a *DEXSeq* analysis, combined with metadata on parts of transcripts, such as protein domains, form the basis for exploring a biological phenomenon, differential exon usage, that is currently not well understood and whose study may reveal many surprises.

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)

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A. *et al.* 2011 The evolution of gene expression levels in mammalian organs. *Nature*, **478**(7369), 343–348. (doi:10.1038/nature10532)

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)

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

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

Cline, M. S., Blume, J., Cawley, S., Clark, T. A., Hu, J.-S., Lu, G., Salomonis, N., Wang, H. and Williams, A. 2005 ANOSVA: a statistical method for detecting splice variation from expression data. *Bioinformatics (Oxford, England)*, **21 Suppl 1**, i107–15. (doi:10.1093/bioinformatics/bti1010)

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.

de Lecea, L., Ruiz-Lozano, P., Danielson, P. E., Peelle-Kirley, J., Foye, P. E., Frankel, W. N. and Sutcliffe, J. G. 1997 Cloning, mRNA expression, and chromosomal mapping of mouse and human preprocortistatin. *Genomics*, **42**(3), 499–506. (doi:10.1006/geno.1997.4763)

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)

ENCODE Project Consortium 2011 A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biology*, **9**(4), e1001 046. (doi:10.1371/journal.pbio.1001046)

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G. *et al.* 2010 The Pfam protein families database. *Nucleic acids research*, **38**(Database issue), D211–22. (doi:10.1093/nar/gkp985)

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)

Hirai, T. and Chida, K. 2003 Protein Kinase Czeta (PKCzeta): Activation Mechanisms and Cellular Functions. *Journal of Biochemistry*, **133**(1), 1–7. (doi:10.1093/jb/mvg017)

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)

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

Huber, P. J. 1981 *Robust statistics.* Wiley.

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)

Katz, Y., Wang, E. T., Airoldi, E. M. and Burge, C. B. 2010 Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, **7**(12), 1009–1015. (doi:10.1038/nmeth.1528)

Letunic, I., Doerks, T. and Bork, P. 2012 SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Research*, **40**, D302–D305. (doi: 10.1093/nar/gkr931)

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. 2012 Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research.* Advance online publication. (doi:10.1093/nar/gks042)

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

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B. 2008 Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, **5**(7), 621–8. (doi:10.1038/nmeth.1226)

Nakahara, M., Shimozawa, M., Nakamura, Y., Irino, Y., Morita, M., Kudo, Y. and Fukami, K. 2005 A novel phospholipase C, PLCη2, is a neuron-specific isozyme. *The Journal of Biological Chemistry*, **280**(32), 29 128–34. (doi:10.1074/jbc.M503817200)

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)

Purdom, E., Simpson, K. M., Robinson, M. D., Conboy, J. G., Lapuk, A. V. and Speed, T. P. 2008 FIRMA: a method for detection of alternative splicing from exon array data. *Bioinformatics*, **24**(15), 1707–14. (doi:10.1093/bioinformatics/btn284)

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 Oshlack, A. 2010 A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, **11**(3), R25. (doi:10.1186/gb-2010-11-3-r25)

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)

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

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)

Sigrist, C. J. A., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A. and Hulo, N. 2010 PROSITE, a protein domain database for functional characterization and annotation. *Nucleic acids research*, **38**(Database issue), D161–6. (doi:10.1093/nar/gkp885)

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)

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)

Wu, T. D. and Nacu, S. 2010 Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**(7), 873–81. (doi:10.1093/bioinformatics/btq057)

Zhou, Y., Sondek, J. and Harden, T. K. 2008 Activation of human phospholipase C-η2 by G$\beta\gamma$. *Biochemistry*, **47**(15), 4410–7. (doi:10.1021/bi800044n)

Anders, Reyes, Huber:
Detecting differential usage of exons from RNA-Seq data

Supplement

Supplementary Notes

S.1 Normalization for sequencing depth

To normalize the read counts for sequencing depth, we use the method that we introduced for *DESeq* (Anders and Huber, 2010) and which we describe here in more detail.

As before, we denote by K_{ijl} the number of reads mapping to counting bin l of gene i in sample j. For a counting bin (il) that is not subject to differential exon usage or differential expression, we assume that the expectation values $\mathrm{E}(K_{ijl})$ in the different samples j are –to sufficient approximation– proportional to each other, with proportionality factors that only depend on j, but not on i and l. This assumption is natural given the random sampling that underlies the shotgun sequencing of cDNA. The assumption is also supported by many data sets; if deviations were observed, we could relate them to data quality problems or “batch effects”. We denote the proportionality factor associated with sample j by $\tilde{s}_j$. For instance, if $\tilde{s}_2/\tilde{s}_1 = 1.5$, then we expect the counts for the bins that are not subject to differential exon usage or differential expression to be 1.5 times higher in sample 2 than what we expect for sample 1.

We first note that the total number of aligned reads in sample j is not a good estimator for $\tilde{s}_j$. This is because often a sizable fraction of the total number of reads originate from a small number of strongly expressed genes, and if these are differentially expressed, the estimate will be biased away from the true value. Thus, when testing for differential expression it is inadvisable to normalize read counts by dividing by the total number of reads, as done for example in the *RPKM* measure of Mortazavi *et al.* (2008). This was pointed out by Anders and Huber (2010) and independently by Robinson and Oshlack (2010).

A better alternative for an estimator of $\tilde{s}_j/\tilde{s}_{j'}$ is the median of all observed ratios,

$$\operatorname*{median}_{i,l} \left\{ k_{ijl}/k_{ij'l} \right\}. \tag{S1}$$

The fact that expression (S1) is computed over all counting bins, hence also over potentially differentially expressed ones, is harmless if there are enough non-differential counting bins, since the median will not be affected by a few outliers. The expression (S1) works well for a pair of samples, but what do we do for experiments with more than two samples? Computing (S1) cannot assure transitivity, i. e. the scaling factor determined in this way between samples j and j'' would not be the ratio of those for j and j' and for j' and j''. A more elegant solution is possible. We can construct a *virtual reference* sample by computing for each counting bin (il) the geometric mean of the counts of all samples:

$$k_{il}^{\mathrm{ref}} = \left(\prod_{j=1}^{m} k_{ijl} \right)^{1/m}. \tag{S2}$$

Then, we assign to each sample j a *size factor* s_j as the median of the ratio of this sample's counts to the virtual reference,

$$\hat{s}_j = \operatorname{median} \left\{ \frac{k_{ijl}}{k_{il}^{\mathrm{ref}}} \,\middle|\, i,l:\ k_{il}^{\mathrm{ref}} > 0 \right\}. \tag{S3}$$

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

Using the geometric mean in Equation (S2) rather than, say, the arithmetic mean, ensures that the scaling ratio of two samples j, j', now given by $\hat{s}_j/\hat{s}_{j'}$ is close to the expression (S1) discussed first.

Finally, we note that since we only care about ratios of size factors, we multiply them with a suitably chosen constant c and define $s_j = c\,\hat{s}_j$ such that $\prod_{j=1}^{m} s_j = 1$. This keeps the size factors close to unity and ensures that normalized count values k_{ijl}/s_j remain close to their original values, making their interpretation, e. g. in plots, easier.

S.2 Motivation of the use of 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 here we briefly summarise their argumentation.

Let us denote by $\tilde{Q}_{ijl}$ the concentration of cDNA fragments mapping to counting bin l of gene i in sample j. Then we can assume that the number of counts K_{ijl}, conditioned on $\tilde{Q}_{ijl}$, is Poisson-distributed with mean $s_j \eta_{il} \tilde{Q}_{ijl}$. The Poisson distribution follows from the fact that the probability that a given fragment is sequenced is small, is independent of which other fragments are sequenced, and depends only on its availability and the total number of reads produced. The factor η_{il} represents detection efficiency, which can depend on factors such as GC content, length, secondary structure or the like. The important point is that η_{il} is determined by the identity of gene i and exon l and may hence be assumed to not depend on the sample j. We can absorb η_{il} by defining the *effective concentration* $Q_{ijl} = \eta_{il}\tilde{Q}_{ijl}$. $\tilde{s}_j$ is a size factor as discussed in Note S.1.

As variance and mean are equal in a Poisson distribution, we have

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

and, by the law of total variance,

$$\operatorname{Var}\left(K_{ijl}\right) = s_j E(Q_{ijl}) + s_j^2 \operatorname{Var}(Q_{ijl}). \tag{S4}$$

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 models Q_{ijl} with 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 $\tilde{\mu}$ of a NB distribution is commonly parametrized as $v = \tilde{\mu} + \alpha\tilde{\mu}^2$, where the constant α is known as the *dispersion parameter*. Comparing this relation with Equation (S4) (with $\tilde{\mu} = \operatorname{E}(K_{ijl})$ and $v = \operatorname{Var}(K_{ijl})$) shows that the dispersion parameter for K_{ijl} can be interpreted as the squared coefficient of variation (SCV) of the effective concentration Q_{ijl}.

It is instructive to also consider the SCV of the count value K_{ijl}, which can be found (using Equation (S4)) to be

$$\operatorname{SCV}\left(K_{ijl}\right) = \frac{\operatorname{Var}(K_{ijl})}{\left(\operatorname{E}(K_{ijl})\right)^2} = \frac{1}{\operatorname{E}(K_{ijl})} + \alpha$$

This shows that the SCV of a negative binomial distribution can be decomposed into two terms, the first corresponding to a Poisson noise component and the second to overdispersion, i. e., a noise component that is in excess of the Poisson noise. If we use the parametrization $\alpha(\mu) = a_1/\mu + a_0$ (Equation (4)) for the dependence of α on the normalized mean count $\mu = \mathrm{E}(K_{ijl}/s_j) = \mathrm{E}(Q_{ijl})$, we can write the SCV as

$$\mathrm{SCV}\,(K_{ijl}) = \frac{1 + s_j a_1}{\mathrm{E}(K_{ijl})} + a_0.$$

Hence, the coefficient a_0 in our parametrization is the asymptotic value of the dispersion and of the SCV for large count values and a_1 causes noise with a dependence on the mean that is proportional to that of the Poisson noise.

S.3 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 (8), 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 counting-bin and counting-bin–condition–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. (A counting bin 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 counting bin 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 counting bin l, averaged over all conditions,

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

(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. (See Section *Visualization* in the description of the *DEXSeq* method.)

S.4 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.

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

Group 1	Group 2	DEXSeq 1.1.5	cuffdiff 1.1.0	cuffdiff 1.2.0	cuffdiff 1.3.0
proper comparisons, treatment (knock-down) vs control:					
T1 – T3	C1 – C4	159	145	69	50
T1, T2	C2, C3	52	323	120	578
mock comparisons, control vs control:					
C1, C3	C2, C4	8	314	650	639
C1, C4	C2, C3	7	392	724	728

Table S1: Results of the comparison for the Brooks et al. data.

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 an 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.

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

S.5 Variance stabilizing transformation

To achieve the axis warping described in the main text, at the end of the Section on visualization, 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 $\tau(k_{ijl}/s_j)$ are 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 $\beta^{\mathrm{EC}}_{i,2,l} - \beta^{\mathrm{EC}}_{i,1,l}$ 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 $\tau\left(\exp\beta^{\mathrm{EC}}_{i,2,l}\right) - \tau\left(\exp\beta^{\mathrm{EC}}_{i,1,l}\right)$ gives more informative results.

S.6 Additional same-vs-same comparisons

Brooks et al. data

In the Applications section, we assessed the reliability of type-I error control in *DEXSeq* and *cuffdiff* by tasking both tools to perform a mock comparison of two versus two samples from the set of four control replicates. *Cuffdiff* found more genes to significantly differ in this mock comparison than in the comparison of treated versus control samples, and we concluded from this that *cuffdiff*'s assessment of biological variability was not working well for this data set. Here, we provide further results in support of this conclusion, summarised in Table S1. It shows for each comparison the number of genes reported as affected by differential exon usage (*DEXSeq*) or differential splicing (*cuffdiff*) at a nominal false discovery rate of 10%:

The table only shows two of the three possible 2-vs-2 mock comparisons, because we know that the third one is affected by a confounder, namely library type: samples C1 and C2 are single-end, samples C3 and C4 are paired-end. The mock comparisons shown in the table are balanced for library type, as is the 2-vs-2 proper comparison. The latter is instructive since it informs on the effect of sample size on the tools' performance. The table also includes the results from earlier versions of *cuffdiff*, indicating that the issue is not tied to a specific release.

The table indicates that *cuffdiff*, and in particular its most recent version, identifies a large number of differential splicing events in each of the mock comparisons, while it finds fewer in the proper comparisons. In contrast, *DEXSeq*'s discovery rates are consistent with the experimental design: few hits are found between replicates, while more are found for the proper comparisons.

While these results are instructive, they are from a single, small and specific data set, and as such should not be over-interpreted. In the next section, we repeat the analysis on a second data set, whose characteristics and experimental design are quite different; and interested readers are encouraged to perform similar benchmarks on their data sets.

Brawand et al. data

In the chimpanzee data set, there are five similar samples which we may consider replicates, namely the five prefrontal cortex (PFC) samples from males. We performed all 15 possible two-versus-two mock comparisons.

Table S2 shows, as before, the number of genes reported as affected by differential exon usage (*DEXSeq*) or differential splicing (*cuffdiff*) at a nominal false discovery rate of 10%. The table also includes the full comparison of all six PFC samples (including the one from a female chimpanzee) with the two cerebellum (CB) samples (which were taken from one male and one female animal).

The numbers of genes reported by *cuffdiff* in the mock comparisons were large, and overall were even higher than for the proper comparisons. In contrast, *DEXSeq* reported, with one exception, only small numbers in the mock comparisons, most of them less than a hundredth of the number of hits in the full comparison.

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

Group 1	Group 2	DEXSeq 1.1.5	cuffdiff 1.3.0
proper comparison, PFC vs CB:			
PFC 1 – PFC 6	CB 1, CB 2	650	114
PFC 1, PFC 2	CB 1, CB 2	56	230
PFC 1, PFC 3	CB 1, CB 2	18	361
PFC 1, PFC 4	CB 1, CB 2	26	370
PFC 1, PFC 5	CB 1, CB 2	32	215
PFC 1, PFC 6	CB 1, CB 2	27	380
mock comparisons, PFC vs PFC :			
PFC 1, PFC 3	PFC 2, PFC 4	3	405
PFC 1, PFC 2	PFC 3, PFC 4	0	399
PFC 1, PFC 4	PFC 2, PFC 3	244	590
PFC 1, PFC 3	PFC 2, PFC 5	2	628
PFC 1, PFC 2	PFC 3, PFC 5	1	499
PFC 1, PFC 5	PFC 2, PFC 3	2	555
PFC 1, PFC 4	PFC 2, PFC 5	2	460
PFC 1, PFC 2	PFC 4, PFC 5	2	504
PFC 1, PFC 5	PFC 2, PFC 4	2	308
PFC 1, PFC 4	PFC 3, PFC 5	10	497
PFC 1, PFC 3	PFC 4, PFC 5	5	554
PFC 1, PFC 5	PFC 3, PFC 4	0	353
PFC 2, PFC 4	PFC 3, PFC 5	1	476
PFC 2, PFC 3	PFC 4, PFC 5	10	823
PFC 2, PFC 5	PFC 3, PFC 4	0	526

Table S2: Results of the comparison for the Brawand et al. data.

Supplementary Figures

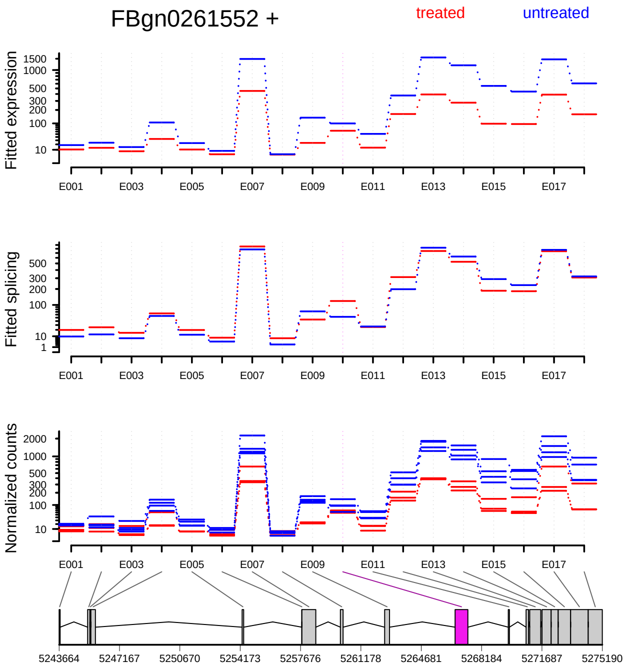

Figure S1: As pasilla is knocked down, its counts are lower in treatment then in control samples (first and third panel). This makes it difficult to see why *DEXSeq* detected differential exon usage for counting bin E010 (see highlight in the bottom panel). In the second panel, the data are shown in a different manner: the overall differential expression effect for the whole gene is removed (the per-condition expression coefficient $\beta^{C}_{i\rho_j}$ of the gene is replaced by its mean, see text for details), and the exon-specific effect for E010 is more apparent. Colours are as in Figure 3. The data suggest two possible biological interpretations: either, pasilla influences its own splicing, or the RNAi knockdown has different efficiency for the gene's different isoforms.

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

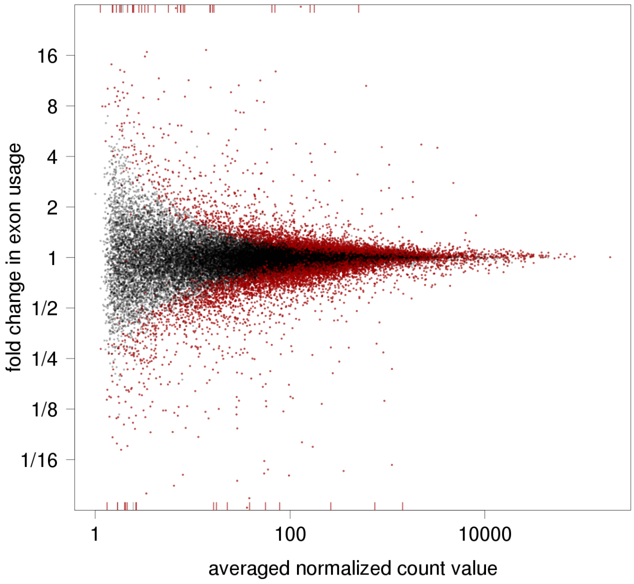

Figure S2: The same plot as in Figure 4, but with the red colour now indicating counting bins which appear to show significant differential exon usage when neglecting to account for biological variation in the test.

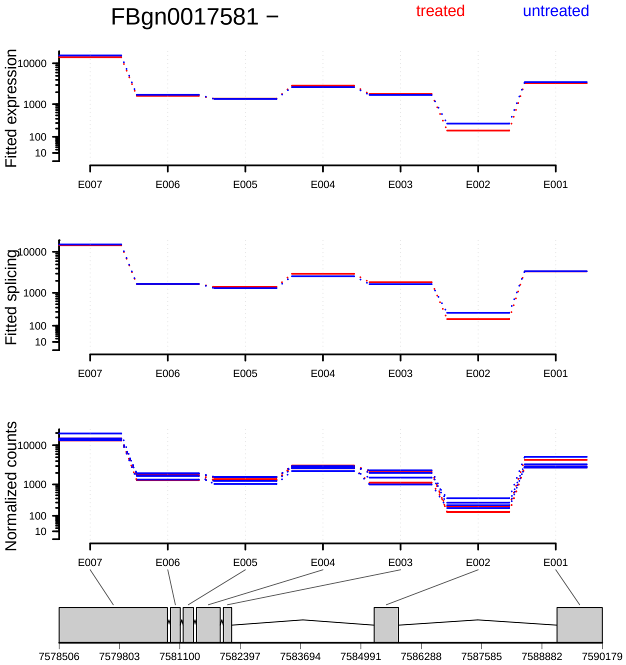

Figure S3: For this gene, *Lk6* (CG17342, in Brooks et al.'s annotation SG11207) Brooks et al. report a significant change in category *alternative first exon*. In fact, the usage of the two isoforms seems to change from sample to sample. However, due to the high within-group variation, the data do not support that this difference can be attributed to the treatment. Colours are as in Figure 3.

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

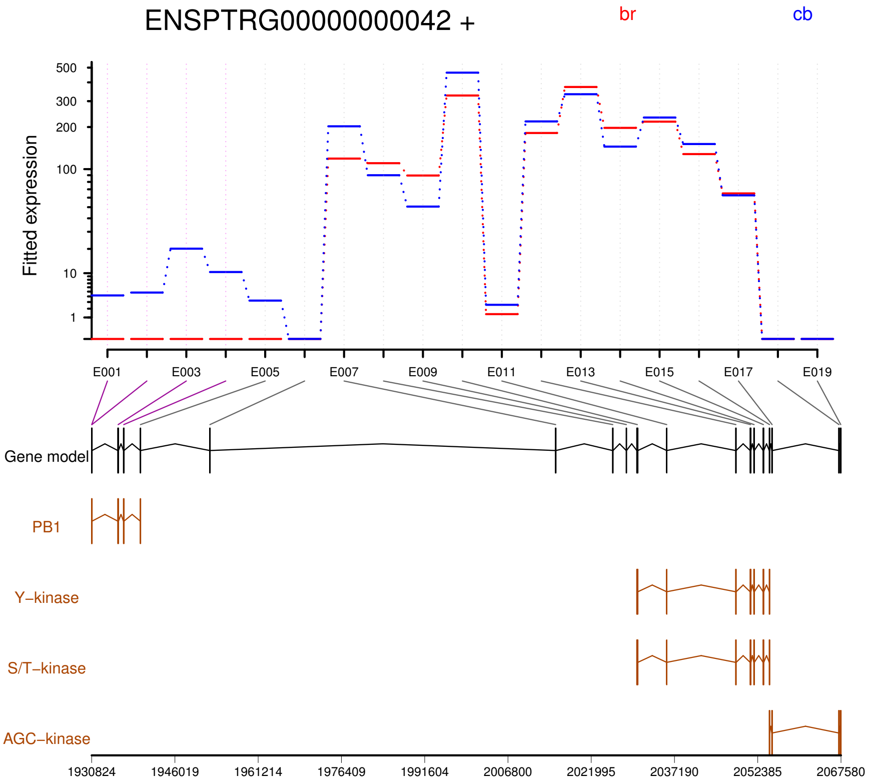

Figure S4: Protein kinase C ζ does not use its first four exons in the prefrontal cortex (red), while they are expressed in the cerebellum (blue). The brown rows below the flattened gene model show the protein domain annotation provided by the *SMART* database in the *splice variants* display function of the *Ensembl* genome browser. SMART indicates that the cerebellum-specific exons encode a *PB1* domain (SMART accession number SM00666), while the exons present in both tissues contain the catalytic domain of a tyrosine or serine/threonin protein kinase (SM00219 or SM00220). The domain to the very right is annotated as an extension to Ser/Thr-type kinases (SM00133).

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012

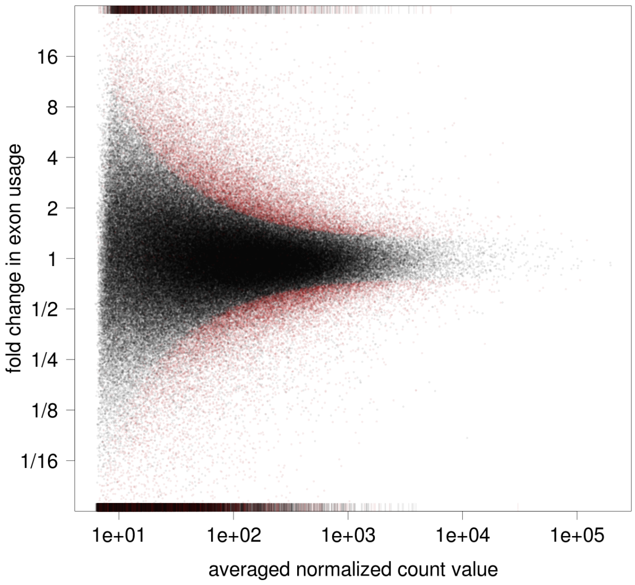

Figure S5: Plot of usage fold change versus average read counts, as in Figure 4, but here for the counting bins in the *Encode* data set, i.e., testing for changes in differential exon usage between HUVEC and H1-hESC cell lines.

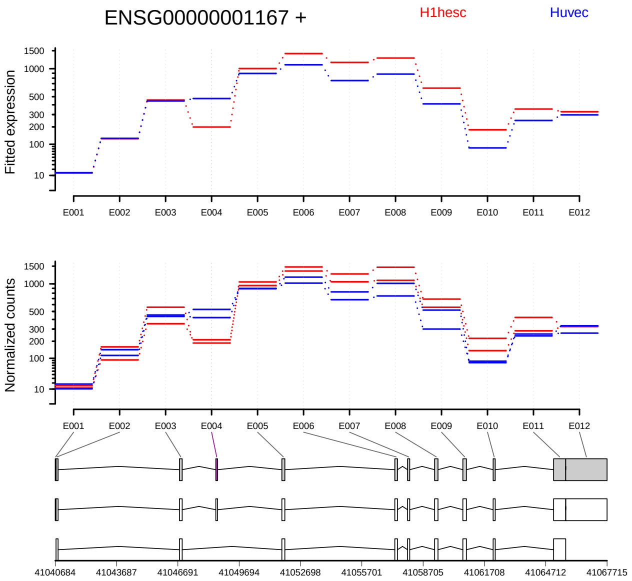

Figure S6: As an example for a result from the comparison of H1-hESC versus HUVEC cells, this figures show the gene NFYA (nuclear transcription factor Y alpha; ENSG00000001167). Its third exon, which has been found by *DEXSeq* as differentially used, is already annotated in *Ensembl* as a casette exon. (The two transcripts given by *Ensembl* are shown below the flattened gene model.) It is part of a glutamin-rich region (entry PS50322 in *Prosite*), formed by the first six exon.

Supplement II

Please find Supplement II at
http://www-huber.embl.de/pub/DEXSeq/Supplement_II_v2.html.

Nature Precedings : doi:10.1038/npre.2012.6837.2 : Posted 4 Apr 2012