Supplemental material:
Accurate Detection of Differential RNA Processing

Philipp Drewe,[1,2] Oliver Stegle,[3,4] Lisa Hartmann,[2,5] André Kahles,[1,2] Regina Bohnert,[2]
Andreas Wachter,[5] Karsten Borgwardt,[3,4,6] Gunnar Rätsch[1,2]

[1]Computational Biology Center, Sloan-Kettering Institute, 1275 York Avenue, New York, NY 10065, USA,
[2]Friedrich Miescher Laboratory of the Max-Planck Society, Spemannstrasse 39, 72076 Tübingen, Germany,
[3]Max Planck Institute for Intelligent Systems, Spemannstrasse 38, 72076 Tübingen, Germany,
[4]Max Planck Institute for Developmental Biology, Spemannstrasse 38, 72076 Tübingen, Germany,
[5]Center for Plant Mol. Biology, University of Tübingen, Auf der Morgenstelle 28, 72076 Tübingen, Germany,
[6]Center for Bioinformatics, University of Tübingen, Sand 14, 72076 Tübingen, Germany

S1 STATISTICAL MODELS FOR READ COUNTS

The probability density function of the Negative Binomial distribution $\mathcal{NB}(k|r,p)$ is given by:

$$\mathcal{NB}(k|r,p) = \binom{k+r-1}{k}(1-p)^r p^k, \tag{A1}$$

where $k \in \mathcal{N}_0$ is the number of observed reads, $r>0$ and $p \in (0,1)$. It can be seen as the distribution the sum of r independent random variable which follow a geometric distribution with parameter $1-p$. A more natural parameterization is in terms of mean and variances of, $\mathcal{NB}\mu,\sigma^2$, which will be used in the following. The exact parameterization follow from the relationships $\mu = \frac{rp}{(1-p)}$ and $\sigma^2 = \frac{rp}{(1-p)^2}$. Therefore, the effective variance can be written as a function of μ and r:

$$\sigma^2 = \mu + \frac{1}{r}\mu^2, \tag{A2}$$

showing how the limiting case of no overdispersion is realized for $r \to \infty$. In this case the Negative Binomial distribution converges to the poisson distribution. In order to simulate reads with biological variance we will use that the Negative Binomial distribution can also be seen as the distribution of a poisson variable which has a intensity that is gamma distributed. with $\alpha = r$ and $\beta = \frac{p}{(1-p)}$ (see for example (7)). This will allow to separate the variances of the sequencing which is supposed to be poisson and the biological variance.

S1.1 Accounting for fitted variance function

In general we assume that we have a set of counts $(c_g^r)_{r \in R}$ for regions j in a gene g in sample R as well as an estimate of the gene expression $(N^r)_{r \in R}$. We then compute a normalizing constant to decouple the splicing rate from the gene expression

$$s_g^r := \frac{|R| N_g^r}{\sum_{j \in R} N_g^j} \tag{A3}$$

After this we compute the set of normalize the counts $\hat{c}_{g,j}^R$ to

$$\hat{c}_{g,j}^r := \frac{c_{g,j}^r}{s_g^r} \tag{A4}$$

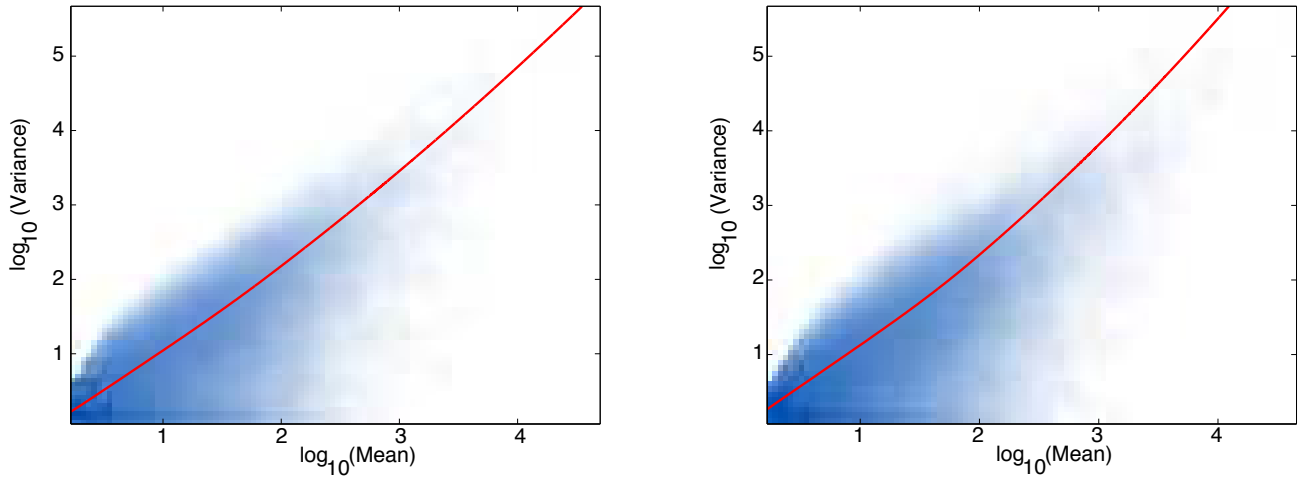
For each region j in each gene g we then compute the mean counts

$$\mu_{g,j} = \frac{1}{|R|} \sum_{r \in R} \hat{c}_{g,j}^r \tag{A5}$$

as well as the empirical variance:

$$\sigma_{g,j}^{2^R} = \sigma_{r \in R}^2(\hat{c}_{g,j}^r) \tag{A6}$$

Once we have the normalized counts we perform a local regression on the points $(\mu_{g,j}^R, \sigma_{g,j}^{2^R})$. By this we obtain a function mapping the empirical mean to the expected variance. This was done using the Locfit (27) package (See for details). It should be



(a) Variance fit on real data between two biological replicates

(b) Variance fit on real data between WT 0 h and WT 1 h

Figure S1. Log-Log plot of the variance fit on real data used for the simulation of the artificial dataset. The density of the empirical mean variance pairs is shown in (blue). The estimated variance function is shown in (red). In (A) The fit is shown for the biological replicate at 0h in (B) the fit is shown when considering WT 0h and 1 as biological replicates.

noted that one could also perform a higher dimensional variance estimation by including also a bias model or the length of the regions from which the counts have been used.

S1.2 Variance function estimation

Given a pairs of counts, normalized to account for gene expression, we estimated the variance function by fitting a function on the empirical mean and variance. We used the Locfit package that is part of Chronux 2.00 obtained from http://chronux.org. As parameters we used for bandwidth selection Mallows's CP criterion, local polynomials of degree two and gamma distribution as local likelihood function.

S1.3 Working without replicates

If replicate data is not available, conservative estimates of the variance function can be obtained from within-sample fits. Following (7), we consider the two samples A and B as replicates to fit the variance function. If there were not differential sites, this approximation would be fully legitimate, whereas in the presence of true differences we expect an over-estimation of the variance fits, leading to a conservative approximation. Another possibility is to use an estimated variance function from a similar sample as the ones under investigation.

S2 RDIFF

S2.1 rDiff.nonparametric

The test statistic for MMD is computed in two steps. First, all reads from both samples $A_g, B_g \subset \mathcal{X}_g$ for a gene g are mapped to a feature space $\mathcal{H}_g$ (a so-called "reproducing Kernel Hilbert Space") via a mapping function $\phi : \mathcal{X}_g \to \mathcal{H}_g$. Second, one computes the means of A_g and B_g in $\mathcal{H}_g$ by

$$\mu_r = \frac{1}{N^r} \sum_{i=1}^{N^i} \phi(\mathbf{x}_i^r), \quad r \in \{A_g, B_g\}, \tag{A7}$$

where $\mathbf{x}_i^r$ is the i-th example, i.e. read, in sample r and N^r is the cardinality of the set of reads ∇ for sample r. The test statistic is then the distance between these means of A_g and B_g (discrepancy) in the norm of $\mathcal{H}_g$, $D = \|\mu_{A_g} - \mu_{B_g}\|_{\mathcal{H}_g}$. The larger this distance, the less likely it is that both samples originated from the same distribution. p-values for the null hypothesis, of both samples being drawn from the same distribution, can be computed with a range of different strategies (see **(author?)** (29)). Here, we employ bootstrapping, where the reads are randomly shuffled among the two samples T times, computing the discrepancies

D_t for each permutation $t=1,\ldots,T$. Based on this empirical null discrepancy distribution we can compute the P-value for the actual observed discrepancy between the two samples. The corresponding P-value follows as

$$p_g = \frac{1}{T}\sum_{t=1}^{T}\mathbf{I}(D \leq D_t), \tag{A8}$$

where $\mathbf{I}(\text{true})=1$ and 0 otherwise. The minimal p-value that can be obtained by this strategy is limited by the inverse of the number of permutations and especially for low p-values the ranking for different genes is not very conclusive. To improve this ranking we resolved ties as described in Sec. S4.2.4.

There is a considerable freedom in designing the mapping function $\phi(\cdot)$. This choice is possible via the use of kernel functions $k(\cdot,\cdot)$. Kernels compute inner products between two elements, i.e., $k(\mathbf{x},\mathbf{x}')=\langle\phi(\mathbf{x}),\phi(\mathbf{x}')\rangle$, and can be used to efficiently deal with high dimensional representations of the elements. Here, each element, i.e. mapped read, is represented as a binary vector of length $|P|$, which is 1 at the positions where the read maps to and 0 otherwise. Therefore, the feature space is P dimensional and the mean computation in (A7) amounts to computing the average read coverage for each position given a set of reads. The MMD test strategy for this representation therefore boils down to testing whether the difference between the average read coverages is significantly greater than expected under a random assignment of all reads to samples.

S2.2 Biological variance The biological variance can be made allowance for by modeling samples of the null distribution to have a variance according to a given variance such as the one derived for rDiff.parametric. This is done by choosing the size of the random sample such that the variance induced by the subsampling matches the biological variance expected at each position. This was done as described in the following paragraphs.

Bootstrapping variance When drawing a subsample of n reads from the total of N^r reads the distribution of the mean coverage C_p^r at a position p follows a hypergeometric distribution $\mathcal{H}_N(N^r,n,C_p^r)$, where C_p^r is the fraction of reads covering the position p, N^r is the number of reads in the sample r and n^r is the size of a subsample. This is because we draw samples from a finite set without placing them back, which results in the aforementioned distribution. The variance $\sigma^{2^r}_{\text{subsample}}$ of the coverage of a subsample of size n^r is then given by:

$$\sigma^{2^r}_{\text{subsample}} = n^r \frac{C_p^r}{N}\frac{N^r-C_p^r}{N^r}\frac{N^r-n^r}{N^r-1} \tag{A9}$$

The variance of the read density reduces therefore to:

$$\sigma^{2^r}_{\text{subsample-density}} = n^r \frac{\frac{C_p^r}{N^r}\frac{N^r-C_p^r}{N^r}\frac{N^r-n^r}{N^r-1}}{(n^r)^2} \tag{A10}$$

$$= \frac{f_r(1-f_r)}{N^r-1}\frac{N^r-n^r}{n^r} \tag{A11}$$

where $f_r = \frac{C_p^r}{N^r}$ is the fractions of reads covering position p.

Matching the bootstrapping and biological variances In order to obtain null samples with variance $\sigma_{\text{biological variance}}$, we matched the two variances at a position p. To determine the necessary subsample size n^r for the variances to match, we solve the following equation for n^r:

$$\sigma^{2^r}_{\text{biological variance}} = \sigma^{2^r}_{\text{subsample-density}} \tag{A12}$$

$$\frac{f(C_p^r)}{(N^r)^2} = \frac{f_r(1-f_r)}{N^r-1}\frac{N^r-n^r}{n^r} \tag{A13}$$

We simplify by $c^r = \frac{f_r(1-f_r)}{N^r-1}$ which leads to a sample size n^r:

$$n^r = \frac{c^r f_r}{c^r + \frac{f(\text{median}_p(C_p^r))}{(N^r)^2}} \tag{A14}$$

In order to match the variances at not only one position of the coverage $C=C_A+C_B$ we define 10 equally sized bins of position $b_j, j\in\{1,\ldots,10\}$ where the coverage is in the same 10% quantile of positive coverage. For each of those bins and all samples r,

we determine a subsample rate n_j^r by matching the variances at the median of the Coverage in that bin. We then compute a new mean for the null distribution by:

$$\mu_r = \sum_{j=1}^{10} \frac{\sum_p C_p|b_j}{\sum_p C_p} \frac{1}{n_j} \sum_{r=1}^{n_j^r} \phi(\mathbf{x}^r_{\sigma(r)}|b_j), \tag{A15}$$

where σ is a permutation of $1,\ldots,N_A+N_B$.

S2.3 Alternative embeddings To further strengthen the information of the splice sites one can include this by the following mapping function. Let K be the number of observed introns in the reads, that is the number of unique pairs of intron starts and intron ends. Then we can define $\phi: \mathcal{X} \to \mathbb{R}^K$ as: $\phi(r)_i = 1$ if r supports intron i and 0 otherwise.

S3 ARTIFICIAL DATA SIMULATION

First we simulated the transcript expressions of each gene in both samples and replicated. For a gene g_j for each transcript $j \in \{1,\ldots,k\}$ we drew a relative intensity $e_i^j \in [0,1]$ from a uniform distribution. We then normalized the vector $e^j = (e_1^j,\ldots,e_k^j)$ such that $\|e_j\| = 1$ in order to get the relative abundance of each transcript in the gene. In order to generate the relative transcript abundances for the two samples, called A, and B, we changed for half of the genes the relative transcript abundance. This was done by, first choosing a vector $v_j^{A,B} \in [-0.5,0.5]^k$ for each sample, which determined the change of the relative transcript abundances and by choosing the strength of the change $c_j \in [0,1]$. Both the strength and the change vector were drawn from uniform distributions. For the sample A we adjusted e_j by adding $c_j v_j^A$ from it and for the gene in sample B we adjusted e_j by adding $c_j v_j^B$ to it. If any e_i^j was negative we set it to zero and if all e_i^j were negative we repeated the procedure above. After that we again normalized e_j, to get the final relative transcript abundances for the two samples. From those relative abundances we calculated the actual transcript abundances by multiplying the relative transcript abundances with the measured expressions from our experiments. We did this by choosing without replacement for each gene the expression estimates for a gene in the top 5875 expressed genes. We then used the estimate from the first sample for sample A and the ones from the second sample for sample B. After that we simulated the biological variance. We assumed that the transcript abundance is gamma distributed. Therefore, we drew from a Gamma distribution for each biological replicate new transcript abundances. The Gamma distribution was for a transcript with expressions e^j was $\Gamma\left(\frac{e_j^2}{f(e_j)-e_j}, \frac{f(e_j)-e_j}{e_j}\right)$. This Gamma distribution had the property that it was in concordance with our variance model when adding a poisson noise to it and that the mean transcript abundance was unchanged. We then used FluxSimulator (build 20100611) obtained from http://code.google.com/p/fluxcapacitor/downloads/list in order to simulated reads for our expression model.

S4 APPLICATIONS OF METHODS

S4.1 rDiff

S4.1.1 rDiff.parametric Since we assume that the gene model is complete we discarded all reads which are not in accordance with the gene model. For rDiff.parametric we estimated the variance function on counts in alternative regions. The fitting of the variance was performed as explained above

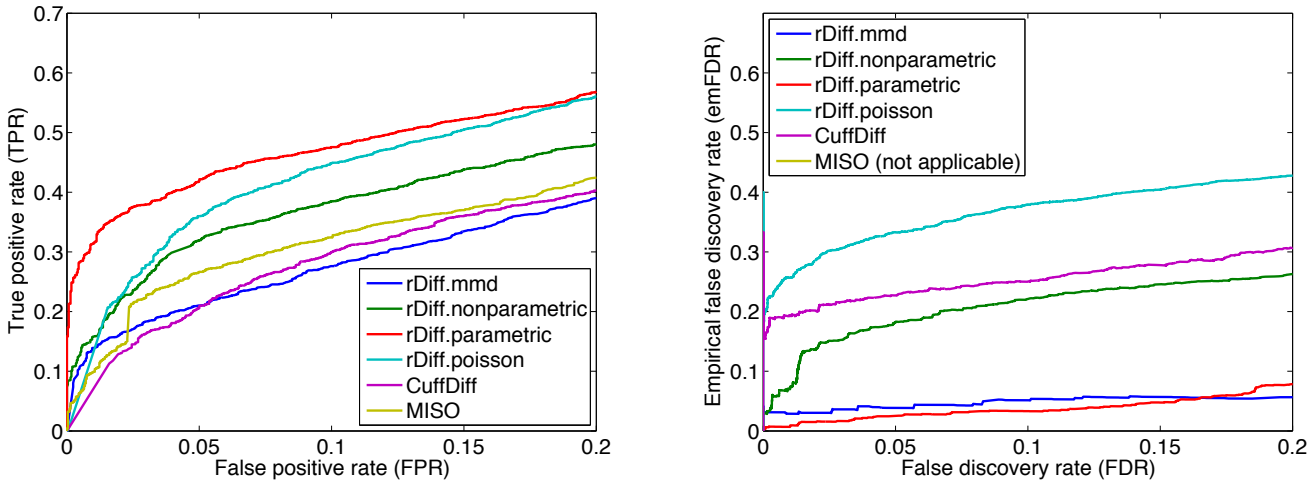
S4.1.2 rDiff.nonparametric For rDiff.nonparametric we estimated the variance function on every counts at each position of the genes. We used 1000 permutations for each gene. In order to speed up the computations on the real data we randomly sampled 10000 reads whenever there were more than 10000 reads mapping to a gene. For rDiff.nonparametric we estimated the variance function on the Coverage per position. The fitting of the variance function was done as described above.

MMD-tie breaking In order to resolve ties for genes that had the same p-value we added to each p-value a small value $\frac{\max_{j=1,\ldots,10} p_j}{\text{number of permutations}+1}$. This value is always smaller then the absolute difference between to of the raw p-values.

S4.2 Application of other methods

S4.2.1 CuffDiff We used for all our experiments CuffDiff from cufflinks-1.3.0 to detect differential splicing, which we obtained from http://cufflinks.cbcb.umd.edu. Contrasting previous findings (14) which have found that version 0.9.3 performed better then version 1.3.0 in identifying differential transcript expression, we have found that CuffDiff 1.3.0 performed better in detecting differential splicing than the version 0.9.3. For our experiments we used default parameter except for the following parameters, where we increased the iterations by a factor of 10:

```
--num-bootstrap-samples 200
```



(a) ROC-Curve for different methods

(b) Calibration of the methods

Figure S2. Comparison when the biological variance is large. (A) ROC-curve of rDiff, MISO and CuffDiff. (B) Comparison of the empirical false discovery rate (empFDR) and the FDR computed out of the p-values provided by the methods, if given.

```
--num-importance-samples 10000
--max-mle-iterations 50000
```

The resulting p-values for each test locus in a gene have then been combined using Bonferroni's correction for multiple testing.

S4.2.2 MISO We used for all our experiments MISO which we downloaded from the MISO website on 8/6/2011. For our experiments we used the default parameters. For the computation of the ROC-curve we used the Bayes factor as the ranking criterion. In the replicate setting we merged the replicated and proceeded as described above.

S5 EXPERIMENTAL DATA

S5.1 RNA isolation

RNA was isolated using the EURx RNA isolation kit and the following protocol. After grinding tissue in a Retsch Mill the sample is thawed while vortexing in 400 μl RL buffer supplemented with 4 μl β-mercaptoethanol. After spinning for 3 min at maximum speed the supernatant is transferred to a homogenization column. Spin for 2 min at maximum speed. 350 μl of 70 % ethanol is added to the flow through and mixed by pipetting. The mixture is transferred to an RNA binding column that is then spun at 11000 x g for 1 min. All following centrifugation steps are done at this speed and duration unless noted otherwise. The column is washed using 400 μl DN1. 50 μl DNR buffer mixed with 1 μl DNase I (Fermentas) are added to the column and incubated for 10 min at room temperature. After adding 400 μL the column is spun. The column is washed with first 650 μL, then 350 μL RBW and subsequently spun with the cap left open to dry. The RNA is eluted in 40 μL RNase-free water.

S5.2 RNA-seq library construction and sequencing

mRNA libraries were prepared using the Illumina mRNA-Seq 8-sample Prep kit according to the manufacturer's instructions, with exception to the size selected, which was around 300 bp, and an additional gel purification on a 3% agarose gel after the final PCR. Sequencing was run on the Genome Analzyer IIx using version 4 kits.

MAPPING

S5.3 Read alignment

The reads were aligned using Palmapper with the following parameter settings: Max number of mismatches: 6; Max number of gaps: 1; Max edit operations: 6; Minimal considered hit length: 15; Mini- mum length of long hit: 25; Minimum length of short hit: 8; Minimum combined length: 35; Longest intron length: 25,000; Maximum number of introns in spliced alignments: 2; Maximum number of spliced alignments per read: 5; CT: 10; Report a number of top scoring alignments: 10; Report spliced alignments; Number of hits of a seed that lead to it being ignored: 10,000; Report splice sites with confidence not less than a threshold: 0.9; Trigger spliced alignment, if unspliced alignment has at least this many mismatches: 2; Trigger spliced alignment,

if unspliced alignment has at least this many gaps: 0; filter- splice-min-edit: 1; filter-splice-region: 5; qpalma-use-map-max-len 1,000; polytrim: 40. Splice site predictions based on the annotated genome were used for the alignment. For spliced alignments, six bases on both sides of a splice site had to match perfectly. For all analyses we only considered reads which were longer than 70 bp. In order to further decrease the influence of suboptimally mapped read ends we clipped the three bases at the ends of all reads. For each gene we also removed all reads which could have stemmed from other genes.

S5.4 Validation of splicing events

RNA was isolated as described above and reverse transcribed using RevertAid^{TM}Premium Reverse Transcriptase (Fermentas) according to the manufacturers instructions with the exception of RNase inhibitor, which was omitted, and using the maximum amount of RNA possible in a reaction. qPCR reactions were prepared in 8 μL using MESA BLUE MasterMix plus for SYBR©(Eurogentec), 1 μL of cDNA diluted $1:100$ and 0.2 μM of each primer. The qPCR was run on a BioRad CFX384 under following cycling samples: initial denaturation at $95°$C for 5 min, 40 cycles of $95°$C for 15 sec and $60°$C for 45 sec. Fluorescence levels were measured at the end of every cycle, and at the end of the qPCR a melting curve for the products was recorded. Every measurement was done in technical triplicates, and the average of the triplicates was formed and used for downstream calculations. C_T values deviating from their technical replicates by more than 0.5 were treated as measurement errors and thus omitted from calculations. To determine the efficiency of every reaction 5 cDNA dilutions of a reference sample were measured using every primer pair. The amount of cDNA in the reactions was set arbitrarily. The efficiency was then determined by $10^{-\frac{1}{m}}-1$ where m is the slope of a linear regression fitted to the data points of the different dilutions. The formula $10^{\frac{C_T-x}{m}}$, where C_T is the average value of technical triplicates and x denotes the x-axis intercept of the aforementioned linear regression, gave the amount of cDNA transcript present in each sample given the arbitrarily set amount of cDNA transcript. Then the values were normalized to the amount of transcript calculated for total gene expression. The total gene expression was measured in a transcript region that is identical for all isoforms. The value at the time point 0 h was set to 1, and fold changes in transcript levels were calculated relative to this value.For each gene in every pair of comparison we have chosen the maximum fold change as being representative for the change in the respective gene.

S5.5 Genes found by MMD

Classification of rDiff.nonparametric hits

In order to determine which region was the most causative for genes detect by rDiff.nonparametric we first computed the squared difference of the mean coverages. We then reported the position which had the highest value when averaging the over the neighboring covered 50b. Following this we determined, using the annotation, from which parts of the gene the reported position came from. With this approach several regions can be detected for a gene, namely when the annotated regions overlap.

S5.6 Oversensitivity for highly expressed genes

We have found that rDiff.poisson showed a oversensitivity for high expressed genes. This is reflected by Fig S3. It shoes that the high expressed genes are enriched in the upper part of the ranking. An example for a highly expressed gene which is detected by the rDiff.poisson ($p \leq 2.67 * 10^{-07}$) but not by rDiff.parametric ($p \leq 0.897$) between WT 0 h and WT 1 h. As it can be seen in the Fig S4 for the rDiff.poisson, the high expressed genes are enriched in the genes with a low p-value when compared to rDiff.parametric.

Table S1. Table of the regions which contained the most differential 100 bp in the three comparisons between the three time points, in genes found by rDiff.nonparametric with a FDR smaller than 10%.

Event	WT 0 h vs WT 1 h	WT 0 h vs WT 6 h	WT 1 h vs WT 6 h
Intronic regions	118	126	77
5' UTR	30	36	23
3' UTR	46	47	23
First exon	29	22	13
Last exon	30	32	13
Other exons	18	10	14

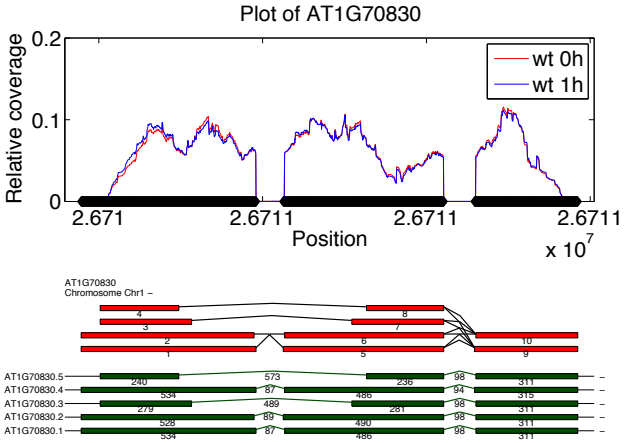


Figure S3. Example of a high expressed gene which is detected by rDiff.poisson but not by rDiff.parametric. Shown is the relative coverage and gene structure for AT1G70830

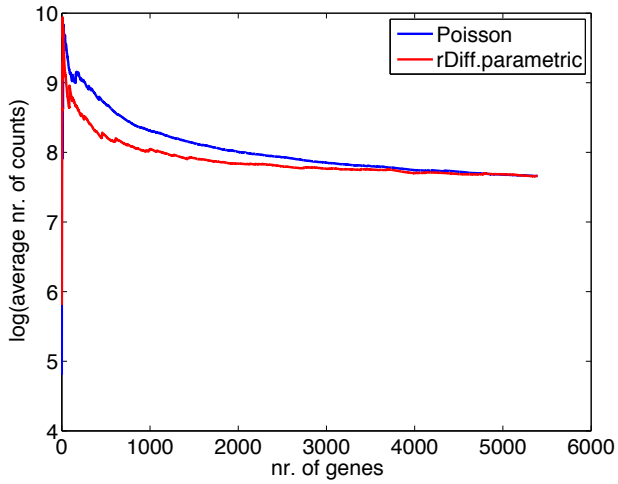


Figure S4. Log-mean expression of the top genes with the lowest p-value for the rDiff.poisson (blue) and rDiff.parametric (red). On the x-axis the number of genes is shown which were used to compute the log-mean.

S6 DIFFERENTIAL RNA PROCESSING IN D. MELANOGASTER

We downloaded the paired end read sequences from the NCBI Gene Expression Omnibus (libraries GSM461177, GSM461178, GSM461180, GSM461181). We trimmed the reads to 36b from the end. We then aligned the reads using Tophat 1.3.1 and the following strict parameters:

```
--segment-length 18
--max-insertion-length 0
--max-deletion-length 0
-g 10
```

using the genome Flybase, r5.22. We applied rDiff.nonparametric and rDiff.parametric as described in section S4. For the analysis we treated the both ends of the read-pairs as independent single-end reads.