Differential expression analysis for sequence count data

Simon Anders Wolfgang Huber

European Molecular Biology Laboratory (EMBL), Heidelberg, Germany

6 April 2010

Abstract

Motivation: High-throughput nucleotide sequencing provides quantitative readouts in assays for RNA expression (RNA-Seq), protein-DNA binding (ChIP-Seq) or cell counting (barcode sequencing). Statistical inference of differential signal in such data requires estimation of their variability throughout the dynamic range. When the number of replicates is small, error modelling is needed to achieve statistical power.

Results: We propose an error model that uses the negative binomial distribution, with variance and mean linked by local regression, to model the null distribution of the count data. The method controls type-I error and provides good detection power.

Availability: A free open-source R software package, *DESeq*, is available from the Bioconductor project and from
http://www-huber.embl.de/users/anders/DESeq.

Contact: sanders@fs.tum.de

1 Introduction

High-throughput sequencing of DNA fragments allows monitoring RNA abundance and protein-DNA binding, including the possibility to discover novel sequence variants and to dissect allele-specific effects and genetic variation. There is a range of technologies; a common feature between them is that they produce large amounts of sequence reads sampled from a preparation of DNA fragments that reflects, e.g., a biological system's repertoire of RNA molecules (RNA-Seq, Nagalakshmi et al. (2008); Mortazavi et al. (2008)) or the DNA or RNA interaction regions of nucleotide binding molecules (ChIP-Seq, Robertson et al. (2007); HITS-CLIP, Licatalosi et al. (2008)). Typically, these reads are classified based on their mapping to a common region of the target genome, where each class represents a target transcript, in the case of RNA-Seq, or a binding region, in the case of ChIP-Seq. An important summary statistic is the number of reads in a class; for RNA-Seq, this *read count* has been found to be (to good approximation) linearly related to the abundance of the target transcript (Mortazavi et al., 2008). Interest lies in comparing read counts between different biological conditions or between different genetic variants. In the simplest case, the comparison is done separately, class by class. We will use the term *gene* synonymously to *class*, even though a class may also refer to, e.g., a transcription factor binding site, or even a barcode (Smith et al., 2009).

We would like to use statistical testing to decide whether, for a given gene, an observed difference in read counts is significant, i.e., whether it is greater than what would be expected just due to natural random variation.

If reads were independently sampled from a population with given, fixed fractions of genes, the read counts would follow a multinomial distribution, which can be approximated by Poisson distributions. Consequently, Poisson distributions have been used to test for differential expression (Marioni et al., 2008; Wang et al., 2010). The single parameter of a Poisson distribution is determined by its mean, and its variance and all other properties follow from it; in particular, the variance is equal to the mean. However, it has been noted (Robinson and Smyth, 2007; Nagalakshmi et al., 2008) that the assumption of Poisson distribution is too tight: because it ignores the extra variation due to actual differences in replicate samples it predicts smaller variations than what is seen in the data. The resulting statistical test does therefore not control type-I error (the probability of false discoveries) as advertised. We show instances for this in Section 5.1.

To address this so-called overdispersion problem, it has been proposed to model count data with negative binomial (NB) distributions (Whitaker, 1914), and this approach is used in the *edgeR* package for analysis of SAGE and RNA-Seq (Robinson and Smyth, 2007; Robinson et al., 2010). The NB distributions are a family with two parameters, which are uniquely determined by mean μ and variance v. However, the number of replicates in data sets of interest is often too small to estimate both of those two parameters, mean and variance, reliably for each gene. For *edgeR*, Robinson and Smyth (2008) proposed to assume that mean and variance are related by $v = \mu + \alpha\mu^2$, with a single proportionality constant α that is the same throughout the experiment and that can be estimated from the data. Hence, only one parameter needs to be estimated for each gene, allowing application to experiments with small numbers of replicates.

In this paper, we extend this model by allowing more general, data-driven relationships of variance and mean, provide an effective algorithm for fitting the model to data, and show that it provides better fits (Section 2).

As a result, more balanced selection of differentially expressed genes throughout the dynamic range of the data can be obtained (Section 3). We demonstrate the method by applying it to four data sets (Section 4) and discuss how it compares to alternative approaches (Section 5).

2 Model

2.1 Description

We assume that the number of reads in sample j that are assigned to gene i can be modelled by a negative binomial (NB) distribution,

$$K_{ij} \sim \mathrm{NB}(\mu_{ij},\, \sigma^2_{ij}), \tag{1}$$

which has two parameters, the mean μ_{ij} and the variance σ^2_{ij}. The read counts K_{ij} are non-negative integers. The probabilities of the distribution are given in Supplementary Note A. This family of distributions is commonly used to model count data when overdispersion is present (Cameron and Trivedi, 1998).

In practice, we do not know the parameters μ_{ij} and σ^2_{ij}, and we need to estimate them from the data. Typically, the number of replicates is small, and further modelling assumptions need to be made in order to obtain useful estimates. In this paper, we develop a method that is based on the following three assumptions.

First, the mean parameter μ_{ij}, that is, the expectation value of the observed counts for gene i in sample j, is the product of a condition-dependent per-gene value $q_{i,\rho(j)}$ (where $\rho(j)$ is the experimental condition of sample j) and a library size parameter s_j,

$$\mu_{ij} = q_{i,\rho(j)}\, s_j. \tag{2}$$

$q_{i,\rho(j)}$ is proportional to the expectation value of the true (but unknown) concentration of fragments from gene i under condition $\rho(j)$. The library size parameter s_j is proportional to the coverage, or sampling depth, of library j, and we will use the term *common scale* for quantities, such as $q_{i,\rho(j)}$, that are adjusted for coverage by dividing by s_j.

Second, the variance σ^2_{ij} is the sum of a *shot noise term* and a *raw variance term*,

$$\sigma^2_{ij} = \underbrace{\mu_{ij}}_{\text{shot noise}} + \underbrace{s_j^2 v_{i,\rho(j)}}_{\text{raw variance}}. \tag{3}$$

Third, the per-gene raw variance parameter $v_{i,\rho(j)}$ is a smooth function v_ρ of the per-gene abundance $q_{i,\rho(j)}$,

$$v_{i,\rho(j)} = v_\rho(q_{i,\rho(j)}). \tag{4}$$

The decomposition of the variance in Equation (3) is motivated by the following hierarchical model: We assume that the actual concentration of fragments from gene i in sample j is proportional to a random variable R_{ij}, such that the rate that fragments from gene i are sequenced is $s_j r_{ij}$. For each gene i and all samples j of condition ρ, the R_{ij} are i.i.d. with mean $q_{i\rho}$ and variance $v_{i\rho}$. Thus, the count value K_{ij}, conditioned on $R_{ij} = r_{ij}$, is Poisson distributed with rate $s_j r_{ij}$. The marginal distribution of K_{ij} –when allowing for variation in R_{ij}– has the mean μ_{ij} and (according to the law of total variance) the variance given in Equation (3). Furthermore, if the higher moments of the distribution of R_{ij} are modelled according to a gamma distribution, the marginal distribution of K_{ij} is NB (see e.g. Cameron and Trivedi (1998, Sec. 4.2.2)).

2.2 Fitting

We now describe how the model can be fitted to data. The data are an $n \times m$ table of counts, k_{ij}, where $i = 1, \ldots, n$ indexes the genes, and $j = 1, \ldots, m$ indexes the samples. The model has three sets of parameters:

1. m library size parameters s_j; the expectation values of all counts from sample j are proportional to s_j.
2. for each experimental condition ρ, n gene abundance parameters $q_{i\rho}$; they reflect the expected abundance of fragments from gene i under condition ρ, i.e., expectation values of counts for gene i are proportional to $q_{i\rho}$.
3. The smooth functions $v_\rho : \mathbb{R}^+ \to \mathbb{R}^+$; they model the dependence of the raw variance $v_{i\rho}$ on the expected mean $q_{i\rho}$.

To estimate the size parameters, we use

$$\hat{s}_j = \operatorname*{median}_i \frac{k_{ij}}{\left(\prod_{\nu=1}^m k_{i\nu}\right)^{1/m}}. \tag{5}$$

The denominator of this expression can be interpreted as a pseudo-reference sample obtained by taking the geometric mean across samples. Each library size parameter estimate $\hat{s}_j$ is then computed as the median of the ratios of the j-th sample's counts to those of the pseudo-reference. In many cases, the values $\hat{s}_j$ are proportional to, and thus equivalent to, the sums $\sum_i k_{ij}$. However, it is not uncommon for the sums to be dominated by the counts for a few, highly abundant genes. In such cases, the estimator (5) is more robust, and produces a better library size adjustment for the majority of genes.

To estimate $q_{i\rho}$, we use the average of the counts from the samples j corresponding to condition ρ, transformed to the common scale:

$$\hat{q}_{i\rho} = \frac{1}{m_\rho} \sum_{j:\, \rho_j = \rho} \frac{k_{ij}}{\hat{s}_j}, \tag{6}$$

where m_ρ is the number of replicates of condition ρ and the sum runs over these replicates.

To estimate the functions v_ρ, we first calculate sample variances on the common scale

$$w_{i\rho} = \frac{1}{m_\rho - 1} \sum_{j:\, \rho(j) = \rho} \left(\frac{k_{ij}}{\hat{s}_j} - \hat{q}_{i\rho} \right)^2 \tag{7}$$

and define

$$z_{i\rho} = \frac{\hat{q}_{i\rho}}{m_\rho} \sum_{j:\,\rho(j)=\rho} \frac{1}{\hat{s}_j}. \tag{8}$$

In Supplementary Note B, we show that $w_{i\rho} - z_{i\rho}$ is an unbiased estimator for the raw variance parameter $v_{i\rho}$ of Equation (3).

However, for small numbers of replicates, m_ρ, as is typically the case in applications, the values $w_{i\rho}$ are highly variable, and $w_{i\rho} - z_{i\rho}$ would not be a useful variance estimator for statistical inference. Instead, we use local regression (Loader, 1999) on the graph $(\hat{q}_{i\rho}, w_{i\rho})$ to obtain a smooth function $w_\rho(q)$, with

$$\hat{v}_\rho(\hat{q}_{i\rho}) = w_\rho(\hat{q}_{i\rho}) - z_{i\rho} \tag{9}$$

as our estimate for the raw variance.

Some attention is needed to avoid estimation biases in the local regression. $w_{i\rho}$ is a sum of squared random variables, and the residuals $w_{i\rho} - w(\hat{q}_{i\rho})$ are strongly skewed. Following McCullagh and Nelder (1989, Ch. 8) and Loader (1999, Sec. 9.1.2), we use a generalised linear model of the gamma family for the local regression, using the implementation in the *locfit* package (Loader, 2007).

3 Testing for differential expression

Suppose that we have m_A replicate samples for biological condition A and m_B samples for condition B. For each gene i, we would like to weigh the evidence in the data for or against differential abundance of that gene between the two conditions. In particular, we would like to test the null hypothesis $q_{i\mathrm{A}} = q_{i\mathrm{B}}$, where $q_{i\mathrm{A}}$ is the gene abundance parameter for the samples of condition A, and $q_{i\mathrm{B}}$ for condition B. To this end, we define, as test statistic, the total counts in each condition,

$$K_{i\mathrm{A}} = \sum_{j:\rho(j)=\mathrm{A}} K_{ij} \qquad K_{i\mathrm{B}} = \sum_{j:\rho(j)=\mathrm{B}} K_{ij}, \tag{10}$$

and their overall sum $K_{i\mathrm{S}} = K_{i\mathrm{A}} + K_{i\mathrm{B}}$. From the error model of Section 2, we show below that we can compute the probabilities of the events $K_{i\mathrm{A}} = a$ and $K_{i\mathrm{B}} = b$ for any pair of numbers a and b. We denote this probability by $p(a, b)$. The p value of a pair of observed count sums $(k_{i\mathrm{A}},\, k_{i\mathrm{B}})$ is then the sum of all probabilities less or equal to $p(k_{i\mathrm{A}},\, k_{i\mathrm{B}})$, given that the overall sum is k_{iS}:

$$p_i = \frac{\sum\limits_{a+b=k_{i\mathrm{S}},\; p(a,b)\le p(k_{i\mathrm{A}},k_{i\mathrm{B}})} p(a,b)}{\sum\limits_{a+b=k_{i\mathrm{S}}} p(a,b)} \tag{11}$$

The variables a and b in the above sums take the values $0, \ldots, k_{i\mathrm{S}}$. The approach presented so far follows that of Robinson and Smyth (2008) and is analogous to that taken by other conditioned tests, such as Fisher's exact test. (See Agresti (2002, Ch. 2) for a discussion of the merits of conditioning in tests.)

Computation of $p(a, b)$. First, assume that, under the null hypothesis, counts from different samples are independent. Then, $p(a, b) = \Pr(K_{i\mathrm{A}} = a)\, \Pr(K_{i\mathrm{B}} = b)$. The problem thus is computing the probability of the event $K_{i\mathrm{A}} = a$, and, analogously, of $K_{i\mathrm{B}} = b$. The random variable $K_{i\mathrm{A}}$ is the sum of m_A NB-distributed random variables. We approximate its distribution by a NB distribution whose parameters we obtain from those of the K_{ij}. To this end, we first compute the pooled mean estimate from the counts of both conditions,

$$\hat{q}_{i0} = \sum_{j:\rho(j)\in\{A,B\}} k_{ij}/s_j, \tag{12}$$

which accounts for the fact that the null hypothesis stipulates that $q_{i\mathrm{A}} = q_{i\mathrm{B}}$. The summed mean and variance for condition A are

$$\hat{\mu}_{i\mathrm{A}} = \sum_{j\in\mathrm{A}} s_j \hat{q}_{i0}, \tag{13}$$

$$\hat{\sigma}^2_{i\mathrm{A}} = \sum_{j\in\mathrm{A}} \hat{s}_j \hat{q}_{i0} + \hat{s}_j^2 \hat{v}_\mathrm{A}(\hat{q}_{i0}), \tag{14}$$

Supplementary Note C describes how the distribution parameters of the NB for $K_{i\mathrm{A}}$ can be determined from $\hat{\mu}_{i\mathrm{A}}$ and $\hat{\sigma}^2_{i\mathrm{A}}$. (To avoid bias, we do not match the moments directly, but instead match a different pair of distribution statistics.) The parameters of $K_{i\mathrm{B}}$ are obtained analogously. Supplementary Note D explains how we evaluate the sums in Equation (11).

4 Applications

4.1 Data sets

There are only few published RNA-Seq data sets with biological replication. Here, we present results based on the following data sets:

RNA-Seq in fly embryos. B. Wilczynski, Y.-H. Liu, N. Delhomme and E. Furlong have conducted RNA-Seq experiments in fly embryos and kindly shared part of their data with us ahead of publication. In each sample of this data set, a gene was engineered to be overexpressed, and we compare two biological replicates each of two such conditions, in the following denoted as "A" and "B".

Tag-Seq of neural stem cells. Engström et al. (2010) performed Tag-Seq (Morrissy et al., 2009) for tissue cultures of neural cells, including four from glioblastoma-derived neural stem-cells ("GNS") and two from non-cancerous neural stem ("NS") cells. As each tissue culture is derived from a different subject and so has a different genotype, these data show high variability.

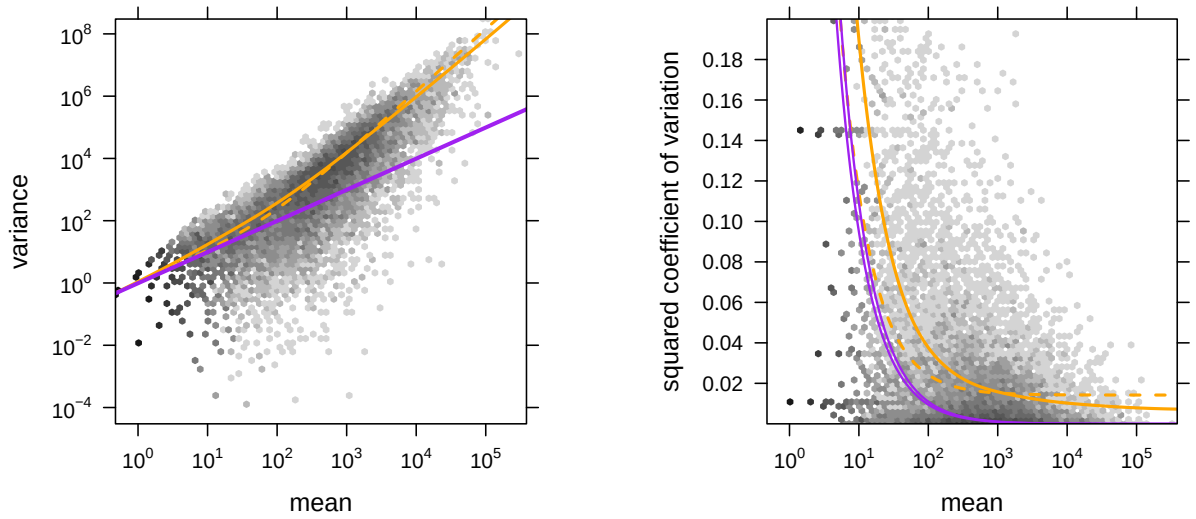

Figure 1: (*a*) Dependence of the variance on the mean for condition A in the fly RNA-Seq data. The scatter plot shows the common-scale sample variances (Equation (7)) plotted against the common-scale means (Equation (6)). The orange line is the fit $w(q)$. The purple lines show the variance implied by the Poisson distribution for each of the two samples, i.e., $\hat{s}_j\hat{q}_{i,A}$. The dashed orange line is the variance estimate used by *edgeR*. (*b*) Same data as in (*a*), with the y-axis rescaled to show the squared coefficient of variation (SCV), i.e. all quantities are divided by the square of the mean. The solid orange line is computed using the bias correction described in Supplementary Note C. (The plot only shows SCV values in the range $[0, 0.2]$. For a zoom-out to the full range, see Supplementary Figure S8).

RNA-Seq of yeast. Nagalakshmi et al. (2008) performed RNA-Seq on replicates of *Saccharomyces cerevisiae* cultures. They tested two library preparation protocols, *dT* and *RH*, and obtained three sequencing runs for each protocol, such that for the first run of each protocol, they had one further technical replicate (same culture, replicated library preparation) and one further biological replicate (different culture).

ChIP-Seq of HapMap samples. Kasowski et al. (2010) compared DNA-protein binding between 10 human individuals by ChIP-Seq. They compiled a list of binding regions for polymerase II and NF-κB, and counted, for each sample, the number of reads that mapped onto each binding region. The aim of the study was to investigate how much the binding differs between individuals.

4.2 Variance estimation

We start by demonstrating the variance estimation. Figure 1a shows the sample variances $w_{i\rho}$ (Equation (7)) plotted against the means $\hat{q}_{i\rho}$ (Equation (6)) for condition A in the fly RNA-Seq data. Also shown is the local regression fit $w_\rho(q)$ and the shot noise $\hat{s}_j\hat{q}_{i\rho}$. In Figure 1b, we plotted the squared coefficient of variation (SCV), i.e. the ratio of the variance to the mean squared. In this plot, the distance between the orange and the purple line is the SCV of the noise due to biological sampling (cf. Equation (3)).

The many data points in Figure 1b that lie far above the fitted orange curve may let the fit of the local regression appear poor. However, a strong skew of the residual distribution is to be expected. See Supplementary Note E for details and a discussion of diagnostics suitable to verify the fit.

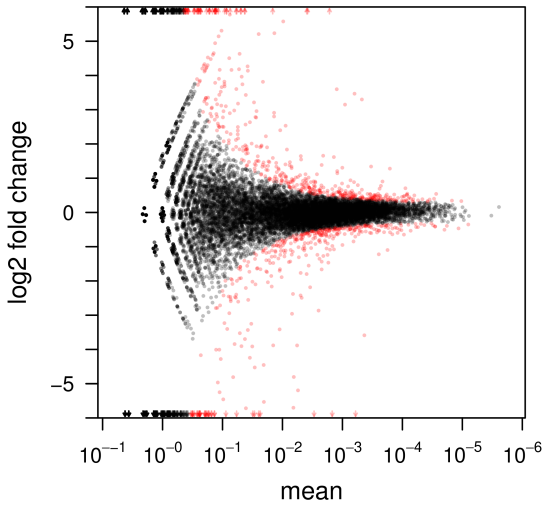

Figure 2: Testing for differential expression between conditions A and B. (a) Scatter plot of $\log_2$ ratio (fold change) versus mean. The red colour marks genes detected as differentially expressed at 10% false discovery rate when Benjamini-Hochberg multiple testing adjustment is used. The symbols at the upper and lower plot border indicate genes with very large or infinite fold changes. (b) For the corresponding volcano plot, see Supplementary Figure S9.

4.3 Testing

In order to verify that *DESeq* maintains control of type-I error, we contrasted one of the replicates for condition A against the other one, using for both samples the variance function estimated for condition A. In this case, we expect to find uniformly distributed p values. Figure 3 shows this to be the case.

Next, we contrasted the two A samples against the two

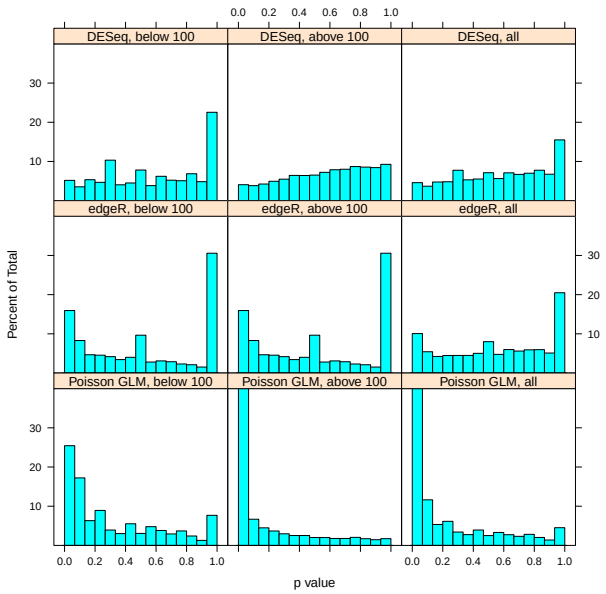

Figure 3: Type-I error control by *DESeq*, *edgeR* and a Poisson-based test. The histograms show p values from a comparison of one replicate for condition A with the other one. Between replicates, no genes are truly differentially expressed, and the distribution of p-values is expected to be uniform in the interval $[0, 1]$. Left and middle column show the distributions separately for genes below and above a mean of 100, right column for all genes. While both *DESeq* (top row) and *edgeR* (middle row) are slightly conservative for high counts, *DESeq* has an excess of too low p values for low counts. (The peak at 1 for low counts is, for both methods, due to discreteness: for low counts, the count sums from both conditions are frequently exactly equal, forcing $p = 1$.) The bottom row shows that a Poisson-based χ^2 test fails to maintain type-I error control.

B samples. Using the procedure described in Section 3, we computed a p value for each gene.

Figure 2 shows the obtained fold changes and p values. 14% of the p values are below 5%. Adjustment for multiple-testing with the false discovery rate (FDR) controlling procedure of Benjamini and Hochberg (1995) yielded significant differential expression at a 10% FDR level for a set 867 genes (of 17,605). These are marked in red in the figure.

Figure 2 demonstrates how the power to detect differential expression depends on overall counts. Observe how even a small increase in counts drastically improves a detection power for counts below ~ 100, while at higher counts, when shot noise becomes unimportant (cf. Figure 1b), power depends only weakly on count level. These plots are helpful to guide experiment design: For weakly expressed genes, in the region where shot noise is important, deeper sequencing would be helpful to increase power if the current detection power were considered unsatisfactory, while for the higher-count regimes, only further biological replicates would allow further improvement.

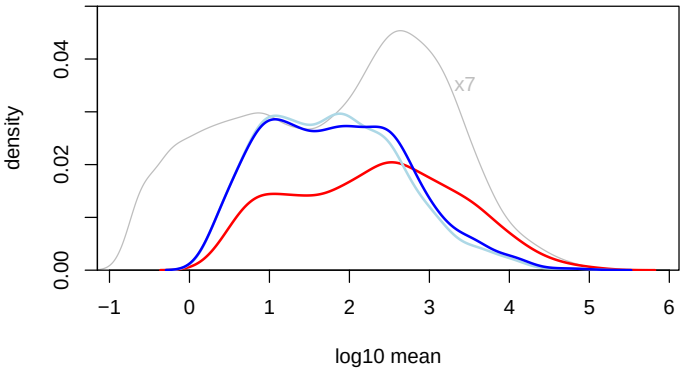

Figure 4: The density of common-scale mean values q_i for all genes (grey line, scaled down by a factor of 7), for the hits reported by *DESeq* (red line) and by *edgeR* (dark blue line: with tag-wise dispersion estimation; light blue line: common dispersion mode)

4.4 Comparison with edgeR

We also performed the test for differential expression with *edgeR* (version 1.5.11; Robinson and Smyth (2007, 2008); Robinson et al. (2010)). While *DESeq* reports 867 genes at Benjamini-Hochberg adjusted FDR of 10%, *edgeR* finds 1,116 genes significant. The two hit lists share 705 genes.

From these numbers alone, it would be hard to see whether and in what way the differences between *DESeq* and *edgeR* are consequential. However, the difference between the results does not merely lie in the number of genes, but also in their properties. As can be seen from Figure 4, the distributions of expression strengths of the significant genes are quite different. A priori, one would expect the distribution to roughly follow the expression strength distribution for all genes (expect for low counts, where detection power is poorer). For *DESeq*'s hits, this is indeed the case, while *edgeR*'s hits tend to concentrate at lower abundance. *edgeR*'s bias is likely not a reflection of biology, but rather a result of the rigidity of its error model: *edgeR* estimates a common dispersion of 0.014. The dashed orange line in Figure 1a and b shows the variance implied by a raw SCV of this value. As one can see, it is lower than *DESeq*'s estimate (solid orange line) for the lower part of the the dynamic range, and higher in the upper range. Figure 3 shows how this leads to an anti-conservative behaviour of *edgeR* in the low count range, leading *edgeR* to call too many hits among genes with low counts. This matches the observation from Figure 4. On average, over the whole dynamic range, FDR control is of course maintained, albeit at the cost of detection power.

In Supplementary Note F, we observe a similar effect with the neural stem cell data.

4.5 Working without replicates

DESeq allows analysis of experiments with no biological replicates in one or even in both of the conditions. While

one may not want to draw strong conclusions from such an analysis, it may still be useful for exploration and hypothesis generation.

If replicates are available only for one of the conditions, one might choose to assume that the variance-mean dependence estimated from the data for that condition holds as well for the unreplicated one.

If neither condition has replicates, one can still perform an analysis based on the assumption that for most genes, there is no true differential abundance, and that a valid mean-variance relationship can be estimated from treating the two samples as if they were replicates. A minority of differentially abundant genes will act as outliers; however, they will not have a severe impact on the gamma-family GLM fit, as the gamma distribution for low values of the shape parameter $(m-1)/2$ has a heavy right-hand tail. Some overestimation of the variance may be expected, which will make that approach conservative.

We performed such an analysis for the fly RNA-Seq and the neural cell Tag-Seq data, by restricting both data sets to only two samples, one from each condition. For the neural cell data, the estimated variance function was, as expected, above the two functions estimated from the *GNS* and *NS* replicates. Using it to test for differential abundance still finds a 271 hits at 10% FDR, and these hits have good overlap (202 hits) with those found from the more reliable analysis with all available samples, i.e., about a third of the hits are recovered. In the case of the fly RNA-Seq data, however, only 93 of the 867 hits (i.e., only 11%) are recovered (with 2 new hits). This is explained by the fact that the low noise in the fly data makes the use of proper replication much more beneficial than in case of the neural cell data, where the noise between replicates is hardly lower than between conditions (see also Section 4.6).

4.6 Variance-stabilising transformation

Given a variance-mean dependence, a variance-stabilising transformation (VST) is a monotonous bijection such that for the transformed values, the variance is (approximately) independent of the mean. Using the variance-mean dependence $w(q)$ estimated by *DESeq*, a VST is given by

$$\tau(\kappa) = \int^{\kappa} \frac{\mathrm{d}q}{\sqrt{w(q)}}. \qquad (15)$$

Applying the transformation τ to the common-scale count data, k_{ij}/s_j, yields new data values whose variances are approximately the same throughout the dynamic range.

One application of VST is sample clustering, as in Figure 5; such an approach is more straightforward than, say, defining a suitable distance metric on the untransformed count data, whose choice is not obvious, and may not be easy to combine with available clustering or classification algorithms (which tend to be designed for variables with similar distributional properties).

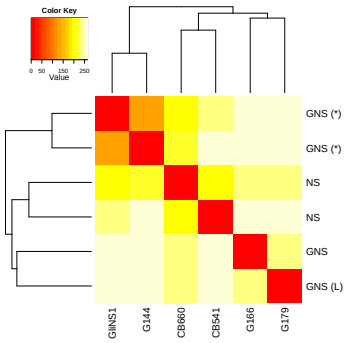

Figure 5: Sample clustering for the data of Engström et al. (2010). A common variance function was estimated for all samples and used to apply a variance-stabilising transformation. The heat map shows a false colour representation of the Euclidean distance matrix, and the dendrogram represents a hierarchical clustering. Two *GNS* samples were derived from the same patient (marked with "(*)") and show the highest degree of similarity. The two other *GNS* samples (including one with atypically large cells, marked "(L)") are as dissimilar from the former as the two *NS* samples.

4.7 ChIP-Seq

DESeq can also be used to analyse comparative ChIP-Seq assays. Kasowski et al. (2010) have analysed transcription factor binding for HapMap individuals, and counted how many reads mapped for each sample to pre-determined binding regions. We considered two individuals from their data set, HapMap IDs GM12878 and GM12891, for both of which at least four replicates had been done, and tested for differential occupation of binding regions. Type-I error control was maintained by *DESeq*: the lower left two panels of Figure 6 show approximately uniform p value histograms for comparisons within the same individual, and no binding region was significant at 10% FDR using Benjamini-Hochberg adjustment. Differential binding was found, however, when contrasting the two individuals, with 6,450 binding regions significant when only two replicates each were used and 9,415 when four replicates were used.

Using an alternative approach, Kasowski et al. (2010) fitted generalised linear models (GLMs) of the Poisson family, i.e., the noise was not estimated from the data but assumed to be shot noise only. This (upper row of Figure 6) results in an enrichment of small p values even for comparisons within the same individual, indicating that the variance is underestimated by the Poisson GLM, and literal use of the p values would lead to anti-conservative (overly optimistic) bias. Kasowski et al. (2010) addressed this by adjusting for this bias and using additional criteria for calling differential binding regions.

5 Discussion

Why is it necessary to develop new statistical methodology for sequence count data? If large numbers of replicates were available, questions of data distribution could be avoided by using non-parametric methods, such as

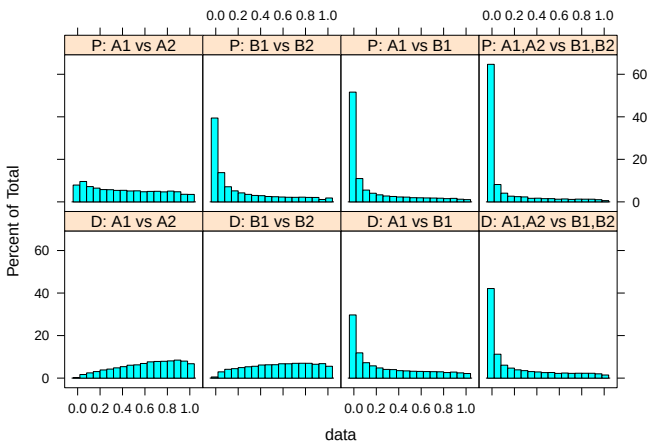

Figure 6: Application to ChIP-Seq data. Shown are p value histograms resulting from comparisons of Pol-II ChIP-Seq data between replicates of the same individual (first and second column) and between two different individuals (third and forth column). The upper row corresponds to an analysis based on Poisson GLMs ("P"), the bottom row to analysis with *DESeq* ("D"). In the first column, two replicates from HapMap individual GM12878 (*A1*) are compared against two further replicates from the same individual (*A2*). As expected, the p value histograms are approximately flat, indicating no significant differences. In the second column, two replicates from individual GM12891 (*B1*) are compared against two further replicates from the same individual (*B2*). While no significant differences are expected, the Poisson GLM analysis finds an enrichment of small p values; this is a reflection of overdispersion in the data, that is, the variance in the data is larger than what the Poisson GLM assumes (see also Section 5.1). The third column compares two replicates from individual GM12878 (*A1*) against two from the other individual (*B1*). True binding differences are expected, and both methods result in enrichment of small p values. The forth column shows the comparison of four replicates of GM12878 (*A1* combined with *A2*) against four replicates of GM12891 (*B1*, *B2*); increased sample size leads to higher detection power.

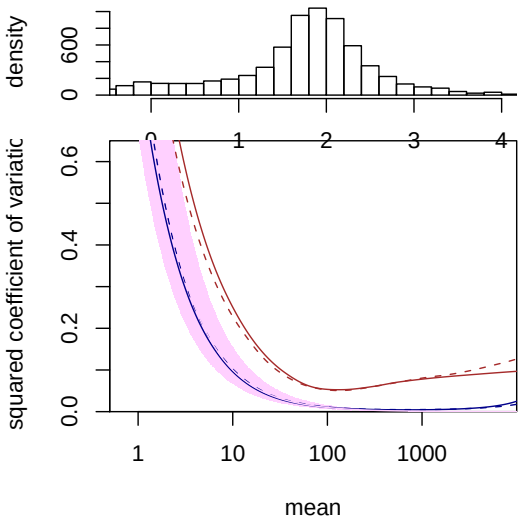

Figure 7: Noise estimates for the data of Nagalakshmi et al. (2008). The data allow assessment of technical variability (between library preparations from aliquots of the same yeast culture) and biological variability (between two independently grown cultures). The blue curves depict the squared coefficient of variation at the common scale, $w_\rho(q)/q^2$ (see Equation (9)) for technical replicates, the red curves for biological replicates (solid lines, dT data set, dashed lines, RH data set). The data density is shown by the histogram in the top panel. The purple area marks the range of the shot noise for the range of library sizes in the data set. One can see that the noise between technical replicates follows closely the shot noise limit, while the noise between biological replicates exceeds shot noise already for low count values.

Wilcoxon or permutation tests. However, it is desirable (and possible) to consider experiments with smaller numbers of replicates per condition. In order to compare an observed difference with the to be expected random variation, we can improve our picture of the latter in two ways: first, we can use distribution families, such as normal, Poisson and negative binomial distributions, in order to determine the higher moments, and hence the tail behaviour, of statistics for differential abundance, based on observed low order moments such as mean and variance. Second, we can share information, for instance, distributional parameters, between genes, based on the notion that data from different genes follow similar patterns of variability. Here, we have described an instance of such an approach, and we will now discuss the choices we have made.

5.1 Distributional family

While for large counts, the normal distributions might provide a good approximation of between-replicate variability, this is not the case for lower count values, whose discreteness and skewness mean that probability estimates computed from a normal approximation would be inadequate.

For the Poisson approximation, a key paper is the work by Marioni et al. (2008), who studied the *technical* reproducibility of RNA-Seq. They extracted total RNA from two tissue samples, one from the liver and one from the kidneys of the same individuum. From each RNA sample they took seven aliquots, prepared a library from each aliquot according to the protocol recommended by Illumina and sampled each library on one lane of a Solexa genome analyser. For each gene, they then calculated the variance of the seven counts from the same tissue sample and found very good agreement with the variance predicted by a Poisson model. In line with our arguments in Section 2, Poisson shot noise is the minimum amount of variation to expect in a counting process. Thus, Marioni et al. (2008) concluded that the technical reproducibility of RNA-Seq is excellent, and that the variation between technical replicates is close to the shot noise limit.

From this vantage point, Marioni et al. (2008) (and similarly Bullard et al. (2010)) suggested to use the Poisson model (and Fisher's exact test, or a likelihood ratio test as an approximation to it) to test whether a gene is differentially expressed between their two samples. It is now vital to notice that a rejection from such a test only informs us that the difference between the average

counts in the two samples is larger than one would expect between *technical* replicates. Hence, we do not know whether this difference is due to the different tissue type, kidney instead of liver, or whether a difference of the same magnitude could have been found as well if one had compared two samples from different parts of the same liver, or from livers of two individuals.

Figure 1 shows that shot noise is only dominant for very low count values, while already for moderate counts, the effect of the biological variation between samples exceeds the shot noise by many orders of magnitude. This is confirmed by comparison of technical with biological replicates (Nagalakshmi et al., 2008). In Figure 7, we used *DESeq* to obtain variance estimates for the data of Nagalakshmi et al. (2008). The analysis indicates that the difference between technical replicates barely exceeds shot noise level, while biological replicates differ much more.

Tests for differential abundance that are based on a Poisson model, such as those discussed by Marioni et al. (2008), Wang et al. (2010), Bloom et al. (2009), Bullard et al. (2010) or Kasowski et al. (2010) should thus be interpreted with caution, as they may severely underestimate the effect of biological variability.

Consequently, it is preferable to use a model that allows for overdispersion. While for the Poisson distributions, variance and mean are equal, the negative binomial distributions are a generalisation that allow for the variance to be larger. The most advanced of the published methods using this family of distributions is likely *edgeR* (Robinson and Smyth, 2007). *DESeq* owes its basic idea to a good part to *edgeR*, but differs in several aspects.

5.2 Sharing of information between genes

First, we discovered that the use of total read counts as estimates of sequencing depth, and hence for the adjustment of observed counts between samples (as recommended by Robinson and Smyth (2007) and other authors) may result in high apparent differences between replicates, and hence in poor power to detect true differences. *DESeq* uses the more robust size estimate Equation (5); in fact, *edgeR*'s power increases when it is supplied with those size estimates instead.

For small numbers of replicates such as often encountered in practice, it is not possible to obtain simultaneously reliable estimates of the variance and mean parameters of the NB distribution. *edgeR* addresses this problem by estimating a single *common dispersion* parameter. In our method, we make use of the possibility to estimate a more flexible, mean-dependent local regression. The amount of data available in typical experiments is large enough to allow for sufficiently precise local estimation of the dispersion. Over the large dynamic range that is typical for RNA-Seq, the raw SCV often appears to change noticeably, and taking this into account allows *DESeq* to avoid bias towards certain areas of the dynamic range in its differential-expression calls (see Figures 3 and 4).

This flexibility is the most substantial difference between *DESeq* and *edgeR*, as simulations show that *edgeR* and *DESeq* perform comparably if provided with artificial data with constant SCV (Supplementary Note G). *EdgeR* attempts to make up for the rigidity of the single-parameter noise model by allowing for an adjustment of the model-based variance estimate with the per-gene empirical variance. An empirical Bayes procedure, which was originally developed for the *limma* package (Smyth, 2005, 2004; Lönnstedt and Speed, 2002), determines how to combine these two sources of information optimally. However, for typical low replicate numbers, this so-called tagwise dispersion mode seems to have little effect (Figure 4) or even reduces *edgeR*'s power (Supplementary Note F).

Third, we have suggested a simple and robust way of estimating the raw variance from the data. Robinson and Smyth (2008) employed a technique they called quantile-adjusted conditional maximum likelihood to find an unbiased estimate for the raw SCV. The *quantile adjustment* refers to a rank-based procedure that modifies the data such that the data seem to stem from samples of equal library size. In *DESeq*, differing library sizes are simply addressed by linear scaling (Equations (2) and (3)), suggesting that quantile adjustment is an unnecessary complication. The price we pay for this is that we need to make the approximation that the sum of NB variables in Equation (10) is itself NB distributed. While it seems that neither the quantile adjustment nor our approximation pose reason for concern in practice, *DESeq*'s approach computationally is faster and, perhaps, conceptually simpler.

Fourth, our approach provides useful diagnostics. Plots such as Supplementary Figure S2 are helpful to judge the reliability of the tests. In Figures 1b and 7, it is easy to see at which mean value biological variability dominates over shot noise; this information is valuable to decide whether the sequencing depth or the number of biological replicates is the limiting factor for detection power, and so helps in planning experiments. A heat map as in Figure 5 is useful as data quality control.

6 The R package *DESeq*

We implemented our method as a package for the statistical environment R (R Development Core Team, 2009) and distribute it within the Bioconductor project (Gentleman et al., 2004). As input, it expects a table of count data. The data, as well as meta-data, such as sample and gene annotation, are managed with the S4 class *CountDataSet*, which is derived from *eSet*, Bioconductor's standard data type for table-like data. The package provides high-level functions to perform analyses such as in Section 4 with only a few commands, allowing researchers with little knowledge of R to use it. This is demonstrated with examples in the documentation (the so-called pack-

age vignette). Furthermore, lower-level functions are supplied for more experienced users who wish to deviate from the standard work flow. A typical calculation, such as the analysis shown in Section 4.2, takes a few minutes of computation time on a desktop computer.

Acknowledgements

We are grateful to Paul Bertone for sharing the neural stem cells data ahead of publication, and to Bartek Wilczyński, Ya-Hsin Liu, Nicolas Delhomme and Eileen Furlong likewise for sharing the fly RNA-Seq data. We thank Nicolas Delhomme and Julien Gagneur for helpful comments on the manuscript. S. An. has been partially funded by the European Union Research and Training Network "Chromatin Plasticity".

References

Agresti, A. (2002). *Categorical Data Analysis*. Wiley, 2nd edition.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B*, 57, 289.

Bliss, C. I. and Fisher, R. A. (1953). Fitting the negative binomial distribution to biological data. *Biometrics*, 9, 176.

Bloom, J. S. et al. (2009). Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics*, 10, 221.

Bullard, J. et al. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11, 94.

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

Clark, S. J. and Perry, J. N. (1989). Estimation of the negative binomial parameter κ by maximum quasi-likelihood. *Biometrics*, 45, 309.

Engström, P. et al. (2010). Transcriptional characterization of glioblastoma stem cell lines using tag sequencing. In preparation. [Full author list: P. Engström, D. Tommei, S. Stricker, A. Smith, S. Pollard, P. Bertone].

Gentleman, R. C. et al. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5, R80.

Kasowski, M. et al. (2010). Variation in transcription factor binding among humans. *Science*, 328, 232.

Lawless, J. F. (1987). Negative binomial and mixed poisson regression. *The Canadian Journal of Statistics*, 15, 209.

Licatalosi, D. D. et al. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456, 464.

Loader, C. (1999). *Local Regression and Likelihood*. Springer.

Loader, C. (2000). Fast and accurate computation of binomial probabilities. http://projects.scipy.org/scipy/raw-attachment/ticket/620/loader2000Fast.pdf (Note: This is a copy of the original paper, which is no longer available online.).

Loader, C. (2007). *locfit: Local Regression, Likelihood and Density Estimation*. R package version 1.5-4.

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

Marioni, J. C. et al. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, 18, 1509.

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

Morrissy, A. S. et al. (2009). Next-generation tag sequencing for cancer gene expression profiling. *Genome Research*, 19, 1825.

Mortazavi, A. et al. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5, 621.

Nagalakshmi, U. et al. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320, 1344.

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.

Robertson, G. et al. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Meth.*, 4, 651 .

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinf.*, 26, 139.

Robinson, M. D. and Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23, 2881.

Robinson, M. D. and Smyth, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostat*, 9, 321.

Saha, K. and Paul, S. (2005). Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics*, 61, 179.

Smith, A. M. et al. (2009). Quantitative phenotyping via deep barcode sequencing. *Genome Research*, 19, 1836.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Gen. Mol. Biol.*, 3. Article 3.

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

Wang, L. et al. (2010). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, 26, 136.

Whitaker, L. (1914). On the Poisson law of small numbers. *Biometrika*, 10, 36.

Differential expression analysis for sequence count data

Simon Anders, Wolfgang Huber

Supplement

A Parametrisation of the negative binomial distribution

An integer valued random variable K is said to follow a negative binomial distribution with parameters $p \in]0,1[$ and $r \in]0,\infty[$ if (Cameron and Trivedi, 1998)

$$\Pr(K = k) = \binom{k+r-1}{r-1} p^r (1-p)^k. \qquad (16)$$

This two-parametric distribution can, equivalently, be parametrised in terms of its mean μ and variance σ^2, via

$$p = \frac{\mu}{\sigma^2} \quad \text{and} \quad r = \frac{\mu^2}{\sigma^2 - \mu}. \qquad (17)$$

B Variance estimator

In Section 2.2, we claim that $\hat{w}_{i\rho} - z_{i\rho}$, as defined in Eqs. (7, 8), is an unbiased estimator for the raw variance $v_{i\rho}$. To show this, we compute the expectation value of $\hat{w}_{i\rho}$. To simplify notation, we suppress the indices i and ρ in the following. Furthermore, we neglect differences between the true library sizes s_j and their estimates $\hat{s}_j$. Then,

$$\hat{Q} = \frac{1}{m} \sum_{j=1}^{m} \frac{K_j}{s_j}$$

is an unbiased estimator of q, because, due to Equation (2), $\mathbb{E}\, K_j = s_j q_0$. Next, we examine

$$(m-1)\, \hat{w} = \sum_{j:\, \rho_j = \rho} \left(\frac{k_j}{s_j} - \hat{q} \right)^2 .$$

Taking expectations on both sides yields

$$(m-1)\, \mathbb{E}\, \hat{w} = \left(1 - \frac{1}{m}\right) \sum_j \frac{\mathbb{E}\, K_j^2}{s_j^2} - \frac{1}{m} \sum_{\substack{j,l \\ j \neq l}} \frac{\mathbb{E}\, K_j K_l}{s_j s_l}$$

For $j \neq k$, K_j and K_l are independent, and hence $\mathbb{E}\, K_j K_l = s_j s_l \hat{q}^2$, while for $j = l$, we have $\mathbb{E}\, K_j^2 = (\mathbb{E}\, K_j)^2 + \operatorname{Var} K_j = s_j^2 \hat{q}^2 + s_j \hat{q} + s_j^2 v$ by the definition of variance and Equation (3). Using this, we find

$$\mathbb{E}\, \hat{w} = v + \underbrace{\frac{\hat{q}}{m} \sum_j \frac{1}{s_j}},$$

where the under-braced part is the bias correction term z.

C Removal of bias due to parametrisation

When estimating distribution parameters for the purpose of calculating p values from the distribution, bias in the parameter estimates can cause problems. As the choice of parameters to characterise a distribution is arbitrary, the question arises for which set of parameters bias should be minimised in order to allow for accurate inference.

For the NB distribution, we investigated this issue: By means of simulations with similar settings as in Supplementary Note G, we found that if we used the unbiased mean and variance estimates $\hat{q}_{i\rho}$ and $w_\rho(\hat{q}_{i\rho})$ from Equations (6) and (9) to calculate p values with Equation (11) for simulated data without any differential expression, the p values were not uniform, but tended to be too small when the number of replicates was low. In previous work on inference based on the NB distribution, the authors usually aimed at getting unbiased estimates for another pair of parameters, namely for the mean and either for the dispersion parameter (e.g., Bliss and Fisher (1953)) or, more recently, for its reciprocal, i.e., the quantity we denote the raw SCV (e.g., Clark and Perry (1989); Lawless (1987); Saha and Paul (2005)). The question why this parameter pair is suitable is discussed by Lawless (1987).

Hence, we now estimate mean and variance from the data as discussed in the main text, calculate the raw SCV from this (i.e., re-parametrise the distribution from being specified by mean and variance to being specified by mean and raw SCV), adjust the raw SCV estimate for the bias that this parametrisation caused, and then use the distribution thus specified as null distribution for the exact test. Our simulations support this approach; the p values become uniform this way.

Numerical bias removal. In order to adjust the raw SCV for the bias, we proceed as follows:

Let f_{mq} be a function that maps a true raw SCV value γ to the expectation of the estimate $\hat{\gamma} = (\hat{\sigma}^2 - \hat{\mu})/\hat{\mu}^2$. $f_{mq}(\gamma)$ approaches its limit for $q \to \infty$ very fast; the changes for $q \gtrsim 30$ are negligible for our purposes, and the values for small q only lead to a conservative overestimation of the variance. Hence, we precalculate f_{mq} for a fixed, large value of q, and all the values $m = 2, 3, \ldots, 15$, at a grid of values for γ, invert it and interpolate in order to bias-correct an estimate $\hat{\gamma}$. For $m \gtrsim 15$, f_{mq} is sufficiently close to the identity function to make a bias correction unnecessary for our purposes.

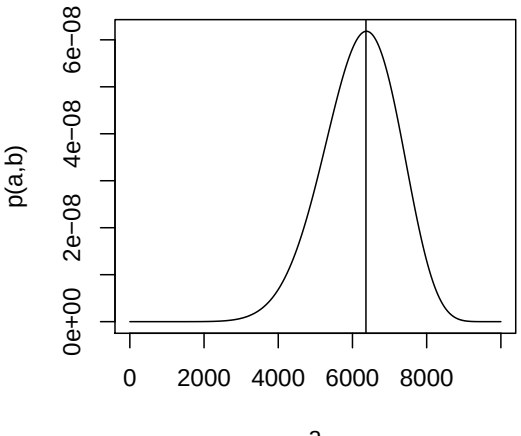

Figure S1: Shape of the function $p(a, b)$, with $k_S = 10,000$, $b = k_{AB} - a$, $\mu_A = 7,000$, $\mu_B = 4,000$, $\sigma^2_A = \mu_A + 0.1\mu^2_A$ and $\sigma^2_B = \mu_B + 0.1\mu^2_B$. The vertical line marks the estimate $k_S\mu_A/(\mu_A + \mu_B)$ for the mode.

D Numerical calculation of the p values

Evaluating the sums in Equation (11) requires some care. In HTS data, the count sum k_S can be large (e.g., millions of counts for a single strongly expressed gene), and calculating all the summands individually may take a long time and result in rounding error accumulation. Figure S1 shows the dependence of $p(a, b)$ (as defined in Section 3 and using Equation (14) for the distribution of K_A) on a for typical parameters. The function is unimodal, with mode approximately at ratio a/b equal to the ratio of the means of K_A and K_B. The function's simple shape allows the following numerical approximation: start at evaluating the sum from the peak (or rather, from its estimated location according to the means) and proceed outwards in two passes, first left, then right. During the summation, watch the changes of the value and keep adapting the step size according to a pre-defined precision goal. The value of p for the observed count values k_A and k_B is calculated beforehand, so that both the sum in the numerator and denominator of Equation (11) can be calculated in the same pass. To compute the density of the NB distribution, we use a function (Loader, 2000) in the C API of R (R Development Core Team, 2009).

E Diagnostics for the local regression

The choice of the gamma family for the local regression can be motivated as follows: If the size-adjusted counts k_{ij}/s_j in the sample variance estimate $w_{i\rho}$ calculated in Equation (7) were normally distributed with true variance σ^2_{ij}, the quantity $(m_\rho - 1)w_{i\rho}/\sigma^2_{ij}$ would follow a χ^2 distribution with $m_\rho - 1$ degrees of freedom, and this

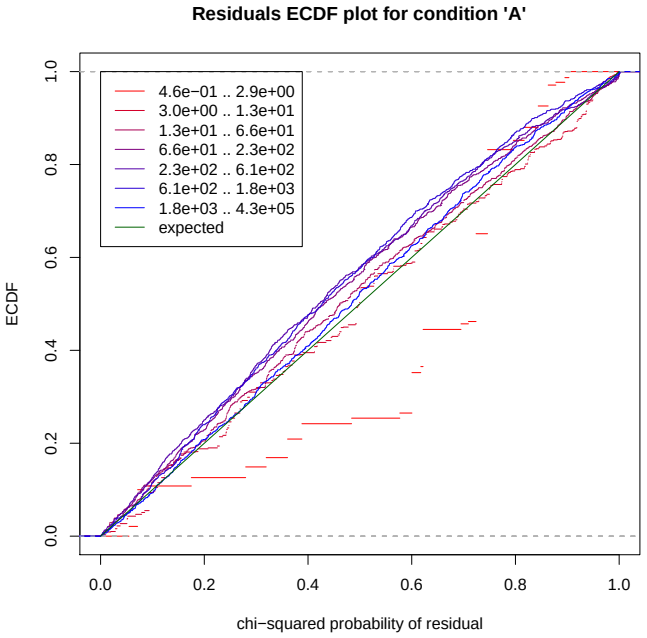

Figure S2: Empirical cumulative density function (ECDF) plots for the χ^2-probabilities of the residuals from the variance fit (orange line in Figure 1), stratified by the mean. The green line is the diagonal, which is the expected curve if the residuals follow the χ^2 distribution with $m_\rho - 1 = 1$ degree of freedom.

should hold as well for the residuals,

$$\xi_{i\rho} = (m_\rho - 1)\frac{w_{i\rho}}{w(\hat{q}_{i\rho})}$$

(where we have replaced the true variance σ^2 with its fitted value $w(\hat{q}_{i\rho})$. Even though the size-adjusted counts are not normally distributed, this is still a useful approximation for GLM local regression. Among the exponential families commonly used with generalised linear models, the gamma family, which includes the χ^2 distributions, is close to the actual distribution of the residuals, and since generalised linear models tend to show robustness against misspecification, we expect a reasonable fit. In order to verify this, we can check how well the residuals $\xi_{i\rho}$ follow a χ^2 distribution. To this end, we calculate the χ^2 quantiles of the $\xi_{i\rho}$ and check them for uniformity by plotting their empirical cumulative density function (ECDF). Figure S2 shows the ECDF curves for the condition ρ = GNS, stratified by the estimated means $\hat{q}_{i\rho}$. As one can see, the residuals follow the distribution reasonably well. Only for extremely low counts (below 5), the fitting quality is reduced. At such low counts, the shot noise dominates (see Figure 1b), and inaccuracies in the estimation of the raw noise are no reason for concern.

It is worth noting that the χ^2 distribution for $m_{GNS} - 1 = 1$ degree of freedom has a heavy right tail. Hence, the fact that in Figure 1 so many points lie far above the fitted line does not imply a bad fit.

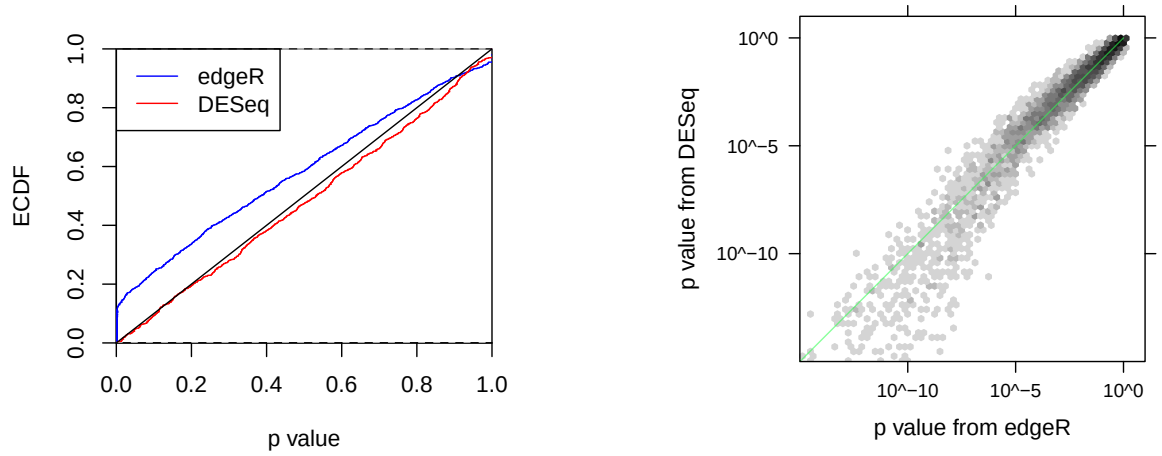

Figure S3: Simulation results. (*a*) Uniformity of the p values calculated for the genes that were not differentially expressed, shown with an ECDF plot. (*b*) Comparison of the p values between the *DESeq* and *edgeR* for the genes that were simulated as differentially expressed.

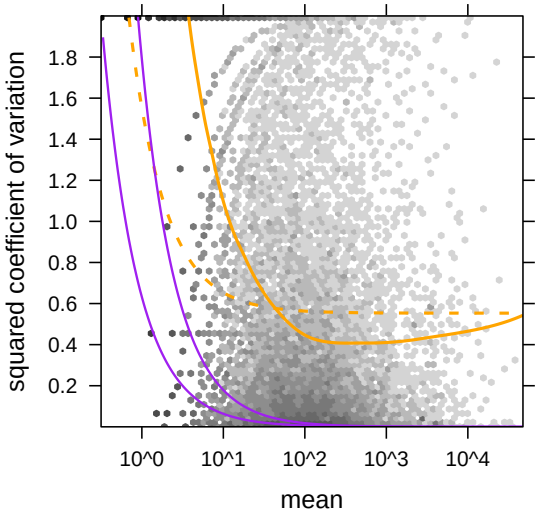

Figure S4: Estimation of the SCV for the neural cell data. Compare with the fly data (Figures 1 and S8. See caption to Figure 1 for further description.)

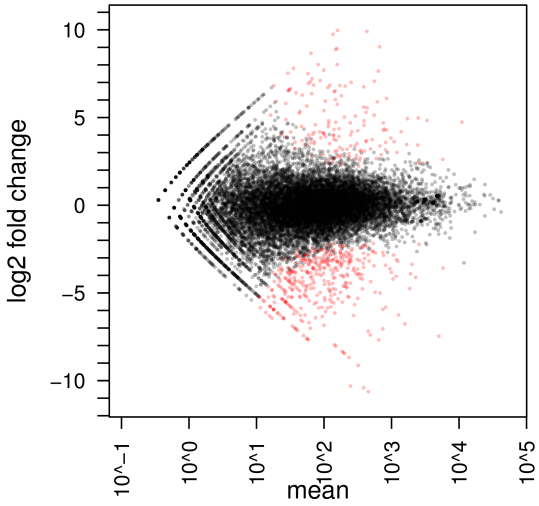

Figure S5: Testing for differential expression between conditions *GNS* and *NS* in the neural cell data. Compare to Figure 2.

F Differential expression analysis for the neural stem cell data set

.

In the main text, we have demonstrated our method with an RNA-Seq data set from *Drosophila* embryos. To show that *DESeq* is flexible in dealing with very different noise types, we performed the same analysis here for the Tag-Seq data set from Engström et al. (2010). The data set comprises of four tissue cultures derived from glioblastoma-derived neural stem cells (condition "GNS") with two tissue cultures derived from non-cancerous neural stem cells (condition "NS").

The number of reads obtained from each library varied from 7.6 millions to 13.6 millions. A good fraction of these (depending on the sample, from 32% to 53%) could be unambiguously assigned to annotated genes, and Engström et al. (2010) summarised the data in a table of counts with six columns for the six samples and 18,760 rows, one for each gene. For the differential expression analysis, we use only two of the four GNS samples (excluding one with a different histopathology and using only one of two samples taken from the same patient).

Each of these samples is taken from a different subject and hence has a different genetic background, giving rise to strong variability. This explains why in the SCV plot for this data set is by orders of magnitude stronger than in the fly case. (Compare Figure S4 with Figure S8.) Nevertheless, a test for differential expression yields useful results: at 10% FDR, *DESeq* calls significant differences for 680 of the 18,392 genes with non-zero counts

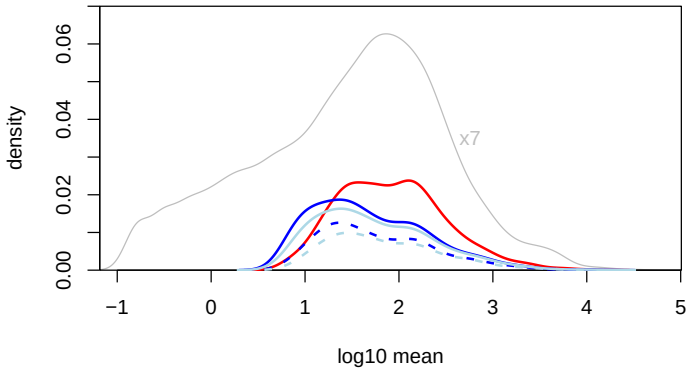

Figure S6: Same plot as Figure 4, now for the neural cell data. Again, red is *DESeq*'s and blue *edgeR*'s result. (light blue, using read count sum for library size adjustment; dark blue, using Equation (5); solid, with common dispersion; dashed, with tagwise dispersion.)

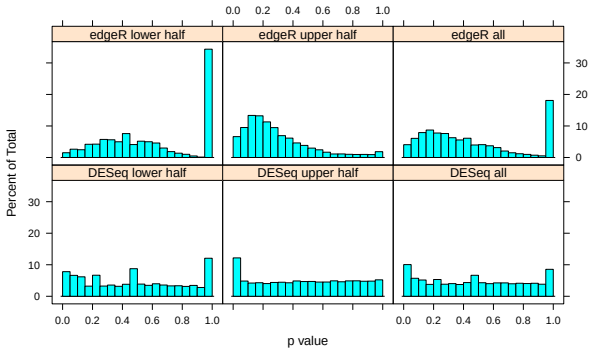

Figure S7: Type-I error control for the neural cell data. Compare with Figure 3.

(Figure S5).

We also compared the two *GNS* samples with the two *NS* samples using *edgeR*. *EdgeR* finds 452 genes when used in *common dispersion* mode and 256 in the *tagwise dispersion* mode. These numbers for *edgeR* were obtained when supplying it with total read counts as library size parameters, as recommended in the documentation; when *DESeq*'s estimates, as in Equation (5), were used, we obtained 525 and 316 genes, respectively. 84% to 96% of *edgeR*'s genes were also reported by *DESeq*, which is consistent with an FDR of 10%.

As with the fly data, the difference between the results of *edgeR* and *DESeq* does not merely lie in the number of genes, but also in their distribution along the abundance scale. We can recover all the observation we made for the fly data: Figure S6 shows that, again, *edgeR*'s hits tend to concentrate at lower abundance, while the hits from *DESeq* are more evenly distributed along the dynamic range, once the mean is above ~ 10. This agrees with Figure S4: *edgeR*'s estimate for the common dispersion (0.56), as indicated by the dashed orange line in Figure S4 is lower than *DESeq*'s estimate (solid orange line) for the lower part of the the dynamic range, and higher in

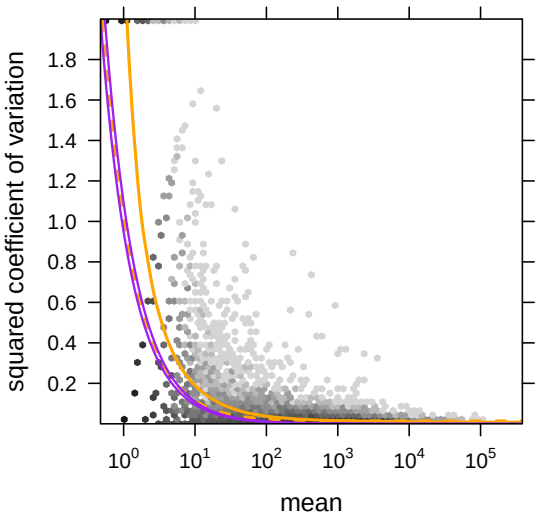

Figure S8: The same plot as in Fig. 1b, but here with the y axis spanning the whole data range.

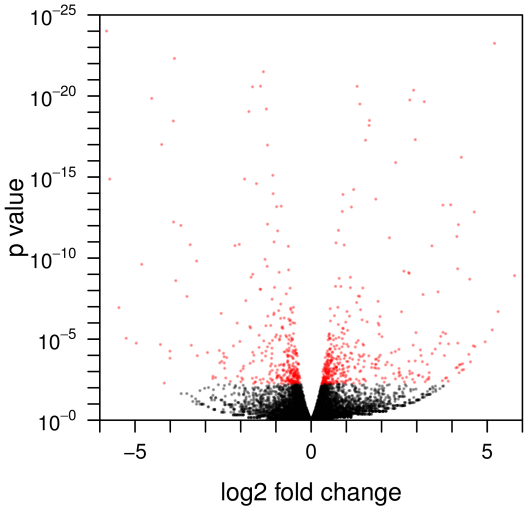

Figure S9: Volcano plot for the comparison of conditions A and B in the fly RNA-Seq data.

the upper range, causing *edgeR* to calls more hits among genes with low counts and be too conservative for genes with high counts. Also, a comparison of the two GNS replicates against each other, shows the same effect: As can be seen in Figure S7, there are either too many high or too many low p values, depending on the range of mean values.

The fact that these differences between *edgeR* and *DESeq* are observed consistently for two very different data sets suggests that our novel variance estimation scheme offers a real advantage over existing methods (of which *edgeR* is arguably the most advanced).

G Simulations

As a check of the correctness of *DESeq* and to further explore its performance in comparison to *edgeR*, we performed simulations. Here, we show a set of typical results

for simulation parameters chosen to resemble the situation in the neural stem cell data set.

We drew true mean values q_i for 20,000 genes from an exponential distribution with rate 1/250. Each gene was considered "truly differentially expressed" (tDE) with probability 30%, and for all tDE genes a $\log_2$ fold change was randomly drawn from a normal distribution with mean 0 and standard deviation 0.7. Finally, four count values were drawn for each gene, two for condition A and two for condition B, from negative binomial distributions, with the given means and variances as below, and multiplied by the size factors, which we chose as 0.5, 1.7, 1.4 and 0.9. For the variances, we catered to *edgeR*'s assumption and set the raw SCV to a constant, 0.015. (All these values were chosen to mimic the situation in the fly RNA-Seq data, except for the size factors, which are taken from the neural cell Tag-Seq data.)

We used both *DESeq* and *edgeR* to test for differential expression. *edgeR* was given the true size factors, while our approach had to estimate them from the data. In this setting, *edgeR* (running in common-dispersion mode) correctly estimated the raw SCV with good accuracy. Both approaches controlled the type-I error rate correctly: the percentage of type-I errors at 5% nominal significance level was (averaged over 10 simulation runs) 3.3% for DESeq and 3.5% for edgeR. (See Figure S3a for a plot with data from one run). At 10% FDR, DESeq discovered 42% of the truly differentially expressed genes, and edgeR found 44%. Finally, both methods stayed below the nominal 10% FDR with an actual FDR of 3.0% (DESeq) and 3.3% (edgeR).