Published online 20 April 2014 *Nucleic Acids Research, 2014, Vol. 42, No. 11* ***e91***
doi: 10.1093/nar/gku310

Robustly detecting differential expression in RNA sequencing data using observation weights

Xiaobei Zhou[1,2], Helen Lindsay[1,2] and Mark D. Robinson[1,2,*]

[1]Institute of Molecular Life Sciences, University of Zurich, CH-8057 Zurich, Switzerland and [2]SIB Swiss Institute of Bioinformatics, University of Zurich, CH-8057 Zurich, Switzerland

Received December 4, 2013; Revised March 10, 2014; Accepted March 31, 2014

ABSTRACT

A popular approach for comparing gene expression levels between (replicated) conditions of RNA sequencing data relies on counting reads that map to features of interest. Within such count-based methods, many flexible and advanced statistical approaches now exist and offer the ability to adjust for covariates (e.g. batch effects). Often, these methods include some sort of 'sharing of information' across features to improve inferences in small samples. It is important to achieve an appropriate tradeoff between statistical power and protection against outliers. Here, we study the robustness of existing approaches for count-based differential expression analysis and propose a new strategy based on observation weights that can be used within existing frameworks. The results suggest that outliers can have a global effect on differential analyses. We demonstrate the effectiveness of our new approach with real data and simulated data that reflects properties of real datasets (e.g. dispersion-mean trend) and develop an extensible framework for comprehensive testing of current and future methods. In addition, we explore the origin of such outliers, in some cases highlighting additional biological or technical factors within the experiment. Further details can be downloaded from the project website: http://imlspenticton.uzh.ch/robinson_lab/edgeR_robust/.

INTRODUCTION

RNA sequencing (RNA-seq) is widely used for numerous biological applications, including the detection of alternative splice forms, ribonucleicacid (RNA) editing, allele-specific expression profiling, novel transcript discovery but most commonly, for detecting changes in expression between experimental conditions or treatments. Compared to microarray technology, RNA-seq offers an open system, higher resolution, lower relative cost and less bias (1). A typical RNA-seq experiment includes: (i) capture of an RNA subpopulation (e.g. polyA-enriched, depleted of ribosomal ribonucleicacid) from cells of interest; (ii) reverse transcription into complementary DNA (cDNA); (iii) preparation and sequencing of millions of short cDNA fragments (∼200 bp); (iv) mapping to a reference genome or (assembled) transcriptome; (v) counting according to a catalog of features. This last counting step can be conducted by excluding ambiguous reads between genes (2), or with advanced tools that portion ambiguous reads to transcripts (3) or can be done in combination with assembly tools (4). The focus here is on methods for count-based differential expression (DE) analyses and the robustness thereof; thus, the starting point here is a count table of features-by-samples, such as those available from the ReCount project (5).

Considerable recent effort has been paid by the statistical community to the discovery of DE features, given a count table; recent comparisons have shown that no method dominates the spectrum of possible situations (6,7). RNA-seq remains expensive and in many cases researchers are studying precious samples or rare cell types, so the number of biological replicates is often limiting. It is clear that the most successful methods implement some form of 'information sharing' across the whole dataset to improve DE inference (2), and this becomes an intricate exercise to tradeoff power, false discovery control and protection against outliers. To highlight this distinction, we describe two popular software implementations for the negative binomial (NB) model, which arguably is the *de facto* standard for accounting for biological variability in such genome-scale count datasets. The latest version of edgeR moderates dispersion estimates toward a trended-by-mean estimate (8), whereas DESeq takes the maximum of a fitted dispersion-mean trend or the individual feature-wise dispersion estimate (9). The effect imposed on features with 'outliers' is illustrated in Figure 1. Ten randomly selected samples from individuals from the HapMap project (denoted as Pickrell (10)) are divided into two groups of 5, forming an artificial 'null' scenario. While very little true differential expression is expected, a low rate of false detections occur; in particular, edgeR detects a small number of genes with low estimated false discovery rate that exhibit one or two observa-

[*]To whom correspondence should be addressed. Tel: +41 44 635 48 48; Fax: +41 44 635 68 68; Email: mark.robinson@imls.uzh.ch

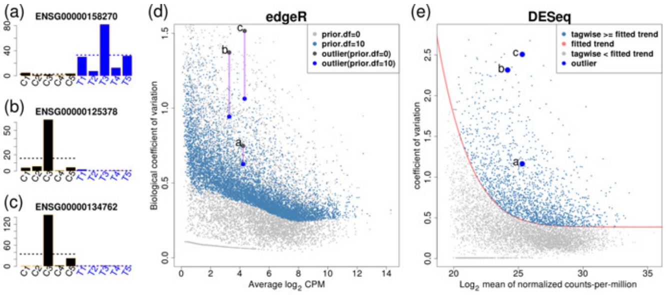

Figure 1. From Pickrell (10) data, 10 randomly selected samples from individuals are divided into two groups of 5, forming an artificial 'null' scenario. (**a**), (**b**) and (**c**) show barplots of log-counts-per-million (CPMs) of three genes from the top 10 DE genes with one or two extremely large observations. Dashed lines represent group-wise average log-CPMs. (**d**) and (**e**) plot genewise biological coefficient of variation (BCV) against gene abundance (in $\log_2$ counts per million) for edgeR and DESeq. In panel (d), gray dots show unmoderated biological BCV estimates ($\sqrt{\phi_i}\sqrt{\phi_i}$) (prior degrees of freedom = 0). Steel blue dots show moderated biological BCV with prior degree 10 (default setting for edgeR). Three outlier genes on (a), (b) and (c) are labeled by large blue dots. For (e), DESeq uses the maximum (steel blue dots) of a fitted dispersion-mean trend (red line) or the individual feature-wise (tagwise) dispersion estimate. Three outlier genes are also pointed out by large blue dots.

tions that are generally much higher in expression (Figure 1a–c). We believe that there are two causes for this: (i) the sensitivity of relative expression estimates to these 'outlying' observations; (ii) moderation of the dispersion estimates toward the trend. In contrast, DESeq remains largely unaffected by these outliers, since the dispersion estimation policy is to keep the maximum; in what follows, we will explore the effect of this maximum policy on power. All computed statistics for this dataset are stored in Supplementary Table S1.

The downstream effect of these dispersion estimation strategies suggest: (i) DESeq is generally conservative but robust; (ii) edgeR can be sensitive to outliers when there is sufficient dispersion smoothing toward the trend (effectively underestimating the dispersion in the shrinking process), but should be more powerful in the absence of such extreme observations (2). Our goal in the current study is to achieve a suitable middle ground, perhaps forfeiting a small amount in statistical efficiency, similar to established robustness frameworks, to reduce the influence of extreme observations in differential expression calls. As hinted above and in general, robustness is not solely determined by the dispersion parameter, but also by controlling the influence of outliers to other parameters in the model (e.g. those representing changes in expression). We explore these aspects in both simulated and real data, provide a extensible framework for evaluating the tradeoffs and highlight some instances of biology or technical factors that may give outliers.

The literature is rich in alternatives for count-based DE analyses and in particular, dispersion estimation, yet it remains increasingly difficult to assess the performance across the range of possibilities. For example, recent evidence suggests that one can suitably transform count data and analyze with methods developed for microarrays, with special treatment (11). The mainstream strategy is to directly fit count data to extensions of the Poisson model and in particular, the NB model. Many implementations are available as R/Bioconductor packages (12), such as edgeR (13), DESeq (9), ShrinkBayes (14), baySeq (15) and variations of dispersion estimation that can be used within existing implementations (16); the main differences lie in the estimation of the dispersion or in the inference machinery (e.g. Bayesian versus frequentist). Recent comparisons and summaries of the methods available can be found in (2), (6) and (7).

Some early and existing count-based DE analysis tools only allowed two-group comparisons. That is, they could not handle more complex situations, such as paired samples, time courses or batch effects. Recently, McCarthy *et al.* developed generalized linear model (GLM) capabilities in edgeR (8), allowing a much broader class of experimental designs to be analyzed and other frameworks have followed suit. However, GLMs require iterative fitting and more complicated dispersion estimation machinery (8). As shown in Figure 1, this framework can suffer a lack of robustness, whereby even a single extreme value (outlier) could largely affect estimates of regression parameters (e.g. mean of experimental condition), as highlighted by recent comparative studies (6) (see also Figure 1). In addition, the moderation of the dispersion parameter toward a trended value is actually contributing to the lack of robustness, forcing the dispersion to be underestimated (Figure 1). DESeq2 (successor of DESeq) takes an altogether different stance on robustness: using a Cook's distance metric, features that exhibit an extreme value are not considered for downstream statistical testing.

The strategy proposed in this paper is that of 'observation weights', effectively down-weighting outliers to dampen their influence. There is already some precedent for doing this in GLM settings: Carroll and Pederson (17) introduced weighted maximum likelihood estimators for the logistic model; Cantoni (18) presented a robust quasi-likelihood approach for inference in binomial and Poisson models; Agostinelli and Alqallaf (19) derived weighted likelihood equation for GLMs by directly inserting 'observation weights' into iterative re-weighted least squares algorithm (IRLS). Of particular importance, after adding observation weights, the asymptotic theory suggests that likelihood ratio statistics of model parameters still converge to approximate chi-squared distributions under the null hypothesis (20). At present, no 'off-the-shelf' robust approach is readily available for the negative binomial model in the context of genome-scale computations. In this paper, we build an outlier-resistant framework that maintains high power and achieves decent false discovery control and make it available in the edgeR software package; the same strategy could be employed in other frameworks. We benchmark its performance on real and simulated data and explore the origins of outlying observations.

MATERIALS AND METHODS

A standard setup of NB model in GLM framework

To most easily explain the addition of observation weights, we follow closely the notation used in McCarthy *et al.* (8). Let the Y_{gi} be the read count in sample i for feature g ($g = 1, \ldots, G$). Assume Y_{gi} follows a NB distribution with mean μ_{gi}

and dispersion ϕ_g, denoted by $Y_{gi} \sim \mathrm{NB}(\mu_{gi}, \phi_g)$. Feature g's variance equals $\mu_{gi} + \phi_g \cdot \mu_{gi}{}^2$, while the dispersion ϕ_g represents the square of the 'biological coefficient of variation' (8). In the GLM setting, the mean response, μ_{gi}, is linked to a linear predictor, here with the canonical logarithm link according to:

$$\log(\mu_{gi}) = X\beta_g + \log N_i, \tag{1}$$

where X is the design matrix containing the covariates (e.g. experimental conditions, batch effects, etc.), β_g is a vector of regression parameters (a subset of which are of interest for differential expression inference) and N_i is the (effective) library size for sample i.

For estimation of the regression parameters, maximum likelihood estimation is used. The derivative of the log-likelihood, $l(\beta_g)$, with respect to the coefficient β_g is $X^T zz_g$, where $zz_{gi} = (y_{gi} - \mu_{gi})/(1 + \phi_g\mu_{gi})$. The estimated value of β_g can be obtained by the IRLS in the form:

$$\beta_g^{\text{new}} = \beta_g^{\text{old}} + (X^T \Omega_g X)^{-1} X^T z_g, \tag{2}$$

where $X^T \Omega_g X$ is the Fisher information matrix (also denoted below as $\mathcal{I}_g\mathcal{I}_g$) and Ω_g is the diagonal matrix of working weights, which are $\mu_g/(1 + \phi_g\mu_g)$ for the NB model.

Moderated and trended dispersion estimates

The adjusted profile likelihood (APL) introduced by Cox and Reid (21) has shown good performance for dispersion estimation in the context of genome-scale count data (8,22). The APL_g is a likelihood in terms of ϕ_g, penalized for the estimation of the regression parameters, β_g, as follows:

$$\mathrm{APL}_g(\phi_g) = \ell(\phi_g; \mathbf{y}_g, \hat{\beta}_g) - \frac{1}{2}\log|\mathcal{I}_g|, \tag{3}$$

where $\mathbf{y}_g\mathbf{y}_g$ is the vector of counts for gene g, $\hat{\beta}_g\hat{\beta}_g$ is the estimated coefficient vector, $\ell(\cdot)$ is the log-likelihood function, $\mathcal{I}_g\mathcal{I}_g$ is the Fisher information matrix and $|\cdot|$ is the determinant. The early strategy to accomplish moderation for the dispersion was by squeezing the tagwise dispersion toward a common dispersion that is estimated over all features (23). This weighted likelihood approach involves maximizing a linear weighting of the individual likelihood and the common (averaged) likelihood, the two terms, respectively, in

$$\arg\max\left\{\mathrm{APL}_g(\phi_g) + \alpha\cdot\frac{1}{G}\sum_{k=1}^{G}\mathrm{APL}_k(\phi_g)\right\}, \tag{4}$$

where α is a suitably chosen weight.

A slight variation on this, which is now commonly applied after experience in many datasets showing a dispersion–mean relationship, is to shrink toward a dispersion estimated from features with similar average expression level (8). This so-called trended dispersion is constructed using local shared log-likelihood for feature g (more precisely, a smooth fit to common dispersions that are calculated in bins of averaged counts per million) and its neighboring features in terms of expression strength. Specifically, individual tagwise estimates for each feature can be estimated by maximizing a linearly weighted function between individual dispersion and local shared dispersion:

$$\hat{\phi}_g = \arg\max\left\{\mathrm{APL}_g(\phi_g) + \gamma\cdot\mathrm{APL}_g^S(\phi_g)\right\}, \tag{5}$$

where $\hat{\phi}_g\hat{\phi}_g$ is moderated tagwise dispersion, γ is the prior degree of freedom afforded to the shared likelihood and

$$\mathrm{APL}_g^S(\phi_g) = \frac{1}{|C|}\sum_{k\in C}\mathrm{APL}_k(\phi_g), \tag{6}$$

where the set C represents features that are close to feature g in average log counts per million.

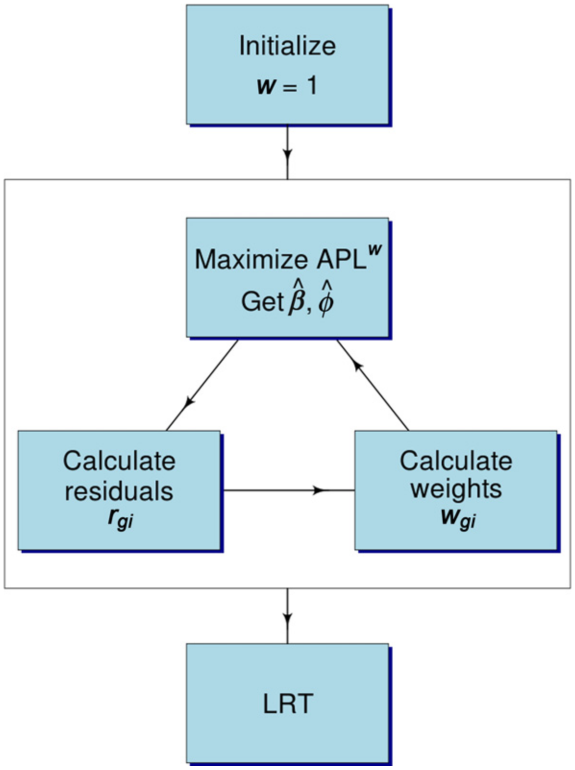

Figure 2. The flow chart of the robust algorithm implemented in edgeR. $\hat{\beta}\hat{\beta}$ is the estimated GLM regression coefficient and $\hat{\phi}\hat{\phi}$ is the moderated dispersion estimate by maximizing APLww (Equation (10)). r_{gi} is the Pearson residual corresponding to count y_{gi} from Equation (7).ww_{gi} is the observation weight from Equation (8). LRT (glmLRT in edgeR) computes likelihood ratio tests using the weights.

A robust negative binomial GLM

Our approach to induce robustness is to attach a weight to each observation; observations that deviate strongly from the model fit are given lower weight. In particular, Pearson residuals from the current fit are sent through a weight function, which gets passed to the next iteration of estimation. The dispersion estimation machinery (i.e. trended APL) also receives the same observation weight, so that the influence of outliers is dampened on both the regression and dispersion estimates.The robust iterative estimation procedure using weights is described in Figure 2. The Pearson residual of an observed count y_{gi} from the NB GLM fit can be calculated as

$$r_{gi} = \frac{y_{gi} - \hat{\mu}_{gi}}{\sqrt{\hat{\mu}_{gi}(1 + \hat{\phi}_g \hat{\mu}_{gi})}} \qquad (7)$$

where $\hat{\mu}_{gi}\hat{\mu}_{gi}$ is the fitted value (from $\hat{\beta}\hat{\beta}$) and $\hat{\phi}_g\hat{\phi}_g$ is the moderated dispersion estimate. The Pearson residual is converted to weights using, e.g. the Huber function:

$$w_{gi} = w(r_{gi}) = \begin{cases} \frac{k}{abs(r_{gi})}, & \text{for abs}(r_{gi}) > k \\ 1, & \text{for abs}(r_{gi}) \leq k \end{cases} \qquad (8)$$

where k represents a tuning constant for Huber estimator and is usually set to 1.345 in normally distributed settings to achieve 95% efficiency (24). This weight, ww_{gi}, gets used in the next iteration of GLM fitting; the IRLS equation becomes:

$$\beta_g^{\text{W-new}} = \beta_g^{\text{W-old}} + (X^T[W_g\Omega_g]X)^{-1} X^T[W_g]z_g \qquad (9)$$

where W_g is the diagonal matrix of observation weights for feature g. The Fisher information matrix with observation weight becomes $\mathcal{I}_g^W = X^T[W_g\Omega_g]X\mathcal{I}_g^W = X^T[W_g\Omega_g]X$. In this approach, the APL for dispersion ϕ_g with observation weights can be written as

$$\text{APL}_g^W(\phi_g) = \ell^W(\phi_g; \mathbf{y}_g, \hat{\beta}_g) - \frac{1}{2}\log|\mathcal{I}_g^W|, \qquad (10)$$

where $\ell^W(\cdot) \equiv \sum_i ww_{gi} l(\cdot)$ is the weighted log-likelihood function and $\mathcal{I}_g^W\mathcal{I}_g^W$ is the Fisher information matrix with observation weights. Then, using these dispersion estimates, the regression parameters are estimated, again using the observation weights.

For users of edgeR, only a small change in the standard pipeline is required.

A simulation framework with parameters based on the joint distribution of mean and dispersion estimates from RNA-seq data

We built a simulation framework that aims to accurately reflect the reality of RNA sequencing data. In order to evaluate the performance of our robust method and other methods across a variety of reasonable conditions, we created several options:

(1) nTags: total number of features,
(2) group: factor containing the experimental conditions,
(3) pDiff: proportion of DE features,
(4) foldDiff: relative expression level of truly DE features,
(5) pUp: proportion of DE features that increase in expression,
(6) dataset: dataset to take model parameters from,
(7) pOutlier: proportion of outliers to introduce,
(8) outlierMech: outlier generation mechanism to use.

We generate true NB model parameters, μ and ϕ, using the joint distribution of estimates, $\hat{\mu}\hat{\mu}$ and $\hat{\phi}\hat{\phi}$, estimated using edgeR from real datasets, such as the published count tables at ReCount (5): Pickrell (10), Cheung *et al.* (25,26). In particular, the joint distribution preserves the dispersion–mean trend, which can vary from dataset to dataset. After the removal of extremely high dispersions and low means (analogous to typical recommended filters; see Supplementary Figure S1), the derived-from-real-data parameters are used to simulate the counts, from a NB distribution and optionally with true DE.

To test robustness, we add outliers to the simulated counts. Outliers are large values and can be produced by two different mechanisms (outlierMech): first, counts are multiplied by a random factor between 1.5 and 10, as employed by Soneson and Delorenzi (6), and includes both the 'simple' (S) and 'random' (R) method. In S, a gene is chosen at some probability to have a single outlier randomly added. In R, each observation can become an outlier with some probability. In the second mechanism, called 'model' (M), each observation can become an outlier with some probability and if so, is sampled from a second NB distribution with larger μ (original μ multiplied by random factor between 1.5 and 10); R and M methods induce the same overall outlier rate.

Recently, van de Wiel *et al.* modeled genome-scale count data as zero-inflated negative binomial model (ZINB), which seemed to explain some of the dispersion–mean relationship (4). We have not considered simulations from ZINB distributions, since they do not appear to explain all of the observed dispersion–mean relationship in the datasets that we tested (see Supplementary Figure S2).

Methods compared

We evaluated and compared several methods for DE analysis, including edgeR, edgeR-robust, limma-voom, DESeq-pool, DESeq-glm, DESeq2, baySeq, SAMseq (27), EBSeq (28) and ShrinkBayes; the performance evaluation system that we developed allows arbitrary additions (assuming they are implemented in R). limma-voom is an extension to DE analysis of RNA-seq count data from limma (11); it transforms the count data with special treatment given to fitting the mean–variance relationship. DESeq is tested as two separate methods: DESeq-pool is the default setting method to estimate the empirical dispersion from all the conditions with replicates; DESeq-glm fits models according to a design matrix and estimates dispersion by maximizing APL. edgeR, DESeq and DESeq2 differ in how the dispersion is estimated: edgeR moderates dispersion toward a trended estimate (8), edgeR-robust expands this with observation weights, DESeq takes the maximum of a fitted trend of dispersion or the individual feature-wise dispersion estimate (9). DESeq2 offers a zero-mean normal prior on the log-fold-changes for moderation and a proper moderation of dispersion estimates to a trended value, except when the feature exhibits variability much greater than other features at the same expression strength; for outlier protection, a Cook's distance is calculated and those features with an extreme value are not promoted to formal statistical testing i.e. *P*-values are set to NA; in our simulations, these *P*-values are set to 1 so as to not remove features. The default normalization method is also different among edgeR, DESeq and DESeq2. edgeR uses trimmed-mean-of-M-values (TMM) (29), while DESeq and DESeq2 use a relative-log-expression approach. SAMseq, a non-parametric method, employs Wilcoxon rank-sum statistics to estimate false discovery rate (FDR) through sample permutations.

baySeq, EBSeq and ShrinkBayes use Bayesian inference. baySeq employs the NB model and assumes that samples can be classified as different groups by their treatment conditions; samples within the same group should follow the same distribution and share parameters. Using an empirical Bayes approach, baySeq estimates the posterior probability of the null state. ShrinkBayes introduces the ZINB and performs inference using integrated nested Laplace approximations (INLA) (30,31) and provides Bayesian FDR and local false discovery rate (lfdr) (32) estimates. Since the computational cost of ShrinkBayes is high, some comparisons are skipped. EBSeq is similar to baySeq, providing posterior probability of DE, as well as EE (equally expressed), based on a parametric mixture model. Compared with other methods tested here, EBSeq can also detect DE isoforms in EE features, yet this is not our primary question here.

Notably, new methods, or variations of existing ones can be easily added to our comparison framework, simply by providing a wrapper to an R function that contains the correct inputs (count table, grouping variable) and outputs (*P*-values). See Supplementary web site for details.

Comparison metrics

To test the performance of each DE method in the presence of outliers, we employ several standard metrics and plots: false discovery (FD) plots, receiver operating characteristic (ROC) curves, partial ROC curves and power curves. Power (TP) curves and (partial) ROC curves (i.e. up to a certain false positive rate) evaluate the ability to distinguish, through statistical evidence, DE and non-DE. FD procedures gauge the control of the expected proportion of incorrectly rejected null hypotheses (33). Another useful plot is the relationship between TP rate and achieved false discovery rate across multiple thresholds.

An open graphical tool and R code for re-analysis: evaluating DE analysis methods

One disadvantage of current method comparisons (e.g. (6,7)) and those that accompany every new method published, is that they are a snapshot in time. If new methods come along, the developer must demonstrate that their method is better, by some metric. This task is important but somewhat repetitive, because many of the same metrics, plots and simulation models are (re-)implemented. We endeavored to create a system for performing standardized simulation-based testing.

In addition, all analyses presented in this paper are freely available from our website. Moreover, our simulation and evaluation framework is made available as a web-sourceable script that consists of three modules: simulation, evaluation (running of the software packages) and metric computation. Each module can be extended, using simple wrapper functions to existing R-based code, ensuring that our comparison results are reproducible, extensible and relatively easy for the user to track exactly what code segments (and versions) were run.

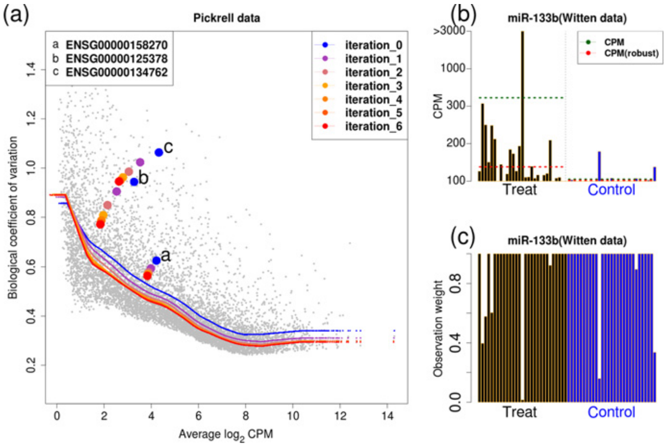

Figure 3. (**a**) For the random 5 versus 5 split of the Pickrell data (10) shown in Figure 1, the trajectories of overall trended dispersion and for the three individual genes are shown over six iterations of the edgeR-robust re-weighted estimation scheme. (**b**) A bar plot of miR-133b expression from Witten *et al*. (25), including an observation with very high count. (**c**) weights for miR-133b after six iterations of the re-estimation from edgeR-robust. Dashed lines in panel (b) shown the group-wise CPM before and after weighting.

In addition to R code, we make available a web-based shiny ‘app’ that can be used to look at simulation results across a wide number of conditions (34). Since there are often too many methods to be easily displayed together, our app gives users the ability to present results for a user-selected subset of methods; the results update automatically as the user selects different simulation settings.

Functional category analysis for outliers

To explore potential biological or technical factors that may manifest as outliers, we performed hypergeometric-based functional category analyses on the set of genes with weights less than some cutoff (here, set to 1) separately for each sample. Our goal with such an analysis is to identify possible biological or technical factors that affect a subset of genes for a particular experimental unit. In some cases, this may shed light on why the expression levels of some genes for a given sample are very different than that of their replicates. Furthermore, we can investigate whether the down-weighting is driven by technical factors. As a positive control for this, we compared the observed weights to the sample-specific guanine-cytosine (GC) effects observed in the Pickrell dataset (10,35).

RESULTS

edgeR-robust dampens the effect of outliers

To highlight how edgeR-robust dampens the influence of outliers, we return to the dataset shown in Figure 1. Figure 3a shows the trajectories for the three outliers in terms of their average log-CPM and dispersion estimates and how the dispersion-mean trend changes over six iterations of the edgeR-robust re-weighted estimation scheme. Although we have not studied convergence in depth, Supplementary Figure S3 highlights the change in parameter estimate by iteration; most features 'converge' after a small number of iterations and we use a fixed number of iterations as a stopping rule. As expected, the outliers appear 'extreme' according to the model, as also reflected by their residuals. Extreme residuals are then down weighted, iteratively, and both the dispersion and average log-CPM estimates are updated (Figure 3a). In particular, we notice large changes to the regression (e.g. log-fold-change) and dispersion parameter estimates, which impose better accordance, in terms of dispersion–mean relationship, with the other features in the dataset. Notably, Figure 3a highlights a global drop in dispersion-mean trend after the iterative robust estimation, which suggests that outliers present in sufficient frequency may have a global effect on the statistical detection of DE within a dataset. Thus, we speculate that gains in statistical power (see sections below) may be achieved in part by this global drop in trended dispersion.

In their manuscript, Li and Tibshirani (27) show some extreme examples of outliers affecting differential count analysis of miRNA-seq data (in particular, see their Figure 2). Figure 3b shows one of those examples, mir-133b, and highlights the estimated mean CPM by group, before and after down-weighting; the observation weights after six iterations are shown in Figure 3c. Notably, for this example, there still exists strong evidence for differential expression, even after careful reassessment of the outlying observations.

Supplementary Table S1 gives the full details of these analyses, before and after re-weighting.

Simulation reflects real data

To test the method on a wide range of simulated settings, we first generate count data from a model that reflects real data as well as possible. As described in the 'Materials and Methods' section, we choose to take the joint distribution of estimated log-CPM and dispersion from a large dataset as the basis for the parameter settings and we use library sizes that mimic those from typical datasets. For example, the Pickrell dataset (10) consists of >50 replicates, which should represent a reasonably accurate reflection of the range of abundances observed, as well as, in particular, the dispersion–mean relationship. We generate all data from the NB model and introduce outliers by various mechanisms (see 'Materials and Methods' section). Supplementary Figure S4 shows the dispersion-mean trend for the Pickrell dataset (top left) and an example simulated dataset based on the estimated parameters (top right), respectively, as well as the marginal distributions of both log-CPMs and dispersion. The framework for these simulations (see 'Materials and Methods' section) is designed to take an initial dataset that seeds the simulation parameters, so datasets spanning the range of biological variation could easily be tested. Notably, we explored the Pickrell dataset for both the frequency of outliers (as detected by down-weighting; Supplementary Figure S5 gives cumulative distributions of weights) and the magnitude of the outliers relative to non-down-weighted observations (Supplementary Figure S6) to justify the use of simulation parameters. In particular, we note that the range of outlier deviations is within the range we use (e.g. multiplication factor between 1.5 and 10; Supplementary Figure S6). Meanwhile, samples from the Pickrell dataset exhibit outlier rates of 2–10% 'per sample' (depending on where a weight threshold is set), suggesting our choice of 10% (of features with a single outlier) is in fact a conservative amount of outliers that may be present.

Standard metrics across various methods for various simulation settings

Next, we present a representative simulation and performance results under a single 'reasonable' setting of the parameters. We sampled NB model parameters μ and ϕ from the joint distribution of estimates from the Pickrell data (10) (dataset); we filtered out the top 10% of the extreme dispersion values (analogous to filtering; see Supplementary Figure S1); 10 000 features were generated (nTags), with a 5 versus 5 two-group comparison (group); 10% of them are defined as DE genes (pDiff=.1), symmetrically (pUp=.5) with fold difference 3 (foldDiff=3); outliers are introduced to 10% of the features (pOutlier=.1) using the 'simple' outlier generation mechanism (outlierMech="S"); outliers are randomly distributed among all features; further details are described in the 'Materials and Methods' section. Original simulated counts and the counts with outliers introduced are separately recorded and all methods were run on both.

Figure 4 shows the set of standard metrics: panels (a)–(c) and (d)–(f) show false discovery plots, ROC curves and power numbers, respectively, for the original and original-with-outliers datasets under the setting of simulation parameters discussed above. Overall, the introduction of outliers results in more false positives (Figure 4a versus d) and/or less true positives at the same false positive rate (Figure 4b versus e). In the absence of outliers, all methods exhibit similar patterns of false discovery rates, with the Bayesian methods, ShrinkBayes and EBSeq having a slightly higher rate. Similarly, in terms of separating the truly DE from non-DE features using a *P*-value (or *P*-value-like score in the case of Bayesian methods), all methods are very close in performance. Furthermore, in the absence of outliers, edgeR, edgeR-robust and DESeq2 appear to have a slight edge in power at the method's 5% FDR, albeit the advantage is small (Figure 4c). When outliers are introduced, edgeR-robust shows some advantages over edgeR. In terms of statistical power, all methods drop in overall power with the introduction of outliers (Figure 4c versus f), while DESeq exhibits a spectacular drop. Notably, DESeq still maintains a good ranking of *P*-values (Figure 4f), but becomes very conservative due to the maximum-of-trend-and-individual dispersion policy; in this respect, presence of outliers affect the whole dataset (see Supplementary Figure S7).

Since the direction of differential expression and the outlier introduction are applied at random, we can further split the DE features according to the position of the outlier relative to the direction of change in abundance (Figure 4g–i); 'DEupOutlier' represents the situation where the outlier is added to the higher expressed group; 'DEdownOutlier' represents those features where the outlier was added to the lower expressed condition; 'DEnoOutlier' represents DE features with no introduced outlier). Notably, edgeR shows the highest power in the 'DEupOutlier' setting, but this is artificial since the introduction of the outliers ac-

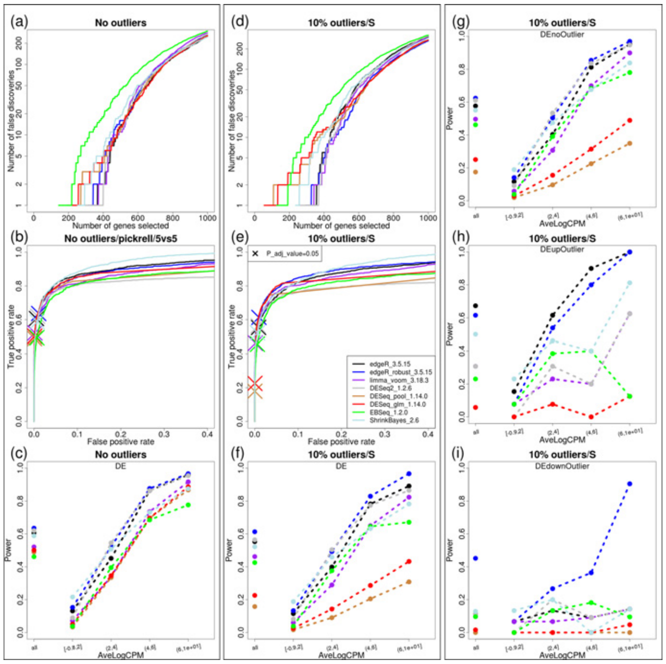

Figure 4. (**a**), (**b**) and (**c**) present FD, partial ROC (up to FP rate of 40%) and power plots (at each methods' 5% FDR) across several tested methods for datasets with no introduced outliers; (**d**), (**e**) and (**f**) show corresponding plots with datasets containing 10% outliers (i.e. 10% of genes have a single outlier) using 'S' method. (**g**), (**h**) and (**i**) split the results from panel (f) into three categories: features without outliers (g); outliers in the higher expression group (h); outliers in the lower expression group (i). All power results are shown as overall (single dot on the left of the plot) and split across five equally-sized average-log-CPM groups. The X on panels (b) and (e) highlights the achieved power (TP) according to each method's 5% FDR cutoff. Note that while panel (g) presents the situation with no outliers, there are outliers present in other features within the dataset and is therefore different from panel (c).

tually helps detection. The 'DEdownOutlier' is the situation where edgeR-robust comes to the forefront, as expected, given that outliers strongly eliminate the differential expression. In the absence of outliers, edgeR-robust still remains a strong competitor, closely followed by DESeq2, ShrinkBayes, limma-voom and edgeR.

It is also interesting, as a byproduct, to consider how well the methods identify outliers. In particular, we compared edgeR-robust's observation weights (using both Pearson and Deviance residuals) with DESeq2's Cook's distance metric (both at observation level and feature-wise maximum) to separate the simulated outliers. Supplementary Figure S8 shows an ROC curve depicting how well the observation weights (and other scores) separate outliers from non-outliers. Similarly, the default setting of DESeq2 leads to a similar tradeoff between false positives (here, falsely detected as an outlier) and false negatives (failing to identify an outlier) and Pearson residuals appear superior and are used for all further analyses with edgeR-robust. Notably, the edgeR-robust strategy smoothly identifies outliers and down-weights them according to the magnitude of discordance, instead of setting a hard threshold where statistical tests are no longer conducted. One byproduct of DESeq2's hard threshold is a loss of power (e.g. Figure 4 panels h and i), since genes with true differential expression as well as outliers are excluded from statistical testing. We also tested DESeq2 after turning off the Cook's distance metric, which results in an expected sensitivity to outliers (Supplementary Figure S9). Although the focus has been on higher-in-magnitude outliers and indeed that is what we see more of (Supplementary Figure S6), lower outliers can be sufficiently detected and down-weighted (e.g. Supplementary Figure S10).

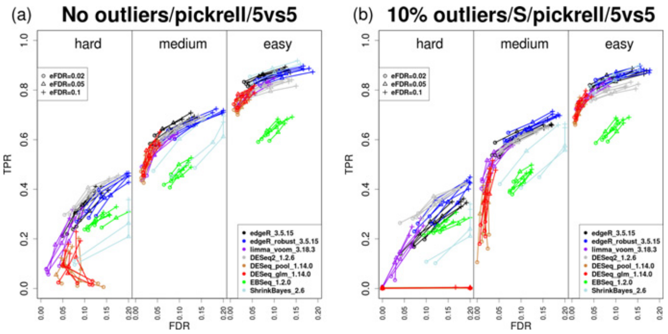

Figure 5. Power-to-achieved-FDR across hard (foldDiff ∈ [2, 2.2]), medium (foldDiff ∈ [3, 3.3]) and easy (foldDiff ∈ [6, 6.6]) simulation settings. (**a**) No outliers; (**b**) 10% outliers. Y-axis shows TP rate and X-axis shows FD rate. Five simulations are shown for each method and each setting. Points are taken according to each method's FDR cutoffs at 0.02, 0.05 and 0.1.

A shiny app to display pre-computed simulation results

The above discussion was in regard to a single dataset under a single set of simulation parameters. To provide a much wider scope of simulation settings, we created a web-based shiny app, that serves up pre-computed results over a range of simulation parameters, including different datasets, sample sizes and so on. In addition, it allows users to plot results for only the subset of desired methods and metrics from Figure 4. While new methods can only be added to the shiny app by us, existing simulations can be easily recreated in a local R environment or additional settings can be added, as described in the Supplementary Note. In general, the conclusions observed from the broader range of simulation settings (e.g. different magnitudes of DE, sample sizes) are in agreement with those mentioned above (see also Supplementary Figure S11).

Across multiple simulations over a range of settings, edgeR-robust is somewhat liberal but maintains a strong power-to-achieved-FDR tradeoff

To complement the simulation results for individual parameter settings, we endeavored to create a compact summary of a wider range of simulations and explore another important aspect of the comparison: do methods accurately control false discovery rate? Figure 5 shows a series of 15 simulations divided into three different blocks based on the degree of difficulty: 'hard' (foldDiff ∈ [2, 2.2]), 'medium' (foldDiff ∈ [3, 3.3]) and 'easy' (foldDiff ∈ [6, 6.6]), including five simulations within each group to illustrate sampling variability. For each dataset, lines connect the true positive rates and achieved FDRs across three thresholds of the estimated

FDR (0.02, 0.05, 0.1). The rest of the simulation parameters are kept fixed: the NB model parameters originate from the Pickrell dataset (10), there are 10 000 features, we consider a two-group comparison (5 versus 5), 10% of features are DE and each dataset contains 10% 'S' outliers; comparisons for 3 versus 3 and 10 versus 10 are shown in Supplementary Figure S12.

Overall, there is a broad range of power-to-achieved-FDR tradeoffs and no method dominates. EBSeq appears to lack power and can be both liberal and conservative. In general, DESeq is conservative and achieves lower power, as reported earlier (6). Altogether, the collection of methods, such as limma-voom, edgeR, edgeR-robust and DESeq2 achieve similar power-to-achieved-FDR tradeoffs across the sample sizes, with perhaps a tendency to be more liberal in large sample sizes for edgeR and edgeR-robust. As expected and as highlighted above, edgeR-robust appears to have advantages in the presence of outliers, with only a minor decrease in power when no outliers are present. Thus, edgeR-robust achieves a good tradeoff between power at the same achieved FDR, even if the target FDR is not quite met. Notably, DESeq2 offers a small advantage in power at low log-fold-changes while suffering a little bit in power for higher log-fold-changes. In all cases, limma-voom controls FDR well and maintains high power.

Outliers may originate from technical or biological sources

While the strategy based on observation weights appears useful for dampening the effect of outliers in differential expression analysis, it may also be of interest to investigate the origin of such outlying observations. In some cases, we know of technical artefacts that affect the profile of RNA-seq expression data, such as sample-specific GC content biases, as highlighted and mitigated by the analyses of the HapMap consortium as well as in follow-up methodology development (e.g. conditional quantile normalization (35)). In this dataset, there are no experimental conditions to detect differential expression, so we fit an intercept-only model, using the iterative robust estimation scheme. Not surprisingly, we first observe that the two samples highlighted by Hansen *et al.* also exhibit a relatively higher number of down weighted observations (Figure 6a). As expected, the degree of down-weighting is strongly related to the GC content of the cDNA sequences of the genes involved (Figure 6b).

In an unrelated dataset from Blekhman (36) comparing expression in human male and female livers, we observe that the most significantly overrepresented functional categories were strongly associated with the set of down-weighted genes from a single sample (SRX014822and3, green circles in Figure 6c). These include several categories involving the extracellular matrix, as well as collagen catabolism and plasma membrane. We show the third most overrepresented category, 'extracellular matrix' (Figure 6c) because the size of this category allows individual genes to be visualized (further details are given in Supplementary Table S2). Although we cannot confirm the exact cause of the overrepresented gene ontology categories, we note that accumulation of collagen and excessive production of extracellular matrix proteins are associated with the development of liver fibrosis (e.g. 37,38), and we suggest that analyses such as these may assist biologists in identifying the source of outliers in gene expression.

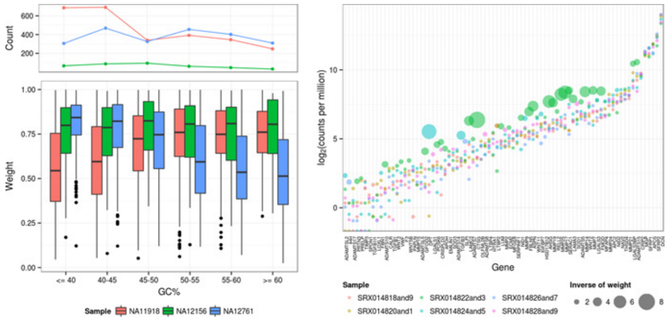

Figure 6. Technical ((**a**) and (**b**)) and biological (**c**) sources of outlier genes. The number of down weighted observations (a) and distribution of outlier weights as a function of the gene GC% in three samples from the HapMap RNA-Seq data (10) are plotted (b). Two of the samples shown (NA11918 and NA12761) were shown by Hansen *et al.* to have strong, opposing relationships between GC% and mapped reads per kilobase per million reads (RPKM). The third sample (NA12156) had the least number of genes down weighted after applying our robust down weighting procedure. (c) The log(CPM) and the inverse of the down weighting value for genes in the 'extracellular matrix' gene ontology category, where a value of one indicates no down weighting and larger inverse weights indicate stronger down weighting.

DISCUSSION

Various method developers have shown that statistical methods for discerning differential expression from RNA-seq data represented as counts can be sensitive to outlying observations. In this report, we have studied in detail the effects of outliers on various approaches and developed a new method based on observation weights that can detect and dampen the effect of outliers. In fact, it requires a delicate tradeoff to maintain high power while at the same time achieving a decent resistance to the presence of outliers. In particular, it is difficult to know exactly what an outlier is and where the line should be drawn to identify it as such. In this respect, we take a 'smooth' approach of dampening their effects, when there is evidence to support departure from the model. We have also explored the origin of such outliers and in some cases, we may be able to identify either a technical or biological effect to explain them. Our robust approach follows the strategy of classical robustness methods that are commonly applied to the linear regression problem. In our approach, we adopted the calculation of the residuals and observation weights to the specifics of the flexible dispersion estimation and standard GLM regression estimation of the negative binomial model.

As mentioned above, one reason that edgeR is sensitive to outlying observations is that the dispersion estimate used in the downstream inference is pulled toward the dispersion-mean trend, which may underestimate the variability. Therefore, another way to dampen the effect of outliers is to decrease the degree of moderation toward the dispersion-mean trend. Although we have not studied it here, there is again a delicate tradeoff between the degree of moderation to use and the average inference performance;

it still remains an open question as to how exactly to set this value for a given dataset.

Though motivated and tested on real datasets, we employed simulations to explore the broad range of possible settings and developed a comprehensive system for such evaluations. Our strategy to mimick real datasets is to take the joint distribution of mean and dispersion estimates from a large dataset as the basis for parameters to sample from. From such a dataset, outliers and differential expression at a specified level can be readily introduced. In fact, because these are estimates and not true values, we expect the sampled dispersion to potentially exhibit more variation than observed in a real dataset. In terms of evaluating the different methods across the spectrum of simulation settings, it is important to consider it from all points of view: false discoveries amongst the list of top called features, the ability to separate the truly differential from non-differential (i.e. ranking by statistical evidence), the statistical power at thresholds that are typically used in practice and the degree to which methods achieve their purported false discovery rates.

Overall, the observation weight robust method performs well and achieves the goal of suffering only minimal loss of power, while maintaining resistance to introduced outliers. We have investigated the outlier policy in other packages and highlight that smoothly down-weighting outlying observations appear preferable. In DESeq, a hard line against outliers is taken by using the maximum of a dispersion-mean trend and the individual estimate; with the addition of outliers, this has a global effect of increasing the variance to all features and gives a resulting loss of power. In DESeq2, a Cook's distance metric is used to remove features with outliers entirely from further consideration; in this case, features that have outliers and differential expression are excluded, with a potential loss of power. It is somewhat of a philosophical decision as to whether to completely filter out features or to down-weight them; the observation weight strategy allows both.

Another important consideration is the required sample size to be able to achieve estimators that are resistant to outliers. Indeed, the lowest levels of replication (e.g. two samples per condition) will not be sufficient. The minimum level of replication to dampen effects of outliers is three samples per condition, but this is the limit of any robust procedure.

With the simulation system that we have created, we can now make a call to the community of both developers and users to check the effect of various settings. All that is required to test a new method and compare it against existing methods is to write a wrapper function with the correct inputs and outputs. In addition, if the exact simulation settings that we use in this report are not adequate, we can easily extend this framework into an open testing system that allows additional variations on the sampling model, perhaps including additional distributions or constructed truths, such as plasmodes (39).

The current edgeR framework does not always achieve its false discovery rate target. However, even if it is forced to be more conservative, it still achieves power as good or better than existing approaches across the simulation settings that we have tested, even with the addition of observation weights. The exact source of the liberality is beyond the scope of the current investigation, but there may be room for improvement, such as borrowing ideas from small sample asymptotic approximations (40).

CONCLUSION

We developed an approach to dampen the effect of outliers on count-based differential expression analyses. Overall, the method appears to achieve the desired 'efficiency': a resistance to outliers while maintaining high power. We provided an implementation for the edgeR Bioconductor package, but the re-weighting idea could easily be adopted to other packages. In addition, we developed an extensible simulation system (at the count table level) that readily performs the simulations based on an existing dataset and provides the infrastructure for producing the standard battery of evaluations. In particular, this allows new methods or variations (e.g. alternative settings) of existing packages to be quickly explored. Instead of preparing a large number of Supplementary Figures, we provide an interactive web-based shiny 'app' to display simulation results across a broad range of simulation settings.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online, including [1,2].

ACKNOWLEDGMENTS

The authors wish to thank all members of the Robinson laboratory for helpful discussions and in particular, Olga Nikolayeva, Gosia Nowicka, Katarina Matthes and Charity Law for careful reading of an earlier version of the manuscript; we also thank members of the Baudis and von Mering groups for useful feedback. We thank Gordon Smyth and Aaron Lun for aspects of the edgeR implementation.

FUNDING

SNSF Project Grant [143883]; European Commission through the 7th Framework Collaborative Project RADIANT [305626]. Funding for open access charge: SNSF Project Grant [143883]; European Commission through the 7th Framework Collaborative Project RADIANT [305626].

Conflict of interest statement. None declared.

REFERENCES

1. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
2. Anders,S., McCarthy,D.J., Chen,Y., Okoniewski,M., Smyth,G.K., Huber,W. and Robinson,M.D. (2013) Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protocols*, **8**, 1765–1786.
3. Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.
4. Grabherr,M.G., Haas,B.J., Yassour,M., Levin,J.Z., Thompson,D.A., Amit,I., Adiconis,X., Fan,L., Raychowdhury,R., Zeng,Q. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, **29**, 644–652.
5. Frazee,A.C., Langmead,B. and Leek,J.T. (2011) ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, **12**, 449.

6. Soneson,C. and Delorenzi,M. (2013) A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, **14**, 91.
7. Rapaport,F., Khanin,R., Liang,Y., Pirun,M., Krek,A., Zumbo,P., Mason,C.E., Socci,N.D. and Betel,D. (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.*, **14**, R95.
8. McCarthy,D.J., Chen,Y. and Smyth,G.K. (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.*, **40**, 1–10.
9. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.
10. Montgomery,S.B., Sammeth,M., Gutierrez-Arcelus,M., Lach,R.P., Ingle,C., Nisbett,J., Guigo,R. and Dermitzakis,E.T. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, **464**, 773–777.
11. Law,C.W., Chen,Y., Shi,W. and Smyth,G.K. (2014) Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, **15**, R29.
12. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
13. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
14. Van De Wiel,M.A., Leday,G. G.R., Pardo,L., Rue,H.v., Van Der Vaart,A.W. and Van Wieringen,W.N. (2013) Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics*, **14**, 113–128.
15. Hardcastle,T.J. and Kelly,K.A. (2010) baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, **11**, 422.
16. Wu,H., Wang,C. and Wu,Z. (2013) A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics*, **14**, 232–243.
17. Carroll,R.J. and Pederson,S. (1993) On robustness in the logistic regression model. *J. R. Stat. Soc. Ser. B*, **84**, 693–706.
18. Cantoni,E. and Ronchetti,E. (2001) Robust inference for generalized linear models. *J. Am. Stat. Assoc.*, **96**, 1022–1030.
19. Alqallaf,F. and Agostinelli,C. (2013) Robust inference in generalized linear models, in press.
20. Agostinelli,C. (2002) Robust model selection in regression via weighted likelihood methodology. *Stat. Probab. Lett.*, **56**, 289–300.
21. Cox,D.R. and Reid,N. (1987) Parameter orthogonality and approximate conditional inference. *J. R. Stat. Soc. Ser. B Methodol.*, **49**, 1–39.
22. Robinson,M.D. and Smyth,G.K. (2008) Small-sample estimation of negative binomial dispersion with applications to SAGE data. *Biostatistics*, **9**, 321–332.
23. Robinson,M.D. and Smyth,G.K. (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**, 2881–2887.
24. Fox,J. (2002) Robust Regression. *Behav. Res. Methods*, **1**, 1–8.
25. Witten,D., Tibshirani,R., Gu,S.G., Fire,A. and Lui,W.-O. (2010) Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. *BMC Biol.*, **8**, 58.
26. Cheung,V.G., Nayak,R.R., Wang,I.X., Elwyn,S., Cousins,S.M., Morley,M. and Spielman,R.S. (2010) Polymorphic Cis- and Trans-Regulation of Human Gene Expression. *PLoS Biol.*, **8**, 14.
27. Li,J. and Tibshirani,R. (2013) Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Stat. Methods Med. Res.*, **22**, 519–39.
28. Leng,N., Dawson,J.A., Thomson,J.A., Ruotti,V., Rissman,A.I., Smits,B. M.G., Haag,J.D., Gould,M.N., Stewart,R.M. and Kendziorski,C. (2013) EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, **29**, 1035–1043.
29. Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, **11**, R25.
30. Rue,H.v., Martino,S. and Chopin,N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B: Stat. Methodol.*, **71**, 319–392.
31. Martins,T.G., Simpson,D., Lindgren,F. and Rue,H. (2012) Bayesian computing with INLA: new features, in press .
32. Efron,B., Tibshirani,R., Storey,J.D. and Tusher,V. (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, **96**, 1151–1160.
33. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B: Methodol.*, **57**, 289–300.
34. RStudio shiny: Web Application Framework for R (Version 0.7.0) (2013).
35. Hansen,K.D., Irizarry,R.A. and Wu,Z. (2012) Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*, **13**, 204–216.
36. Blekhman,R., Marioni,J.C., Zumbo,P., Stephens,M. and Gilad,Y. (2010) Sex-specific and lineage-specific alternative splicing in primates. *Genome Res.*, **20**, 180–189.
37. Herbst,H., Wege,T., Milani,S., Pellegrini,G., Orzechowski,H., Bechstein,W., Neuhaus,P. and Schuppan,D. (1997) Tissue inhibitor of metalloproteinase-1 and -2 RNA expression in rat and human liver fibrosis. *Am. Soc. Invest. Pathol.*, **150**, 51647–51659.
38. Asselah,T., Bièche,I., Laurendeau,I., Paradis,V., Vidaud,D., Degott,C., Martinot,M., Bedossa,P., Valla,D., Vidaud,M. *et al.* (2005) Liver gene expression signature of mild fibrosis in patients with chronic hepatitis C. *Gastroenterology*, **129**, 2064–2075.
39. Reeb,P. and Steibel,J. (2013) Evaluating statistical analysis models for RNA sequencing experiments. *Front. Genet.*, **4**, 178 .
40. Di,Y., Sarah,C.E., Daniel,W.S., Kimbrel,J.A. and Chang,J.H. (2013) Higher order asymptotics for negative binomial regression inferences from RNA-sequencing data.*Stat. App. Genet. Mol. Biol.*, **12** , 49 –70 .