BIOINFORMATICS ORIGINAL PAPER
Vol. 29 no. 10 2013, pages 1275–1282
doi:10.1093/bioinformatics/btt143

Gene expression
Advance Access publication April 14, 2013

Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size

Danni Yu[1,2], Wolfgang Huber[1] and Olga Vitek[2,3,*]
[1]Genome Biology Unit, European Molecular Biology Laboratory, Mayerhofstraße 1, Heidelberg 69117, Germany
[2]Department of Statistics and [3]Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA

Associate Editor: Inanc Birol

ABSTRACT

Motivation: RNA-seq experiments produce digital counts of reads that are affected by both biological and technical variation. To distinguish the systematic changes in expression between conditions from noise, the counts are frequently modeled by the Negative Binomial distribution. However, in experiments with small sample size, the per-gene estimates of the dispersion parameter are unreliable.

Method: We propose a simple and effective approach for estimating the dispersions. First, we obtain the initial estimates for each gene using the method of moments. Second, the estimates are regularized, i.e. shrunk towards a common value that minimizes the average squared difference between the initial estimates and the shrinkage estimates. The approach does not require extra modeling assumptions, is easy to compute and is compatible with the exact test of differential expression.

Results: We evaluated the proposed approach using 10 simulated and experimental datasets and compared its performance with that of currently popular packages edgeR, DESeq, baySeq, BBSeq and SAMseq. For these datasets, sSeq performed favorably for experiments with small sample size in sensitivity, specificity and computational time.

Availability: http://www.stat.purdue.edu/~ovitek/Software.html and Bioconductor.

Contact: ovitek@purdue.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 25, 2012; revised on February 24, 2013; accepted on March 25, 2013

1 INTRODUCTION

Whole transcriptome shotgun sequencing (RNA-seq) technology (Mardis, 2008; Metzker, 2009; Wang *et al.*, 2009) quantifies gene expression in biological samples in counts of transcript reads mapped to the genes. Accurate and comprehensive, it has made a major impact on genomic research (Garber *et al.*, 2011; Oshlack *et al.*, 2010; Pepke *et al.*, 2009). One common goal of RNA-seq experiments is to detect differentially expressed genes, i.e. genes for which the counts of reads change between conditions more systematically than as expected by random chance (Oshlack *et al.*, 2010). Statistical methods for detecting differentially expressed genes must reflect the experimental design and appropriately account for the stochastic variation. Moreover, many RNA-seq experiments serve as high-throughput screens of a small number of samples with the goal of subsequent experimental validation. Therefore, the analysis must handle a relatively small number of biological replicates.

A variety of statistical methods and software has recently been proposed for detecting differentially expressed genes. These include DESeq (Anders and Huber, 2010), edgeR (Robinson and Smyth, 2007; Robinson *et al.*, 2010), baySeq (Hardcastle and Kelly, 2010), SAMseq (Li and Tibshirani, 2011), BBSeq(Zhou *et al.*, 2011). We briefly overview these methods. We further propose a direct and effective approach for characterizing the variation in the counts of reads, which improves the sensitivity and specificity of detecting differentially expressed genes for experiments with small sample size. We support this approach with an open-source R-based software package sSeq.

2 BACKGROUND

2.1 The Negative Binomial distribution

The input to the statistical analysis is a set of discrete counts of reads in each experimental run. Although the counts can be modeled with the one-parameter Poisson or Geometric distributions (Auer and Doerge, 2011; Li *et al.*, 2011; Marioni *et al.*, 2008), it is often advantageous to use the two-parameter Negative Binomial distribution (Anders and Huber, 2010; Hardcastle and Kelly, 2010; Robinson and Smyth, 2007; Robinson *et al.*, 2010). This distribution is more general and flexible and can be viewed as a generalization of both Poisson and Geometric distributions (Supplementary Materials). We focus on the Negative Binomial distribution in what follows.

Denote X_{gij} as the random variable that expresses the counts of reads mapped to gene g ($g = 1, \ldots, G$) in sample (or, equivalently, *library*) j ($j = 1, \ldots, n_i$) in condition i and denote x_{gij} as the observed values. For simplicity, we consider two conditions ($i = A, B$); however, the discussion holds for pairwise comparisons of conditions in experiments with more complex designs. We are particularly interested in situations where n_A and n_B are small, e.g. 1–4. We consider the following parametrization:

$$X_{gij} \sim \mathcal{NB}(\mu_{gi}, \phi_g), \text{ where } \mu_{gi} \geq 0,\ \phi_g \geq 0, \text{ such that} \\ \mathrm{E}\{X_{gij}\} = \mu_{gi},\ \mathrm{Var}\{X_{gij}\} = \mu_{gi} + \mu_{gi}^2\phi_g\ (\overset{\text{denoted}}{=} V_{gi}) \tag{1}$$

The *dispersion* parameter ϕ_g determines the extent to which the variance V_{gi} exceeds the expected value μ_{gi} (Cameron and Trivedi, 1998; McCullagh and Nelder, 1989).

*To whom correspondence should be addressed.

2.2 Motivation for the proposed approach

The estimation of Var{X_{gij}} is the main focus of this work and is based on the following considerations.

1. A naïve approach is to estimate Var{X_{gij}} using the method of moments (i.e. the per-gene sample variance). However, it is highly variable in experiments with a small sample size (Croarkin and Tobias, 2006).
2. RNA-seq experiments simultaneously quantify the expression of many genes. The genes share aspects of biological and technical variation, and therefore a combination of the gene-specific estimates and of consensus estimates can yield better estimates of variation. Such approaches are increasingly popular with RNA-seq experiments (Anders and Huber, 2010; Robinson and Smyth, 2007).
3. The variance of the Negative Binomial distribution is a known function of the expected value μ_{gi} and of the dispersion ϕ_g. Therefore, an accurate estimation of the dispersion (e.g. by combining the gene-specific and consensus estimates, without explicitly modeling its relationship to μ_{gi}) can lead to an accurate estimation of the variance while preserving the mean–variance relationship.
4. Finally, constraints of throughput sample availability or cost may restrict the number of biological replicates. Although experiments with little or no biological replication have poor reproducibility and are undesirable, such under-replicated screens are the only practical option in some situations (Malo *et al.*, 2006; Markowetz, 2010). They can only detect large changes in expression and require an extensive downstream validation with complementary low-throughput experiments and large sample size. To detect differentially expressed genes, we assume that the majority of the genes are not differentially expressed, and that for these genes, the samples from all conditions can be viewed as biological replicates (Robinson and Oshlack, 2010; Robinson *et al.*, 2010). Under this assumption, a consensus estimate of dispersion helps to improve the accuracy of gene-specific estimates of variation.

Our main concern is in how to (i) accurately define the consensus estimate of dispersion and (ii) accurately combine the gene-specific estimates of dispersion with the consensus estimate.

2.3 Existing approaches for RNA-seq experiments

Among the existing methods, edgeR (Robinson and Smyth, 2007; Robinson *et al.*, 2010), DESeq (Anders and Huber, 2010) and baySeq (Hardcastle and Kelly, 2010) assume the Negative Binomial distribution, and SAMseq (Li and Tibshirani, 2011) and BBSeq (Zhou *et al.*, 2011) use other flexible models. The approaches have been extensively evaluated (Soneson and Delorenzi, 2013) and are broadly used. Hardcastle and Kelly (2010) found that the performance of DESeq, edgeR and baySeq is superior to that of DEGseq (Wang *et al.*, 2010), Li and Tibshirani (2011) found that SAMseq improves on PoissonSeq (Li *et al.*, 2011). We briefly overview these approaches in the historical order. Table 1 summarizes the discussion.

Probability model. *edgeR* models the count of reads with the Negative Binomial distribution. It includes normalization, which accounts for the changes in read counts owing to technical artifacts such as different sequencing depth. The normalization factor can be the total library size (i.e. the number of reads in the library). A more accurate normalization factor is the 'effective' library size m_{ij}, which multiplies the size of the library ij by a robust estimate of the log-fold change of the total count in condition i as compared with a reference run (Robinson and Oshlack, 2010). The parameter p_{gi} in row (b) of Table 1 is the probability that a single read maps to gene g for a sample in condition i. The model assumes that the dispersion parameter is gene specific but constant across conditions. For experiments without replication versions up to 2.4.6 assumed a common dispersion in all the genes. The subsequent versions discourage unreplicated experiments. Finally, generalized linear models for the Negative Binomial response are available.

DESeq also models the count of reads with the Negative Binomial distribution. It normalizes the read counts by a size factor s_{ij} (Anders and Huber, 2010).

$$\hat{s}_{ij} = \text{median}_g \frac{x_{gij}}{\left(\prod_{k=1}^{n_A} x_{gAk} \prod_{k=1}^{n_B} x_{gBk}\right)^{1/(n_A+n_B)}} \tag{2}$$

The size factor can be thought of as the 'representative' ratio of counts in the library to the geometric mean of the counts in all the libraries, and differs from the 'effective' library size in edgeR. The parameter μ_{gi} in row (c) of Table 1 is the expected normalized expression of gene g in condition i. DESeq allows specification of separate variances for genes and conditions and models the variances as functions of the expected values. This relationship can be a flexible smooth function (local polynomial) or a parabolic function $\hat{V}_{gi} = s_{ij} \cdot \hat{\mu}_{gi} + s_{ij}^2 \cdot \mu_{gi}^2 \cdot \left(a_0 + a_1/\hat{\mu}_{gi}\right)$, where $a_0, a_1 > 0$ are constants. Alternatives based on generalized linear models for the Negative Binomial response are also available.

The *baySeq* specifies the same probability model as edgeR; however, it proposes a different Empirical Bayes characterization of the between-library variation. The baySeq assumes that subsets of the libraries share the parameters of Negative Binomial distribution and derives an empirical prior distribution for the corresponding parameter sets. After integrating over the empirical priors, the dispersion in the integrated likelihood is constant across conditions and different between the genes. The default normalization parameter is the library size.

BBSeq specifies a Beta-Binomial generalized linear model. Using the logit link, it connects the expected probability of a read for gene g in condition i and sample j to the linear combinations of predictors, such as indicators of conditions and other covariates. The dispersion parameter can be independent from the mean (free model) or dependent on the mean (constrained model). *SAMseq* uses a fully non-parametric approach.

Estimation of dispersion. *edgeR* maximizes a weighted combination of the conditional log-likelihoods with per-gene dispersion and of the conditional log-likelihood with common dispersion. Conditional likelihoods generalize the restricted maximum likelihood estimation for a discrete response by conditioning on the sum of the read counts per class and improve the statistical properties of dispersion estimates. The estimation requires

Table 1. Existing and proposed approaches for differential analysis of RNA-seq experiments with two conditions

	Probability model	Estimation of dispersion	Testing	$n=1$	Time
(a) sSeq (proposed) (this manuscript)	$X_{gij} \sim \mathcal{NB}\big(s_{ij}\mu_{gi}, \phi_g/s_{ij}\big)$	$\hat{\phi}_g^{sSeq} = \delta\xi + (1-\delta)\hat{\phi}_g^{MM}$, where ξ is a common dispersion and δ is a weight	$H_0:\ \mu_{gA} = \mu_{gB}$ Exact test	Yes	min
(b) edgeR (Robinson and Smyth, 2008)	$X_{gij} \sim \mathcal{NB}\big(m_{ij}p_{gi}, \phi_g\big)$	$\hat{\phi}_g^{edgeR}$ maximize linear combination of per-gene and common-dispersion conditional likelihoods	$H_0 : p_{gA} = p_{gB}$ Exact or GLM-based test	Yes*	min
(c) DESeq (Anders and Huber, 2010)	$X_{gij} \sim \mathcal{NB}\big(s_{ij}\mu_{gi}, \phi_{gi}\big)$	$\hat{\phi}_{gi}^{DESeq} = \left(\hat{V}_{gi} - \hat{\mu}_{gi}\frac{1}{n_i}\sum_j \frac{1}{s_{ij}}\right)/\hat{\mu}_{gi}^2$ $\hat{V}_{gi}$ is estimated as function of the mean	$H_0 : \mu_{gA} = \mu_{gB}$ Exact or GLM-based test	Yes	min
(d) baySeq (Hardcastle and Kelly, 2010)	$X_{gij} \sim \mathcal{NB}\big(N_{ij}p_{gi}, \phi_g\big)$ Empirical priors on sets of parameters	$\hat{\phi}_g^{baySeq}$ maximize per-gene integrated quasi-likelihood	$H_0 : p_{gA} = p_{gB}$ Posterior probability cutoff	Yes	h
(e) BBSeq (Zhou *et al.*, 2011)	$X_{gij} \sim \mathcal{B}inom\big(p_{gi}, N_{ij}\big)$ $p_{gi} \sim \mathcal{B}eta,\ logitE\{p_{gi}\} = Z\beta,$ $\mathrm{V}(p_{gi}) = \mathrm{E}(p_{gi})\ (1 - \mathrm{E}(p_{gi}))\phi_g$	$\hat{\phi}_g^{BBSeq}$ maximize per-gene marginal likelihood; is a free parameter or a function of the mean	$H_0 : \beta = 0$ Wald test	Yes	h
(f) SAMseq (Li and Tibshirani, 2011)	Non-parametric		H_0: same distributions A and B Wilcoxon test & resampling	No	min

(a) s_{ij} is the size factor for sample j in condition i as defined in (Anders and Huber, 2010). μ_{gi} is the expected normalized expression of gene g for a sample in condition i. $\hat{\phi}_g^{MM}$ is the per-gene dispersion estimate using the method of moments in Equation (6).
(b) m_{ij} is the 'effective' library size. p_{gi} is the probability that a read in i maps to gene g. *Up to v2.4.6.
(c) ϕ_{gi} is gene- and condition-specific dispersion. $\hat{\mu}_{gi}$ and $\hat{V}_{gi}$ can be estimated by the method of moments or by the Cox-Reid corrected Maximum Likelihood.
(d) N_{ij} is the size of the library i from condition j. p_{gi} is as in (b).
(e) p_{gi} is as in (b). N_{ij} is as in (d). β is the coefficient of the linear predictor associated with an indicator Z of conditions. Column 'Time' is the run time for the experimental datasets in Section 4 on a laptop computer.

calculating pseudocounts of reads that would have been obtained with libraries of equal size and an iterative computational optimization. For experiments with few replicates, the estimates tend to be discrete values (Lloyd-Smith, 2007; Piegorsch, 1990; Toft *et al.*, 2006). For experiments with many replicates, edgeR specifies a generalized linear model. As conditional likelihoods cannot be easily extended to this case, these are further approximated by adjusted profile likelihoods (McCarthy *et al.*, 2012).

DESeq starts by estimating per-gene means and variances of the normalized counts in each gene and condition by the methods of moments. Next, it re-estimates them by fitting the postulated relationship between the expected values and the variances. The estimates of dispersion can be back-calculated from the estimates of variance as shown in row (c) of Table 1. For experiments without replication, DESeq assumes that the majority of the genes are not differentially expressed, and combines the samples across conditions to estimate the variance. The same strategy is used with the generalized linear models.

The *baySeq* relies on an iterative estimation of the relative gene expression and of the dispersion. Given an initial partition of the libraries into subsets and an initial estimate of the relative gene expression, it estimates the dispersion using the quasi-likelihood approach. Given the estimates of dispersion, it re-estimates the relative gene expression by maximizing the integrated likelihood. This is repeated for different partitions of the libraries.

BBSeq estimates the dispersion using maximum likelihood for the free model. For the constrained model, it uses the estimates from the free model for all the genes, fits the postulated relationship to the mean and re-estimates the dispersions. *SAMseq* sidesteps the need to estimate the dispersion by using a fully non-parametric approach.

Testing. For the Negative Binomial model, *edgeR* tests the null hypothesis $H_0:\ p_{gA} = p_{gB}$, and *DESeq* $H_0:\ \mu_{gA} = \mu_{gB}$ separately for each gene. Both edgeR and DESeq use the exact test, which is free from asymptotic arguments and is therefore preferred. The test statistic for a gene is the total (normalized) count of reads in all the replicates of a condition. The *P*-value is the probability of the normalized read counts per group such that under H_0 their probability is same or lower than the probability of the observed counts, conditional on the total counts equal to the observed. With the generalized linear models, edgeR and DESeq use the asymptotic likelihood ratio or Wald tests.

baySeq ranks the genes by their posterior probabilities of differential expression. *BBSeq* tests the coefficient of the linear predictor (i.e. condition) in the generalized linear model with the asymptotic Wald test. *SAMseq* uses a resampling strategy to estimate the distribution of the test statistic and the *P*-values.

3 METHODS

The proposed approach combines aspects of the existing approaches, but is simpler, requires fewer assumptions and streamlines the computation. It is summarized in Table 1a. More details regarding the method and its implementation in *sSeq* are in Supplementary Materials.

Probability model. The model for the counts X_{gij} of gene $g = 1, \ldots, G$, replicate $j = 1, \ldots, n_i$ and condition $i = A, B$ is

$$X_{gij} \sim \mathcal{NB}(\mu_{gi}\ s_{ij},\ \phi_g/s_{ij}),\ \text{such that} \tag{3}$$

$$\mathrm{E}\{X_{gij}\} = \mu_{gi}\; s_{ij}, \text{ and } \mathrm{Var}\{X_{gij}\} = \mu_{gi}\; s_{ij} + \mu_{gi}^2 \phi_g\; s_{ij} \qquad (4)$$

We follow edgeR, DESeq and baySeq by specifying a Negative Binomial distribution. As in experiments with a small sample size, it may be difficult to distinguish the true dependency of dispersions on expected values from artifacts of random variation, the model specifies free gene-specific dispersion parameters ϕ_g. As the initial versions of edgeR, we specify a common dispersion across conditions, i.e. $\phi_{gA} = \phi_{gB} \overset{denoted}{=} \phi_g$. As a consequence, the counts of differentially expressed genes have different variances in each condition.

We follow DESeq in normalizing the counts by the size factor s_{ij}. However, in the proposed normalization, the size factor affects not only the expected value but also the dispersion. Equation (4) shows that under this assumption the size factor linearly scales both $\mathrm{E}\{X_{gij}\}$ and $\mathrm{Var}\{X_{gij}\}$. Such linear scaling is consistent with the technical variation in RNA-seq experiments, which can be characterized by the Poisson distribution (Marioni *et al.*, 2008). As typical size factors are close to 1, the proposed model has little practical difference from the model in DESeq. However, as shown in Supplementary Section 2, it allows us to directly conduct the exact test and contributes to the accuracy of the results.

Estimation of dispersion. Similarly to DESeq, we start by estimating the dispersion parameters by the methods of moments. A conservative estimate of the per-gene variance in experiments with a small sample size is obtained by pooling the samples across conditions, i.e.

$$\hat{V}_g = \frac{\sum_i \sum_j (x_{gij}/\hat{s}_{ij} - \hat{\mu}_g)^2}{\sum_i n_i \; - \; 1}, \text{ with } \hat{\mu}_g = \frac{\sum_i \sum_j x_{gij}/\hat{s}_{ij}}{\sum_i n_i} \qquad (5)$$

and $g = 1, \ldots, G$. The estimate of dispersion $\hat{\phi}_g^{MM}$ is then calculated from Equation (1), and negative values are truncated at zero

$$\hat{\phi}_g^{MM} = max\left(0, \frac{\sum_i n_i V_g - \mu_g \sum_i \sum_j \frac{1}{s_{ij}}}{\mu_g^2 \sum_i \sum_j \frac{1}{s_{ij}}}\right) \qquad (6)$$

Unfortunately, in experiments with small sample size, $\hat{\phi}_g^{MM}$ are unsatisfactory owing to high variance (Bowman, 1984; Clark and Perry, 1989; Willson et al., 1984). Next, we improve the statistical properties of these estimates by introducing shrinkage.

Stein (1956) showed that when we estimate the expected values of three or more independent Normal random variables with known constant variance, shrinking the per-dimension estimates toward a target value ξ produces biased estimates, but reduces the mean squared error (MSE) for all choices of ξ. The shrinkage estimator by James and Stein (1961) (Lehmann and Casella, 1998; Richards, 1999) implements this strategy. More recently, Hansen extended the approach of James and Stein with a generalized shrinkage estimator, (Hansen, 2008). Hansen's shrinkage can be used with any per-dimension estimator with an arbitrary sampling distribution (not necessarily Normal), for which the Central Limit Theorem holds. Specifically, it requires that the true parameter lies in a neighborhood of the restricted parameter space, and that the estimator is asymptotically Normal with a consistent variance. Estimators by the method of moments satisfy these criteria. Applied to the estimation of ϕ_g, and assuming that the per-gene estimates are independent, the generalized shrinkage estimator is

$$\hat{\phi}_g^{sSeq} = (1-\delta)\hat{\phi}_g^{MM} + \delta \cdot \xi = \xi + (1-\delta)(\hat{\phi}_g^{MM} - \xi) \qquad (7)$$

$\hat{\phi}_g^{sSeq}$ is a linear combination of the pre-defined target ξ and of the per-gene methods of moment estimates. The weight δ is defined as

$$\delta = \frac{\sum_g \left(\hat{\phi}_g^{MM} - \bar{\hat{\phi}}^{MM}\right)^2 / (G-1)}{\sum_g (\hat{\phi}_g^{MM} - \xi)^2 / (G-2)}, \text{and } \bar{\hat{\phi}}^{MM} = \frac{1}{G}\sum_g \hat{\phi}_g^{MM} \qquad (8)$$

As $\sum_g \left(\hat{\phi}_g^{MM} - \bar{\hat{\phi}}^{MM}\right)^2 \leq \sum_g (\hat{\phi}_g^{MM} - \xi)^2$, the weight $\delta \in (0, 1)$. Larger values of δ shrink the estimates closer to the pre-defined target ξ.

We use the Hansen's generalized shrinkage estimator $\hat{\phi}_g^{sSeq}$ in conjunction with the Negative Binomial distribution to test genes for differential expression. Although the assumption of $\hat{\phi}_g^{MM}$ being independent variables is simplistic, it is a suitable approximation for experiments with a small sample size. A similar assumption is made, e.g. by DESeq when modeling the variance as function of the mean. Although the asymptotic argument cannot be justified in this context, we show empirically in Section 5 that $\hat{\phi}_g^{sSeq}$ performs well in practice.

Hansen showed that the estimator in Equations (7) and (8) reduces the asymptotic MSE for all choices of targets ξ. However, a good practice is to select a value for ξ that maximizes this reduction. To this end, we approximate the $\mathrm{MSE} = \mathrm{E}\{\sum_{g=1}^{G} (\hat{\phi}_g^{sSeq} - \phi_g)^2\}$ using the average squared difference (ASD) between $\hat{\phi}_g^{sSeq}$ and $\hat{\phi}_g^{MM}$

$$\mathrm{ASD} = \frac{1}{G}\sum_{g=1}^{G} (\hat{\phi}_g^{sSeq} - \hat{\phi}_g^{MM})^2 \qquad (9)$$

Equation (9) substitutes ϕ_g with $\hat{\phi}_g^{MM}$ and divides MSE by the constant G for numeric stability. It is shown that

$$\mathrm{ASD}(\xi) = \frac{\text{constant}}{\sum_{g=1}^{G} (\hat{\phi}_g^{MM} - \xi)^2} \qquad (10)$$

Figure 1a visualizes the functional form of $ASD(\xi)$ for a simulation in Section 4 and shows that the tail of the curve flattens for large ξ. Therefore, we can also minimize the bias by minimizing ξ while enforcing the constraint that $ASD(\xi)$ is comparably small. In practice, *sSeq* estimates $\hat{\xi}$ by calculating the slope of $ASD(\xi)$ and setting

$$\hat{\xi} = \mathrm{argmin}_{\xi}\{-\epsilon < \mathrm{slope}(\mathrm{ASD}(\xi)) < 0\} \qquad (11)$$

for a small constant ϵ such as $\epsilon = 0.05$. The selected value is shown by the vertical line in Figure 1a.

Figure 1b illustrates the fact that the proposed shrinkage estimator is a linear transformation of $\hat{\phi}_g^{MM}$. The slope of the transformation is $(1-\delta) \in (0, 1)$, and the fixed point is the shrinkage target ξ. The shrinkage increases the per-gene estimates of dispersion that are smaller than ξ and decreases the values that are larger than ξ. From our experience with multiple datasets, $\hat{\xi}$ is often around the 95.5th quantile of $\hat{\phi}_g^{MM}$. In other words, it biases the majority of the estimates towards conservative values.

The proposed estimate of dispersion has analogies in methods developed for other high-throughput technologies. For example, it is similar in spirit to the moderated variance estimator in the package Limma (Smyth, 2004, 2005), which is also a linear combination of per-gene and consensus estimates.

Testing. We follow edgeR and DESeq by testing $H_0 : \mu_{gA} = \mu_{gB}$ per gene with the exact test. The test statistic is $X_{gi\cdot}$, i.e. the sum of the read counts in each condition. Under H_0, the *P*-values are calculated with respect to the reference distribution $X_{gi\cdot} \overset{H_0}{\sim} \mathcal{NB}\left(\sum_j s_{ij} \cdot \mu_g,\ \phi_g / \sum_j s_{ij}\right)$ and are adjusted to control the false discovery rate (FDR) (Benjamini and Hochberg, 1995). See Supplementary Materials for more details.

Extensions to experiments with complex designs. The proposed approach can be extended to pairwise comparisons of conditions in experiments with more complex designs without recurring to a generalized linear model and without additional assumptions. See Supplementary Materials for details.

4 DATASETS

We evaluated the proposed approach using 10 simulated and experimental datasets. The first five datasets had an external

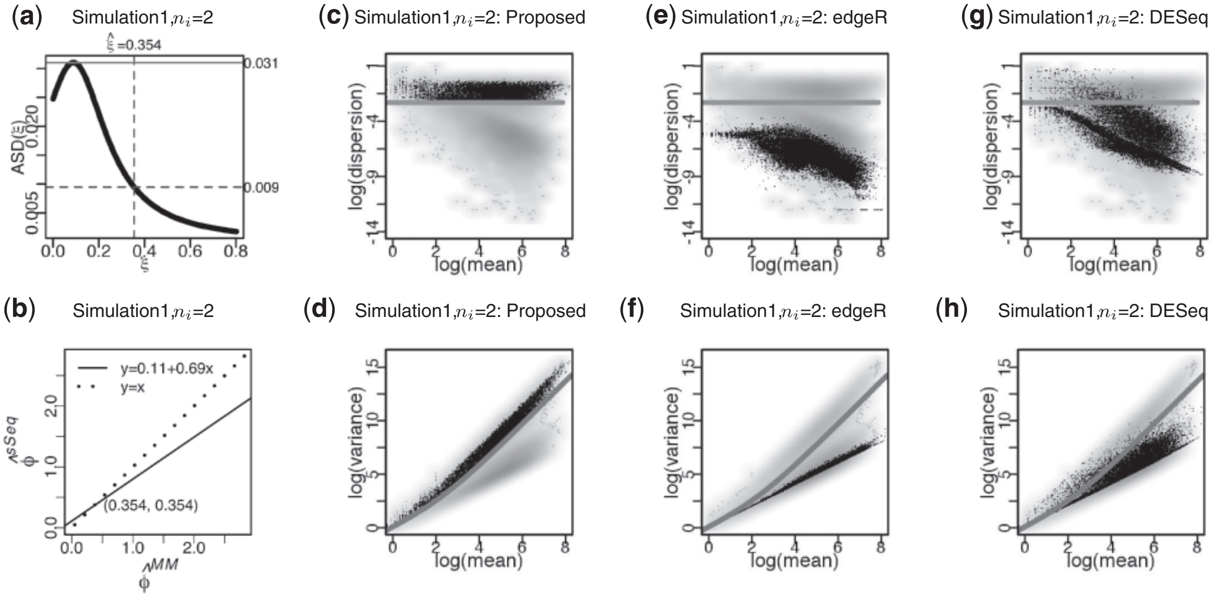

Fig. 1. Dispersion and variance estimation in Simulation1. Similar plots for other datasets are shown in Supplementary Materials. (**a**) ASD versus shrinkage target ξ. ASD is maximized at $\xi = \bar{\hat{\phi}}^{MM}$ (solid horizontal line). The dashed lines are the selected target $\hat{\xi}$ and its ASD. (**b**) The proposed shrinkage estimator is a linear transformation of $\hat{\phi}_g^{MM}$, with the slope $(1-\delta) = 0.69$ and the fixed point $\hat{\xi} = 0.354$. All $\hat{\phi}_g^{MM} \leq 0$ are transformed to $\delta\xi = 0.11$. (**c**, **e** and **g**) Dispersion estimates by sSeq, edgeR and DESeq versus the per-gene mean read counts across conditions. Gray smooth scatter are $\hat{\phi}_g^{MM}$ (same on all the plots). Black dots are $\hat{\phi}_g$ estimated by each method. Gray lines indicate the true dispersion parameters. (**d**, **f** and **h**) Same as above, but for the variances of the read counts

'gold standard' of differential expression, and the last two had experimental designs more complex than a two-group comparison. See Supplementary Materials for more details.

Simulation1, Simulation2 and Simulation3 each generated $G = 20\,000$ genes in conditions A and B, $n_A = n_B = 2$. Thirty percent of the genes were simulated as differentially expressed. Size factors were sampled from the Uniform distribution $s_{ij} \sim Uniform(0.5, 1.7)$.

Simulation1 assumed a constant dispersion parameter across all genes and is favorable to sSeq. Simulation2 assumed that ϕ_g is a non-linear function of μ_g, i.e. $\phi_g = 1/(100 + \mu_g)$ and as such is favorable to DESeq. Simulation3 is most realistic. From the dataset by Bottomly *et al.* (2011) (Frazee *et al.*, 2011), the largest experimental dataset in this manuscript, we selected a subset of non-differentially expressed genes (as determined by a consensus of sSeq, edgeR and DESeq) and sampled pairs $(\hat{\mu}_{gA}^{MM}, \hat{\phi}_g^{MM})$ from this subset as the true parameters for simulating read counts.

MAQC (Shi *et al.*, 2006) is the dataset from the MicroArray Quality Control (MAQC) consortium, comparing three libraries from Ambion human brain reference RNA against two libraries from Stratagene human universal reference RNA. The libraries were sequenced with the Illumina platform, resulting in 19 580 genes. A subset of the genes from four of the libraries was assayed by real-time reverse-transcription PCR (Shi *et al.*, 2006; Zhining *et al.*, 2010). We used the 323 differential genes and 85 non-differentially expressed genes determined by real-time reverse-transcription PCR as the 'gold standard'. Although the dataset only has technical replicates, it has been used extensively as the benchmark in the past (Arikawa *et al.*, 2008; Bullard *et al.*, 2010; Patterson *et al.*, 2006).

Griffith *et al.* (2010) compared fluorouracil (5-FU)-resistant human colorectal cancer cell lines MIP101 against their non-resistant counterpart MIP/5-FU24. One library from each condition was quantified with the paired-end Illumina platform, resulting in 27 145 genes. In all, 197 of these genes from the same samples were assayed by quantitative PCR. We used 12 truly differential genes and 19 truly non-differentially expressed genes as determined by quantitative PCR as the 'gold standard' for method comparison.

Brooks *et al.* (2011) compared untreated cells of *Drosophila melanogaster* against cells cultured in presence of Pasilla, the homologue of the mammalian Nova-1 and Nova-2 protein. Two biological samples per condition were sequenced with the paired-end Illumina platform, resulting in 14 470 genes.

Sultan *et al.* (2008) (Frazee *et al.*, 2011) compared two biological replicates of human cell lines Ramos B and HEK293T with the Illumina platform, yielding 6 573 643 uniquely aligned reads.

Bottomly *et al.* (2011) (Frazee *et al.*, 2011) compared brain tissues of two inbred mouse strains, C57BL/6J (B6) and DBA/2J (D2), using the Illumina platform. The analysis of 10 and 11 biological samples per condition resulted in 13 932 genes.

Hammer *et al.* (2010) (Frazee *et al.*, 2011) compared gene expression in rat strains Sprague Dawley and L5 SNL Sprague Dawley 2, at two times (2 weeks and 2 months) in a factorial design. Two distinct biological libraries per condition and per time slot were quantified using the Illumina platform, resulting in 18 635 genes.

Tuch *et al.* (2010) compared the expression of genes in normal human tissues and in tissues with oral squamous cell carcinoma. The experiment compared pairs of normal and tumor samples from three patients. The six libraries were sequenced using the SOLiD platform, resulting in 10 453 genes.

5 RESULTS

In addition to sSeq, the following versions of the existing packages were used: edgeR v3.0.8 (January 2013), DESeq v1.10.1 (October 2012), baySeq v1.12.0 (October 2012), BBSeq v1.0 (March 2011), SAMSeq as part of the R package samr v2.0 (June 2011).

For sSeq, all the datasets were analyzed with the exact test, and analyses of the Hammer and the Tuch datasets accounted for their experimental designs. For edgeR and DESeq, the datasets with two-group comparisons were analyzed with the exact test, and Hammer and Tuch datasets were analyzed with the glm-based approaches. For edgeR, the estimateCommonDisp function in an older version of edgeR package (v2.4.6) was used to analyze unreplicated datasets. For DESeq, the option fitType='local' was used to estimate the per-group variance. The default parameters were used otherwise. See Supplementary Materials for details and representative R scripts.

5.1 sSeq accurately estimates the variation

As the proposed approach shares most similarities with edgeR and DESeq, we compared their estimates of dispersions and variances in more details. Figure 1c, e and g use the simplest case of Simulation1 to illustrate the estimates by the method of moments and by the three approaches. As expected, $\hat{\phi}_g^{MM}$ have a high variance, which increases with the mean. Also as expected, estimates by $\hat{\phi}_g^{sSeq}$ are biased towards larger values but have smaller deviations from the true values as compared to $\hat{\phi}_g^{MM}$. Estimates by the other two methods fit the pattern of $\hat{\phi}_g^{MM}$.

Figure 1d, f and h show that despite the differences in dispersion estimation, the estimates of variance by the three methods are less different. This is due to the fact that the values of the dispersions are small as compared with the means, and that the variances in Equation (1) are highly influenced by the expected values. As a result, the bias in the estimation of the dispersion has a low impact on the overall estimation of variation. Supplementary Materials provide plots for the other datasets.

The first two columns of Table 2 show that the bias also has little impact on the performance of detecting differentially expressed genes, as the performance of sSeq, edgeR and DESeq are relatively similar. sSeq has a slightly higher area under the ROC curves.

Figure 1d, f and h also provide an insight into why shrinking the method of moments estimates of dispersion is more beneficial than shrinking the method of moments estimates of variance. On the log scale the relationship between the mean and the variance in the Negative Binomial distribution is roughly linear for large mean counts. Mathematically, from Equation (1)

$$\begin{aligned} \log(V_{gi}) &= \log(\mu_{gi} + \mu_{gi}^2\phi_g) = \log(\mu_{gi}) + \log(\mu_{gi}\phi_g + 1) \\ \log(V_{gi}) &\overset{\text{large } \mu_{gi}}{\approx} 2 \cdot \log(\mu_{gi}) + \log(\phi_g) \end{aligned} \tag{12}$$

A shrinkage of the variance estimates would multiply them by $(1 - \delta) \leq 1$ and would distort the slope of the mean–variance relationship in Equation (1) away from Equation (2). The shrinkage of the dispersion parameter, on the other hand, preserves this nominal mean–variance relationship. Our results (shown in Supplementary Materials) confirmed that shrinking the variance leads to inferior performance.

To further investigate the usefulness of multiple shrinkage targets, we partitioned the genes into 10 groups according to the ranges of $\hat{\mu}_g^{MM}$ and applied the shrinkage separately to each group. Our results (not shown here) indicated that there is no advantage in specifying multiple shrinkage targets.

5.2 sSeq accurately detects differential expression

Five datasets with an external 'gold standard' were used to evaluate the sensitivity and the specificity of detecting differentially expressed genes. For each method, the genes were ranked by FDR-adjusted *P*-value of posterior probability and termed 'significant' for varying cutoffs. The sensitivity and the specificity of differential expression was compared with the 'gold standard' and summarized with ROC curves. Table 2 shows that the proposed approach consistently had a similar or a higher accuracy as compared with the existing methods.

Five datasets without an external 'gold standard' were used to evaluate the sensitivity and the specificity less formally, as discussed in (Anders and Huber, 2010). First, comparisons of two conditions ('*AvsB*') had some truly differentially expressed genes. Therefore, methods with higher sensitivity should have higher areas under the empirical cumulative distribution functions (ECDF) of the *P*-values defined as $\hat{F}(\mathrm{p}) = \frac{1}{G}\sum_{g=1}^{G} I_{\{p-\mathrm{value}_g \leq p\}}$.

Table 2. Areas under the ROC curves of detecting differentially expressed genes for the datasets with an external 'gold standard' while varying the FDR-adjusted *P*-value or posterior probability cutoff

Methods		Simulation1		Simulation2		Simulation3		MAQC Project		Griffith *et al.*
		$n_i = 1$	$n_i = 2$	$n_i = 1$	$n_i = 2$	$n_i = 1$	$n_i = 2$	$n_i = 1$	$n_A = 3, n_B = 2$	$n_i = 1$
Proposed	sSeq	0.947	0.962	0.951	0.967	0.856	0.888	0.585	0.911	0.689
Existing	edgeR	0.918	0.948	0.938	0.951	0.840	0.833	0.558	0.850	0.557
	DESeq	0.932	0.940	0.937	0.949	0.842	0.816	0.577	0.867	0.596
	baySeq	0.568	0.711	0.548	0.714	0.558	0.628	0.551	0.852	0.702
	BBSeq	0.675	0.672	0.669	0.674	0.578	0.619	0.560	0.617	0.544
	SAMseq		0.964		0.968		0.882		0.563	

Sub-columns are subsets of the data with one randomly selected replicate per condition and the full datasets. Values closer to 1 indicate higher sensitivity and specificity.

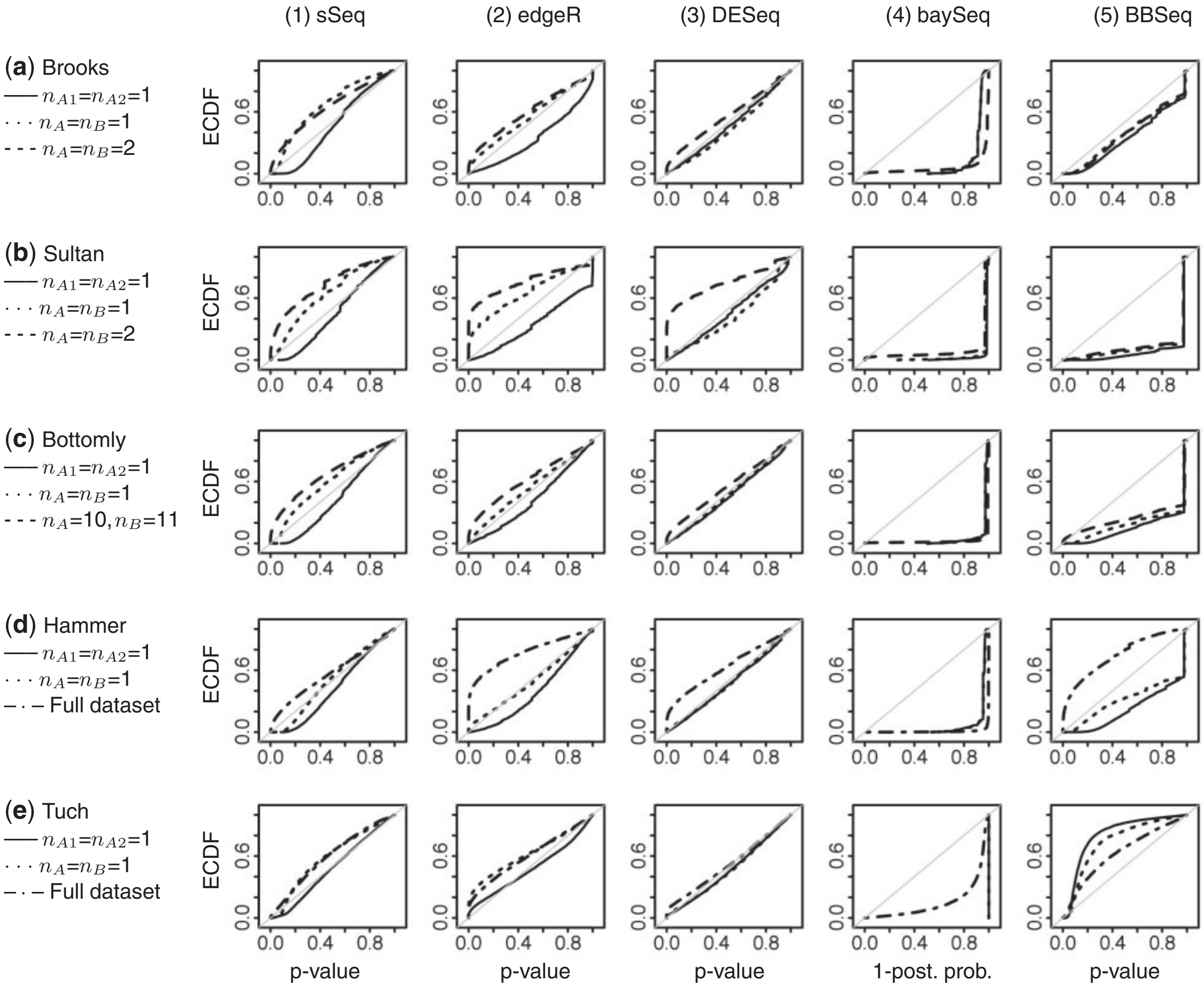

Fig. 2. The ECDF curves of detecting differential expression for the datasets with no external 'gold standard'. Y-axis: ECDF, function of the gene rank. x-axis: P-value or 1 minus posterior probability. Solid line: two randomly selected replicates from a same condition (*AvsA*). Dotted line: one randomly selected replicate from each condition (unreplicated *AvsB*). Dashed line: *AvsB* on the full dataset for two-group designs. Dashed-dotted line: *AvsB* on the full dataset for more complex designs. Gray line: 45 degree. SAMseq is not applicable to unreplicated experiments and is excluded. The desired patterns are high areas under the *AvsB* curves, and *AvsA* curves that are at or below the 45 degree line

Second, comparisons of replicates of a same condition ('*AvsA*') had no differentially expressed genes. Therefore, methods with higher specificity should have ECDF curves at or below the 45 degree line. For baySeq, we expect similar patterns of the ECDF curves based on the posterior probability cutoff. Figure 2 summarizes the curves for the five datasets. It shows that sSeq produced most consistently the expected pattern and had a similar or a higher accuracy as compared with the existing methods.

The effect of sample size and of size factors on the accuracy was investigated using the three simulated datasets. Supplementary Section 4.4 and 4.5 indicate that sSeq is particularly advantageous for experiments when $n \leq 4$.

6 DISCUSSION

In this manuscript, we advocated a model that specifies free per-gene dispersion parameters in the Negative Binomial model for counts of RNA-seq reads. We also advocated a biased estimation of these parameters, which can reduce the variance of the estimates and minimize the overall MSE. Biased estimation is different from specifying a probability model (such as in DESeq) that assumes a true systematic relationship of the true variance and the true mean. It is particularly useful for experiments with a small sample size, where the systematic relationship may be difficult to evaluate. The shrinkage estimates are easy to compute, avoid iterative estimation, minimize the potential for overfitting and do not require extra computation time. They are compatible with the exact test of differential expression. For the datasets in this manuscript, sSeq consistently had a similar or a higher sensitivity and specificity of detecting differential expression than the existing methods. The approach can be generalized to express the dependence of the dispersions on the expected value or on other covariates such as guanine-cytosine (GC) content or Gene Ontology annotations.

sSeq can produce meaningful results in under-replicated RNA-seq screens. However, we stress that RNA-seq screens do not eliminate the biological variation in gene expression Equation (12). As evidenced by Table 2 and Figure 2, the under-replicated screens have lower reproducibility as compared with the replicated studies. Multiple biological replicates are necessary to adequately assess the full extent of the variation in the biological system. Therefore, the under-replicated screens can only be conducted when followed by a rigorous experimental validation with complementary technologies and adequate sample size.

ACKNOWLEDGEMENTS

The authors thank S. Anders, M. E. Bock and the Huber group for helpful discussions.

Funding: NSF CAREER award DBI-1054826 to O.V. (in part).

Conflict of Interest: none declared.

REFERENCES

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol*., **11**, R106.

Arikawa,E. *et al*. (2008) Cross-platform comparison of SYBR Green real-time PCR with TaqMan PCR, microarrays and other gene expression measurement technologies evaluated in the MicroArray Quality Control (MAQC) study. *BMC Genomics*, **9**, 328.

Auer,P. and Doerge,R. (2011) A two-stage Poisson model for testing RNA-seq data. *Stat. Appl. Genet. Mol. Biol*., **10**, 1–26.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R Stat. Soc. B*, **57**, 289–300.

Bottomly,D. *et al*. (2011) Evaluating Gene Expression in C57BL/6J and DBA/2J mouse striatum using RNA-seq and microarrays. *PloS One*, **6**, e17820.

Bowman,K. (1984) Extended moment series and the parameters of the negative binomial distribution. *Biometrics*, **40**, 249–252.

Brooks,A. *et al*. (2011) Conservation of an RNA regulatory map between *Drosophila* and mammals. *Genome Res*., **21**, 193–202.

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

Cameron,A. and Trivedi,P. (1998) *Regression Analysis of Count Data*, Econometric Society Monograph (No.30). Cambridge University Press, Cambridge, UK.

Clark,S. and Perry,J. (1989) Estimation of the negative binomial parameter κ by maximum quasi-likelihood. *Biometrics*, **45**, 309–316.

Croarkin,C. and Tobias,P. (2006) *NIST/SEMATECH e-Handbook of Statistical Methods*. National Institute of Standards and Technology. http://www.itl.nist.gov/div898/handbook/ (15 April 2013, date last accessed).

Frazee,A. *et al*. (2011) ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, **12**, 449.

Garber,M. *et al*. (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods*, **8**, 469–477.

Griffith,M. *et al*. (2010) Alternative expression analysis by RNA sequencing. *Nature Methods*, **7**, 843–847.

Hammer,P. *et al*. (2010) mRNA-seq with agnostic splice site discovery for nervous system transcriptomics tested in chronic pain. *Genome Res*., **20**, 847–860.

Hansen,B. (2008) *Generalized Shrinkage Estimators*. University of Wisconsin.

Hardcastle,T. and Kelly,K. (2010) BaySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, **11**, 422.

James,W. and Stein,C. (1961) Estimation with quadratic loss. In: *Proceedings of the fourth Berkeley Symposium on Mathematical Statistics and Probability Held at the Statistical Laboratory, University of California, June 20-July 30, 1960*. University of California Press, p. 361.

Lehmann,E. and Casella,G. (1998) *Theory of Point Estimation*. Springer, New York.

Li,J. *et al*. (2011) Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, **13**, 523–538.

Li,J. and Tibshirani,R. (2011) Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-seq data. *Stat. Methods Med. Res*., [Epub ahead of print, doi: 10.1177/0962280211428386, November 28, 2011].

Lloyd-Smith,J. (2007) Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. *PLoS One*, **2**, e180.

Malo,N. *et al*. (2006) Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol*., **24**, 167–175.

Mardis,E. (2008) Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet*., **9**, 387–402.

Marioni,J. *et al*. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*., **18**, 1509–1517.

Markowetz,F. (2010) How to understand the cell by breaking it: network analysis of gene perturbation screens. *PLoS Comput. Biol*., **6**, e1000655.

McCarthy,D. *et al*. (2012) Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. *Nucleic Acids Res*., **40**, 4288–4297.

McCullagh,P. and Nelder,J. (1989) *Generalized Linear Models*. Monographs on Statistics and Applied Probability (No. 37), Chapman & Hall/CRC, New York.

Metzker,M. (2009) Sequencing technologies: The next generation. *Nat. Rev. Genetics*, **11**, 31–46.

Oshlack,A. *et al*. (2010) From RNA-seq reads to differential expression results. *Genome Biol*., **11**, 220.

Patterson,T. *et al*. (2006) Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. *Nat. Biotechnol*., **24**, 1140–1150.

Pepke,S. *et al*. (2009) Computation for ChIP-seq and RNA-seq studies. *Nat. Methods*, **6**, S22–S32.

Piegorsch,W. (1990) Maximum likelihood estimation for the negative binomial dispersion parameter. *Biometrics*, **46**, 863–867.

Richards,J. (1999) An Introduction to James-Stein estimation, M.I.T. EECS Area Exam Report, 1999. http://www.yaroslavvb.com/papers/richards-introduction.ps (15 April 2013, date last accessed).

Robinson,M. *et al*. (2010) EdgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Robinson,M. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*., **11**, R25.

Robinson,M. and Smyth,G. (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**, 2881–2887.

Shi,L. *et al*. (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol*., **24**, 1151–1161.

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

Smyth,G. (2005) Limma: Linear models for microarray data. In: *Bioinformatics Computational Biology Solutions Using R and Bioconductor*. Springer, New York, pp. 397–420.

Soneson,C. and Delorenzi,M. (2013) A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, **14**, 91.

Stein,C. (1956) Inadmissibility of the usual estimator for the mean of a multivariate Normal distribution. In: *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*. Vol. 1, University of California Press, Berkeley, pp. 197–206.

Sultan,M. *et al*. (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956–960.

Toft,N. *et al*. (2006) The Gamma-Poisson model as a statistical method to determine if micro-organisms are randomly distributed in a food matrix. *Food Microbiol*., **23**, 90–94.

Tuch,B. *et al*. (2010) Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations. *PloS One*, **5**, e9317.

Wang,L. *et al*. (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, **26**, 136–138.

Wang,Z. *et al*. (2009) RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet*., **10**, 57–63.

Willson,L. *et al*. (1984) Multistage estimation compared with fixed-sample-size estimation of the negative binomial parameter *k*. *Biometrics*, **40**, 109–117.

Zhining,W. *et al*. (2010) Evaluation of gene expression data generated from expired Affymetrix GeneChip microarrays using MAQC reference RNA samples. *BMC Bioinformatics*, **11** (**Suppl. 6**), S10.

Zhou,Y. *et al*. (2011) A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics*, **27**, 2672–2678.