Published online 23 May 2017

Nucleic Acids Research, 2017, Vol. 45, No. 13 ***e127***
doi: 10.1093/nar/gkx456

Gene expression variability and the analysis of large-scale RNA-seq studies with the MDSeq

Di Ran[1] and Z. John Daye[2,*]

[1]Mel and Enid Zuckerman College of Public Health, The University of Arizona, Tucson, AZ 85724, USA and [2]Independent Researcher, Raleigh, NC 27612, USA

Received December 12, 2016; Revised April 10, 2017; Editorial Decision May 07, 2017; Accepted May 19, 2017

ABSTRACT

Rapidly decreasing cost of next-generation sequencing has led to the recent availability of large-scale RNA-seq data, that empowers the analysis of gene expression variability, in addition to gene expression means. In this paper, we present the *MDSeq*, based on the coefficient of dispersion, to provide robust and computationally efficient analysis of both gene expression means and variability on RNA-seq counts. The *MDSeq* utilizes a novel reparametrization of the negative binomial to provide flexible generalized linear models (GLMs) on both the mean and dispersion. We address challenges of analyzing large-scale RNA-seq data via several new developments to provide a comprehensive toolset that models technical excess zeros, identifies outliers efficiently, and evaluates differential expressions at biologically interesting levels. We evaluated performances of the *MDSeq* using simulated data when the ground truths are known. Results suggest that the *MDSeq* often outperforms current methods for the analysis of gene expression mean and variability. Moreover, the *MDSeq* is applied in two real RNA-seq studies, in which we identified functionally relevant genes and gene pathways. Specifically, the analysis of gene expression variability with the *MDSeq* on the GTEx human brain tissue data has identified pathways associated with common neurodegenerative disorders when gene expression means were conserved.

INTRODUCTION

The analysis of gene expressions via hybridization-based microarray technologies has enjoyed much success in the last two decades. Genes regulating a myriad of human diseases have been identified in microarray studies, including those for brain tumors (1), breast cancer (2–4), skin tumors (5), and a number of neurological disorders (6–8). Nonetheless, microarray experiments can be limited by the presence of cross-hybridization artifacts (9), intensity variability at low expression levels (10), signal saturation of highly expressed genes (11), and partial assessment of genes restricted to annotated transcripts (12). Utilizing next-generation sequencing (NGS) technologies, RNA sequencing (RNA-seq) has largely improved upon the limitations of microarray technologies and rapidly emerged as the preferred tool for transcriptome analysis (12). Ever decreasing costs of high-throughput sequencing have led to the recent availability of large-scale RNA-seq studies by providing datasets with moderate to large sample sizes. These include the Encyclopedia of DNA Elements (ENCODE) (13), the Genotype-Tissue Expression (GTEx) Project (14), the Genetic European Variation in Health and Disease (GEUVADIS) dataset (15), etc. The analysis of large-scale RNA-seq count data presents both new challenges and opportunities.

Gene expression variability can provide important insights on how genes function in biological processes beyond those acquired from standard analysis of gene expression means. For instance, variability analysis of gene expression levels has identified transcriptional regulators in the development of early human embryos (16). Gene expression variability at aberrant levels can suggest disruptions or dysregulations of biological processes (17,18). Recent studies have associated increased levels of expression variability with Schizophrenia (19) and aggressive chronic lymphocytic leukemia (20). A number of methods for the analysis of gene expression variability has been proposed for applications in microarray studies (21–27). However, as early RNA-seq studies often have only a limited number of samples that cannot be reliably applied to assess statistical variability, the analysis of gene expression variability has, so far, been largely ignored in RNA-seq studies that focused on the analysis of gene expression means (28–35). The availability of large-scale RNA-seq studies presents an unprecedented opportunity to evaluate gene expression variability without the encumbrance of the many limitations of microarray technologies.

The analysis of large-scale RNA-seq count data brings about several new challenges that are vital towards the analysis and interpretation of both gene expression mean and

*To whom correspondence should be addressed. Tel: +1 765 418 5894; Email: zhongyindaye@gmail.com

variability. Excess zeros, beyond those realized from biological variations, are often present in a significant proportion of genes in large-scale RNA-seq studies. This is often attributed to technical variations from read failures in low-count samples (36) and has been suggested to contribute to elevated levels of overdispersion in RNA-seq data (37). Furthermore, by evaluating a relatively large number of observations, outlying samples are more likely to be encountered in large-scale studies. As gene expression analysis often seeks to interrogate biologically consistent effects across treatments, it is important to identify and remove outlying observations to achieve robust biological interpretation and reproducibility of results (32,38). Most particularly, procedures for differential gene expression analysis often focus on evaluating the null hypothesis that log fold-changes (FCs) between cases and controls are exactly zero, such that there is no differentiation between cases and controls. However, given moderate to large sample sizes, the null hypothesis is often easily rejected, and this can educe a deluge of statistically significant genes, most of which are differentiated at only very modest levels that are biologically uninteresting (39,40). Effective methods are much needed to overcome these challenges in order to provide robust and biologically meaningful analysis of both gene expression mean and variability in large-scale RNA-seq studies.

In microarray studies, log-transformed intensity levels are often assumed to follow a normal or Gaussian distribution (41), and the analysis of gene expression variability usually involves evaluating the normal variances directly (24–26) or conducting heterogeneity tests under assumptions of continuous and symmetric distributions (21–23,27). Yet, RNA-seq counts are both discrete and asymmetrically distributed, such that the analysis of RNA-seq count data requires a much different approach and interpretation from those of microarray data analysis (42).

The Poisson distribution has been proposed for RNA-seq counts that approximates the binomial probability of independently sampled reads (43,44). Given an expected value of $E(Y) = \mu$, a Poisson random variable Y manifests an intrinsic variance of $Var(Y) = \mu$. The Poisson has been largely accepted as a suitable model for the analysis of technical replicates (45,46). However, in large-scale RNA-seq studies, investigators are typically interested in examining biological replicates at different treatment levels in order to interrogate consistent effects of treatments on gene expressions.

Individual subjects in biological replicates engender additional variability over technical ones. The coefficient of dispersion or variance-to-mean ratio $\phi = Var(Y)/\mu$ is a standard measure of additional variability due to biological variations (26,47,48). It has been found to be advantageous in interpreting variability, free from potential finite-number effects due to varying abundances (49,50). Let Y_{ig} be the read count at subject i and gene g. We employ the mean-dispersion model in this paper based on the coefficient of dispersion, where $E(Y_{ig}) = \mu_{ig}$ and $Var(Y_{ig}) = \phi_{ig}\mu_{ig}$. The variance of Y_{ig} consists of a technical μ_{ig} and biological ϕ_{ig} component, where μ_{ig} represents the intrinsic variability due to independent sampling of reads and is biologically uninteresting whereas ϕ_{ig} characterizes the additional variability arising from biological variations. Thus, interpreting the dispersion ϕ_{ig} will be the focus of gene expression variability analysis in large-scale RNA-seq studies. Current procedures for RNA-seq counts often assume the negative binomial with the mean-variance relationship $Var(Y_{ig}) = \mu_{ig}(1 + \alpha_g\mu_{ig})$, where α_g is invariant across subjects at each gene (28–30,51). Compared to the negative binomial, the mean-dispersion framework can allow for more direct interpretation of variability due to biological variations, such that a log FC in ϕ_{ig} can be explicitly attributed towards a log FC in total variance where $\log[Var(Y_{ig})] = \log(\phi_{ig}) + \log(\mu_{ig})$. An additional advantage is that biological variabilities ϕ_{ig} are allowed to vary over both individual genes and subjects, with which the power of large-scale RNA-seq studies can be exploited to incorporate dynamic and complex biological relationships. We will provide a generalized linear model (GLM) framework that can incorporate the effects of both treatments and additional covariates on the mean μ_{ig} and dispersion ϕ_{ig}. This allows the proposed model to account for a wide array of studies, for example, when cases and controls exhibit different variabilities and when variabilities of gene counts may be influenced by additional covariates, such as age, gender, or different stages in a biological process. In the analysis of RNA-seq data, the average expression strength at a gene has been observed to influence expression variability, where genes with decreased average counts tend to have increased overall variability (51). The mean-dispersion GLM accounts for potential changes in the baseline variability due to differences in average expression counts at each gene by incorporating a gene-wise intercept term in the GLM on dispersion. The incorporation of a dynamic variance model also allows for robust analysis of gene expression means, in addition to enabling the analysis of gene expression variability. In this paper, we present the *MDSeq*, an efficient toolset based on the mean-dispersion GLM for the analysis of large-scale RNA-seq studies.

The *MDSeq* utilizes a novel reparametrization of the negative binomial to allow for robust statistical inference and efficient computations of the mean-dispersion model. It includes several important features to address the needs of large-scale RNA-seq studies. (1) As excess zeros can distort model estimation, it is important to account for technical zero counts in order to achieve robust and biologically interpretable results. The *MDSeq* includes a zero-inflated GLM model, that demarcates an excess zero state for technical zeros due to sequencing failures (36) and a random state that originates all of the nonzero counts and some biological zeros due to probabilistic realizations of the random mean-dispersion model. We will demonstrate that the incorporation of excess zeros in modeling RNA-seq counts can significantly improve power in differential analysis of both gene expression mean and variability. (2) Investigators are often interested in evaluating a given set of parameters of interest. For example, it is often of interest to perform hypothesis tests on treatment effects but not necessarily on those of additional covariates, even though they may be incorporated in the GLM. Cook's distance has been proposed for outlier detection in RNA-seq data analysis (51,52). However, it measures the influences of all parameters simultaneously, such that an observation may be identified indistinguishably as an outlier regardless of whether it is influential on treatment effects or merely the additional covariates. A novel procedure is provided with the *MDSeq* that allows compu-

tationally efficient detection of outliers that are influential for statistical inference on user-specified sets of parameters of interest, that can comprise any or all of the treatment effects and coefficients on additional covariates. (3) It is often of interest to determine genes with differential FCs beyond a given threshold in order to facilitate post-experimental verification and provide reproducible results at measurable expression levels. A common procedure is to select differentially expressed genes that satisfy both a log FC threshold and a *p*-value significance level (39,40). However, the procedure is relatively ad hoc and cannot be used to determine whether differentially expressed genes satisfy the log FC threshold with statistical significance. In the *MDSeq*, we develop statistically rigorous procedures for hypothesis tests of both differential mean and variability of expressions at beyond given threshold levels. The development is quite different from those previously proposed for differential analysis of mean expressions (51,71). We will show that the proposed procedure is powerful while controlling type I errors.

The *MDSeq* is compared with a myriad of existing tools using extensive simulations for both differential expression mean and variability analyses. Results suggest that our procedures are robust and powerful in a wide spectrum of data scenarios. The *MDSeq* is further demonstrated on two large-scale datasets from the GTEx project, where we uncovered functionally relevant genes and gene pathways in the human skin and cerebral cortex. Excess zeros, outliers, and the need for hypothesis tests at beyond given threshold levels are illustrated on these two real datasets with the *MDSeq* in Table 1, Supplementary Table S2, and Table 2, respectively. We implemented the *MDSeq* in a user-friendly R package, freely available at https://github.com/zjdaye/MDSeq. The software allows parallel processing with multiple threads for efficient computations.

MATERIALS AND METHODS

Mean-dispersion model

Let Y_{ig} be the read count for sample i and gene g. The *MDSeq* utilizes a novel reparameterization of the real-valued negative binomial $Y_{ig} \sim \text{NB}(\mu_{ig}, \phi_{ig})$ in order to allow gene expression variability to be modeled explicitly based on the coefficient of dispersion. For notational simplicity, we assume a given gene g throughout the 'Materials and Methods' section and do not specify its index g in the read count Y_{ig}.

Consider the real-valued negative binomial or Pólya distribution,

$$Y_i \sim \text{NB}(\mu_i, \phi_i), \tag{1}$$

where $E(Y_i) = \mu_i$ and $Var(Y_i) = \phi_i\mu_i$ for $\phi_i > 1$. The mean-dispersion formulation has the probability model,

$$\Pr(Y_i{=}y_i|\mu_i, \phi_i){=}\frac{\Gamma(y_i+\theta_i)}{\Gamma(y_i+1)\Gamma(\theta_i)}\left(\frac{1}{\phi_i}\right)^{\theta_i}\left(1-\frac{1}{\phi_i}\right)^{y_i}, \tag{2}$$

where $\theta_i = \theta(\mu_i, \phi_i) = \mu_i/(\phi_i - 1)$.

GLM has been applied to extend the classical linear model for RNA-seq counts at the mean expression level (30,34,51,53–55). In this paper, we define the mean-dispersion GLM based on log-linear relationships on both the mean and dispersion, such that

$$\log\mu_i = \sum_j x_{ij}\beta_j \quad \text{and} \quad \log\phi_i = \sum_k u_{ik}\gamma_k, \tag{3}$$

where x_{ij} and u_{ik} are design matrix elements and β_j and γ_k are coefficients for the mean and dispersion, respectively. Intercept terms are included on both the mean and dispersion, such that the dispersion intercept allows the *MDSeq* to account for potential differences in baseline variability due to variations in average expression strengths at each gene (51). The design matrices $\{x_{ij}\}$ and $\{u_{ik}\}$ include contrasts to indicate treatments on subjects. The *MDSeq* allows for a number of contrast coding schemes described in Supplementary Methods. For example, in a simple case-control study, one can set $x_{i1} = u_{i1} = 1$ for all subjects, $x_{i2} = u_{i2} = 0$ for cases, and $x_{i2} = u_{i2} = 1$ for controls. Additional covariates describing clinical, demographic, and other experimental factors may also be included in the design matrices $\{x_{ij}\}$ and $\{u_{ik}\}$. In many instances, the design matrices $\{x_{ij}\}$ and $\{u_{ik}\}$ can be the same. However, different contrasts and covariates can be applied in the *MDSeq* on the mean and dispersion, respectively, to allow applications in a wide array of studies. For example, data sources can be directly incorporated as additional factors in the dispersion GLM, if it is believed that different labs may contribute data at different quality levels and variations.

We used natural logarithms in Equation (3) according to conventions in the GLM literature (56). The natural logarithm and $\log_2$ are related through the identity $\log_2(\cdot) = \log(\cdot)/\log(2)$, and, for convenience, options to output results in the $\log_2$ scale are provided in the *MDSeq* software. The mean-dispersion GLM can be efficiently estimated via constrained optimization techniques (57,58). Further details are provided in Supplementary Methods.

Modeling excess zero counts

Technical excess zeros are often present in a significant proportion of genes in large-scale RNA-seq data. It is often necessary to incorporate them in order to obtain interpretable results for both gene expression mean and variability analyses. We employ the zero-inflated model (59–62),

$$Y_i \sim \begin{cases} 0 & \text{with probability } s, \\ \text{NB}(\mu_i, \phi_i) & \text{with probability } 1-s, \end{cases} \tag{4}$$

where $0 \le s < 1$ is the probability of technical excess zeros. Equation (4) describes two states from which RNA-seq counts may arise. An excess zero state is observed with probability s that generates only zero counts, and a negative binomial state is observed with probability $1 - s$ that generates all of the nonzero counts and a few of the zero counts. That is, the model aims to partition zero counts probabilistically into those arising from technical variations at the excess zero state and those from biological variations at the negative binomial state.

Maximum likelihood (ML) estimates for the zero-inflated mean-dispersion GLM are computed by developing an expectation-maximization (EM) algorithm under constrained optimization (63). Detailed descriptions of algorithm are provided in Supplementary Methods.

Test to determine presence of excess zeros

We evaluate the presence of excess zeros by testing the hypothesis $H_0^s : s = 0$ against $H_a^s : s \neq 0$ in the zero-inflated model of Equation (4). Under the null hypothesis when $s = 0$, all of the zero counts are assumed to arise from biological variations at the negative binomial state. We apply a likelihood ratio test with the statistic $D_s = -2[\mathcal{L}_{MD}(\hat{\beta}, \hat{\gamma}; \mathbf{y}) - \mathcal{L}_{ZIMD}(\hat{\beta}, \hat{\gamma}, \hat{s}; \mathbf{y})]$, where $\mathcal{L}_{MD}(\hat{\beta}, \hat{\gamma}; \mathbf{y})$ and $\mathcal{L}_{ZIMD}(\hat{\beta}, \hat{\gamma}, \hat{s}; \mathbf{y})$ are the log-likelihoods at ML estimates of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_q)^T$ under hypotheses H_0^s and H_a^s, respectively. The log-likelihoods $\mathcal{L}_{MD}(\beta, \gamma; \mathbf{y})$ and $\mathcal{L}_{ZIMD}(\beta, \gamma, s; \mathbf{y})$ are presented in Equations (S2) and (S4), respectively, of Supplementary Methods. Test statistic D_s follows a χ_1^2 distribution under H_0^s, from which we obtain the *p*-value at the tail distribution. The likelihood ratio test is applied to provide more robust inference, whereas the Wald test has been found to be sometimes unstable for inference on s in zero-inflated models (59,60). When H_0^s is rejected, the zero-inflated model of Equation (4) is applied instead of the usual mean-dispersion GLM (Equation (1-3)) in further analyses.

Wald tests for GLM coefficients

Consider the GLM of Equation (3). We apply Wald tests to evaluate significances of the coefficients β_j and γ_k on log mean and dispersion, respectively, with the statistics $W_{\beta_j} = \hat{\beta}_j^2 / Var(\hat{\beta}_j)$ and $W_{\gamma_k} = \hat{\gamma}_k^2 / Var(\hat{\gamma}_k)$, where $Var(\hat{\beta}_j)$ and $Var(\hat{\gamma}_k)$ are obtained as inverses of observed Fisher informations (64). W_{β_j} and W_{γ_k} follow the χ_1^2 distribution under the null hypotheses $\beta_j = 0$ and $\gamma_k = 0$, respectively, from which we obtain the *p*-values at the tail distributions. Wald tests for GLM coefficients can be used, for example, in evaluating significances of additional covariates on RNA-seq counts.

The observed Fisher informations for the mean-dispersion and zero-inflated mean-dispersion models are provided in Supplementary Methods, using closed-form Hessian matrices that allow for efficient computations of test statistics.

Standard tests of differential expression mean and dispersion

Denote $\mathbf{c}^\ell = (c_1^\ell, c_2^\ell, \ldots, c_{L-1}^\ell)^T$ as the contrast vector at factor level ℓ. Then, based on Equation (S1) of Supplementary Methods, log FCs of expressions from factor level ℓ_0 to ℓ_1 can be estimated as $\log(\hat{\mu}_{\ell_1}/\hat{\mu}_{\ell_0}) = \sum_{j=1}^{L-1}(c_j^{\ell_1} - c_j^{\ell_0})\hat{\beta}_j$ and $\log(\hat{\phi}_{\ell_1}/\hat{\phi}_{\ell_0}) = \sum_{j=1}^{L-1}(c_j^{\ell_1} - c_j^{\ell_0})\hat{\gamma}_j$ for differential mean and dispersion, respectively.

Standard procedures consider hypothesis tests of the alternatives $H_a^\mu : \log(\mu_{\ell_1}/\mu_{\ell_0}) \neq 0$ for differential expression mean and $H_a^\phi : \log(\phi_{\ell_1}/\phi_{\ell_0}) \neq 0$ for differential dispersion. Wald statistics W_μ for testing H_a^μ and W_ϕ for testing H_a^ϕ are provided in Equation (S9) of Supplementary Methods, with which the *p*-values are obtained at χ_1^2 tail distributions.

To correct for multiple hypothesis testing across genes, we apply the Benjamini–Yekutieli false discovery rate (FDR) that allows for arbitrary dependence in this paper (65).

Hypothesis tests at beyond a given log fold-change threshold

In large-scale RNA-seq studies with moderate to large numbers of samples, standard tests that evaluate the compliant hypothesis that any change in differential expressions may occur, such as by testing the alternative $H_a^\mu : \log(\mu_{\ell_1}/\mu_{\ell_0}) \neq 0$, would often result in the selection of a large proportion of genes that are only mildly differentially expressed. To allow for experimental replication and interpretation of results, it is often of interest to identify genes with differential changes beyond a given threshold level. In this paper, we develop rigorous procedures based on one-sided hypothesis tests within restricted parameter spaces (66–68) and union-intersection principle (69,70). The development is quite different from those previously proposed for differential analysis of mean expressions (51,71).

We are interested in evaluating whether the differential mean or dispersion of expressions are significant beyond a given log FC threshold; in other words, we wish to test the alternative hypothesis $H_a^{\tau,\mu} : |\log(\mu_{\ell_1}/\mu_{\ell_0})| > \tau$ or $H_a^{\tau,\phi} : |\log(\phi_{\ell_1}/\phi_{\ell_0})| > \tau$, respectively, for some threshold $\tau > 0$. This is accomplished through a two-step procedure. Consider the analysis of differential expression mean. (1) Test of the alternative $H_a^{\tau,\mu}$ is first evaluated asunder as one-sided hypothesis tests of the alternatives $H_{a+}^{\tau,\mu} : \log(\mu_{\ell_1}/\mu_{\ell_0}) \geq \tau$ and $H_{a-}^{\tau,\mu} : \log(\mu_{\ell_1}/\mu_{\ell_0}) \leq -\tau$. Wald statistics are derived under restricted parameter spaces (66–68), whereas *p*-values are computed for each test using mixture distributions (72–74). (2) The *p*-value for testing the composite alternative hypothesis $H_a^{\tau,\mu} : |\log(\mu_{\ell_1}/\mu_{\ell_0})| > \tau$ is obtained as the minimum of *p*-values for testing the alternatives $H_{a+}^{\tau,\mu}$ and $H_{a-}^{\tau,\mu}$ by the union-intersection principle (69,70). That is, $H_a^{\tau,\mu}$ is accepted if either of the alternative $H_{a+}^{\tau,\mu}$ or $H_{a-}^{\tau,\mu}$ is accepted.

Details of thresholded hypothesis tests for differential mean and dispersion are provided in Supplementary Methods. We note that an asymptotically equivalent approach can also be developed based on the likelihood ratio statistics with log likelihoods maximized under restricted parameter spaces (68,74,75). However, this approach requires recomputing GLM estimates under restricted parameter spaces and is computationally more intensive.

Detection of outliers influential for inference on a given set of parameters of interest

RNA-seq data analyses often focus on a set of parameters of interest. For example, in differential expression analysis, one is mainly interested in hypothesis tests involving treatment effects of cases and controls. On the other hand, it is not necessarily of interest to interpret hypothesis tests of additional covariates, although they are often incorporated in the GLM to mitigate conditional effects. In this case, it can be advantageous to focus on the set of parameters of interest, instead of all the parameters as in the Cook's distance (51,52), for efficient detection of outliers.

Suppose that we are interested in evaluating hypotheses based on subsets $\beta_{\mathcal{I}_\beta}$ and $\gamma_{\mathcal{I}_\gamma}$ of the GLM parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_q)^T$, respectively. For instance, in evaluating treatment factors of L levels, the parameters $\beta_{\mathcal{I}_\beta}$ and $\gamma_{\mathcal{I}_\gamma}$ are defined as coefficients $(\beta_1, \beta_2, \ldots, \beta_{L-1})^T$ and $(\gamma_1, \gamma_2, \ldots, \gamma_{L-1})^T$, respec-

tively, on the contrast matrix. A standard likelihood ratio test can be applied for $\beta_{\mathcal{I}_\beta}$ and $\gamma_{\mathcal{I}_\gamma}$ with the statistic $D(\mathbf{y}) = -2[\mathcal{L}_{MD}(\hat{\beta}_{-\mathcal{I}_\beta}, \hat{\gamma}_{-\mathcal{I}_\gamma}; \mathbf{y}) - \mathcal{L}_{MD}(\hat{\beta}, \hat{\gamma}; \mathbf{y})]$, where $\mathcal{L}_{MD}(\hat{\beta}_{-\mathcal{I}_\beta}, \hat{\gamma}_{-\mathcal{I}_\gamma}; \mathbf{y})$ is the maximum log-likelihood with the coefficients of $\beta_{\mathcal{I}_\beta}$ and $\gamma_{\mathcal{I}_\gamma}$ set equal to 0. For outlier detection, we do not consider zero-valued counts, as excess zeros are already accounted for by the zero-inflated GLM (Equation (4)).

Traditional procedures for outlier detection are often based on the leave-one-out approach (52,76). Consider the change in the likelihood ratio statistic when sample i is removed, defined as $I_i = D(\mathbf{y}) - D(\mathbf{y}_{-i})$ where $\mathbf{y}_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)^T$ is constructed by removing y_i from $\mathbf{y}$ and $D(\mathbf{y}_{-\mathbf{i}})$ is the likelihood ratio statistic computed without sample i. The difference of likelihood ratio statistics I_i provides a natural measure of the influence of the ith sample on inferences based on $\beta_{\mathcal{I}_\beta}$ and $\gamma_{\mathcal{I}_\gamma}$, such that an extreme value of I_i may suggest that sample i is an outlier. To identify all the outliers, the influence measure I_i needs to be estimated at each sample i. However, this involves computing the log-likelihoods $\mathcal{L}_{MD}(\hat{\beta}_{-\mathcal{I}_\beta}, \hat{\gamma}_{-\mathcal{I}_\gamma}; \mathbf{y}_{-i})$ and $\mathcal{L}_{MD}(\hat{\beta}, \hat{\gamma}; \mathbf{y}_{-i})$ repetitively with each sample i removed, which can be computationally prohibitive in large-scale RNA-seq studies.

In this paper, we propose to apply a one-step estimator for I_i based on parameter estimates computed only once on all samples (52,56,77,78). The one-step estimator $\hat{I}_i$ is obtained as a weighted sum of standardized deviance and Pearson residuals. Details are provided in Supplementary Methods. We compare $\hat{I}_i$ to a variance-gamma distribution (79,80) and remove sample i from ensuing analyses if $\hat{I}_i$ is below the $(\alpha_{out}/2)$th-quantile or above the $(1 - \alpha_{out}/2)$th-quantile of the variance-gamma, where we use $\alpha_{out} = 0.05$ in this paper. We note that the estimator $\hat{I}_i$ is not guaranteed to follow a variance-gamma distribution due to potential estimation error from outliers and dependence. However, we found that the procedure works well in practice, as extreme outliers can be easily identified with the computationally efficient one-step estimator (see Results). The proposed procedure can be applied on any set of parameters of interest, including all parameters.

Data normalization and preprocessing

Normalization is necessary to account for technical biases of read counts at different samples, such as those due to varying sequencing depths. In this paper, we applied the trimmed mean of M values (TMM) procedure, that has been found to be robust against technical biases (30,81). We adjusted the raw counts by using TMM normalization factors provided by the TMM procedure together with the library sizes. Adjusted counts were subtracted by 0.5 and then raised to the smallest integers for the downstream expression analysis. The *MDSeq* also allows the user to apply other normalization factors, including the relative log expression (RLE) (29), upper-quartile (46), and conditional quantile normalization (cqn) (82) factors. In addition to normalized counts, the *MDSeq* also provides an option to offset sample-specific normalization factors in the GLM with raw counts. Further details are provided in Supplementary Methods.

RESULTS

MDSeq performs the best for gene expression variability analysis in large-scale studies

We compared the *MDSeq* with six other methods that have been proposed for variance heterogeneity analysis in microarray gene expression studies. Bartlett's test is a classical procedure for evaluating unequal variances between groups (22,23,83). The Levene's test is a robust alternative to Bartlett's under non-normal data (22,23,27,84–87). It is further improved upon for data with outlying samples by using the trimmed-mean, in which we removed the top 10% outlying samples (88). The heteroscedastic regression, that extends the simple linear regression with non-constant error variances, has been proposed for detecting genetic loci controlling gene expression variability in microarray studies (24–26,87). We apply the heteroscedastic regression to include additional covariates in this paper; specifically, a normal probability model $N(\mu_i, \sigma_i^2)$ is applied where μ_i and σ_i^2 are linear and log-linear in treatment factors and additional covariates, respectively. The mean-absolute-deviation (MAD) test has been proposed as a robust procedure for differential variability analysis of microarray gene expressions (21,89). Moreover, the Fligner–Killeen test utilizes a nonparametric approach for variance comparisons (90). Using the R programming language, we applied the Bartlett's test with the *bartlett.test* function, Levene's tests with *leveneTest* from the *car* package, heteroscedastic regression with *dglm* using the Gaussian family and log-link options, our implementation of the MAD according to Ho *et al.* 2008 (21), and the Fligner–Killeen test with *fligner.test*.

Differential variability analysis is often applied in order to obtain additional insights beyond those already acquired from standard differential analysis of gene expression means. Thus, in this study, we focused on evaluating scenarios when gene expression means are consistent across cases and controls. To compare with Bartlett's, Levene's, MAD, and Fligner–Killeen tests that do not allow the incorporation of additional covariates, we simulated count data from $\text{NB}(\mu_0, \phi_0)$ for controls and $\text{NB}(\mu_0, 2^{\log 2\text{FC}}\phi_0)$ for cases without additional covariates, where we set $\mu_0 = \exp(5)$ and $\phi_0 = \exp(4)$ as the constant mean and baseline dispersion, respectively. Excess zeros were incorporated according to the zero-inflated model (see 'Materials and Methods' section). Both empirical powers and type I errors for each procedure were estimated as proportions of p-values < 0.05 from 1,000 repetitions, where type I errors were computed under constant variances over cases and controls with $\log_2\text{FC} = 0$. Figure 1 presents type I errors at varying probability of excess zeros and sample sizes. The *MDSeq*, Levene's tests, and heteroscedastic regression have well controlled type I errors at around the theoretical level of 0.05, except when sample sizes are extremely small. On the other hand, both Bartlett's and MAD tests have inflated type I errors in all scenarios with the worst performance at $s = 0.5$ when the probability of excess zeros is relatively large. This suggests that the Bartlett's test, that tends to be sensitive to departure from normality (84), and the MAD test may be inappropriate for RNA-seq count data, especially

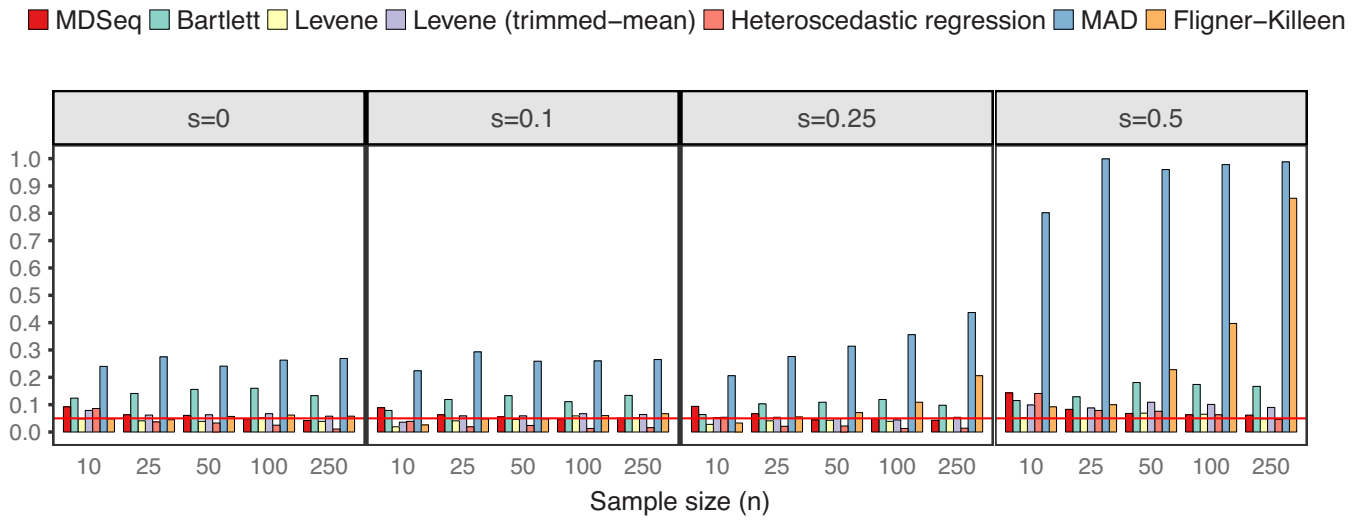

Figure 1. Type I errors in the absence of differential expression variability. There are *n* samples of cases and controls each and varying proportions of excess zeros *s*. The *MDSeq*, Levene's tests, and heteroscedastic regression have well controlled type I errors for moderate to large sample sizes, whereas Bartlett's and MAD tests have highly inflated type I errors. Results are based on 1,000 simulations without additional covariates. Reference lines (in red) are drawn at the 0.05 error rate.

when excess zeros are present. The Fligner-Killeen test has well controlled type I errors at *s* small but can have inflated type I errors at *s* large, suggesting that excess zeros can effect the nonparametric approach. Figure 2 shows that the *MDSeq* dominates Levene's tests and heteroscedastic regression in terms of power, when $|\log_2 FC| = 1$ and $|\log_2 FC| = 2$, in all scenarios, whereas the MDSeq dominates the Fligner-Killeen test except at *s* large when the Fligner-Killeen test can have inflated type I errors. Further, performances of Levene's tests and heteroscedastic regression, that do not incorporate excess zeros, quickly deteriorate with increasing probability of excess zeros *s*, whereas the *MDSeq* remains robust at *s* large. These results suggest that the *MDSeq*, based on a zero-inflated GLM count model, can be advantageous for variability analysis of large-scale RNA-seq count data. Scenarios with additional covariates are presented in Supplementary Figures S1 and S2 for type I errors and powers, respectively.

MDSeq is advantageous for mean expression analysis of counts with excess zeros

The *MDSeq* was compared with six other methods for differential mean analysis. The *DESeq2* (29,51) and *edgeR* (30,34,53–55) are popular procedures for RNA-seq analysis at the mean level. Similar to the *MDSeq*, these methods were developed based on negative-binomial regressions. We compared with the *edgeR* using both conditional maximum likelihood (ML) (55) and quasi-likelihood (QL) (31,91) estimates with the robust option (34). The *voom* from the *limma* package (33,35) circumvents modeling of count data directly by applying the linear model on log-transformed RNA-seq counts. The *tweeDEseq* (92,93) applies a general family of Poisson-Tweedie probability models to better account for heavy tails and scenarios when the amount of excess zeros is modest, whereas these properties of RNA-seq counts were modeled directly using a zero-inflated dispersion model in the *MDSeq*. The *ShrinkBayes* (37,94) incorporates zero-inflated negative binomial models via a Bayesian framework.

We simulated count data without additional covariates from $NB(\mu_0, \phi_0)$ for controls and $NB(2^{\log 2FC}\mu_0, \phi_0)$ for cases, where $\mu_0 = \exp(5)$ and $\phi_0 = \exp(4)$. Excess zeros were incorporated according to the zero-inflated model (see 'Materials and Methods' section). Powers and type I errors were estimated as proportions of *p*-values < 0.05 from 1,000 repetitions, except that Bayesian false discovery rates (computed with the *BFDR* function in *ShrinkBayes* by setting the multcomp option to FALSE) were used for results with the *ShrinkBayes* (37,94). Figure 3 presents type I errors based on scenarios when gene expression means are consistent across cases and controls. The *MDSeq*, *voom* and *tweeDEseq* generally control type I errors well at around 0.05, while, at extremely small sample sizes, the *MDSeq* has moderately inflated type I errors. The *DESeq2* and *edgeR* methods control type I errors well at around 0.05 when no excess zeros are present ($s = 0$), whereas type I errors are nearly 0 when $s > 0$, suggesting that they may be conservative under the presence of excess zeros. Powers are presented in Figure 4. All methods are comparable in terms of power at $s = 0$ when no excess zeros are present. However, the performances of *DESeq2*, *edgeR* methods, and *voom* quickly deteriorate under the presence of excess zeros $s > 0$. This suggests that unaccounted technical zeros may thwart the detection of differential expression means with these methods. The *tweeDEseq*, that applies a general probability model, performs relatively well under the presence of excess zeros. Nonetheless, the *MDSeq* and *ShrinkBayes*, that directly model technical zeros, dominate the *tweeDEseq* in terms of power when the proportion of excess zeros *s* is large. This suggests the need to account for technical excess zeros directly using zero-inflated GLMs.

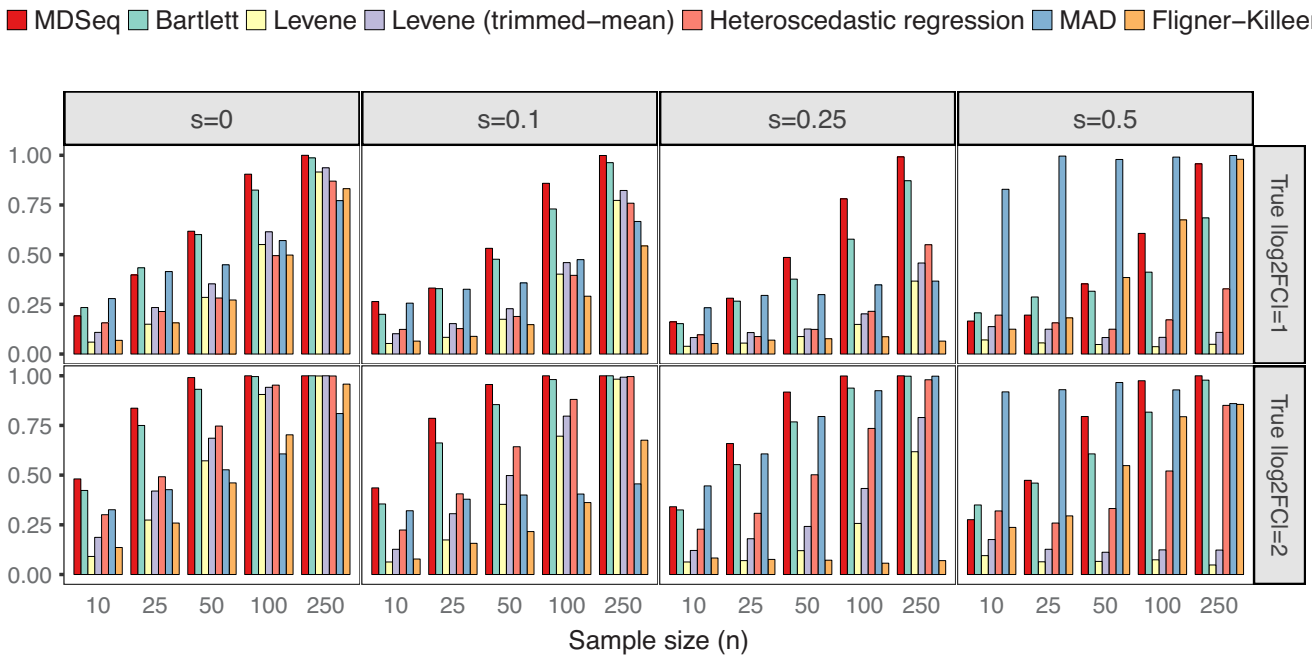

Figure 2. Powers of detecting differential expression variability. There are n samples of cases and controls each and varying proportions of excess zeros s and $\log_2$ fold-changes $\log_2$FC. The *MDSeq* often performs the best when sample sizes are moderate or large. Levene's tests and heteroscedastic regression tend to deteriorate in performance with increasing proportions of excess zeros s. Results are based on 1,000 simulations without additional covariates.

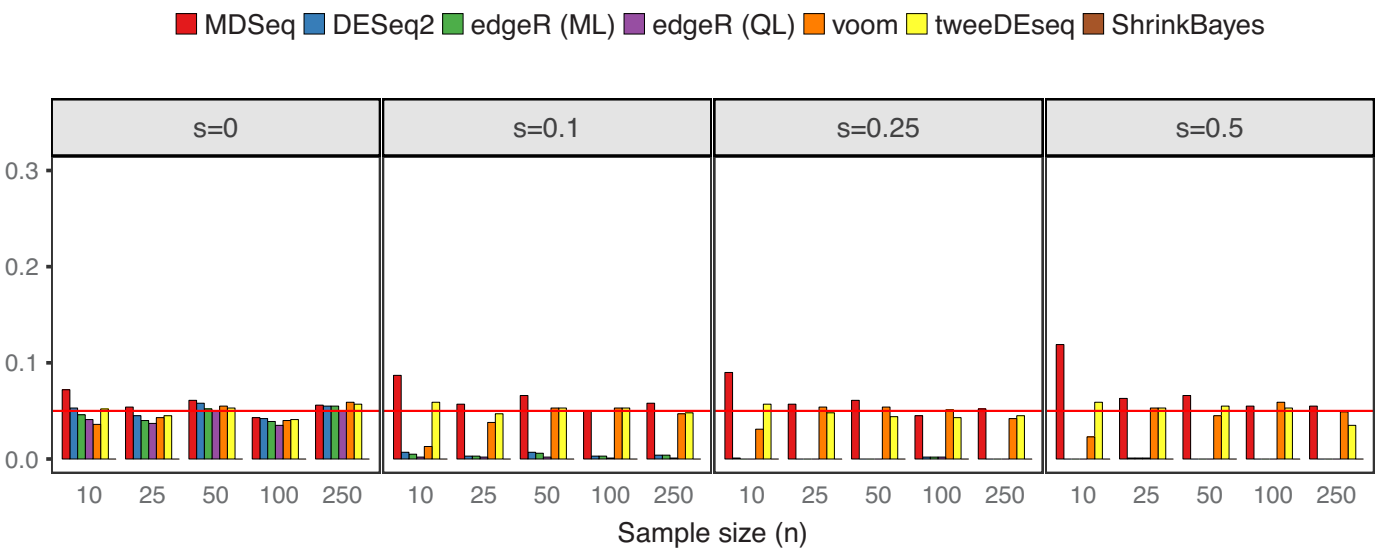

Figure 3. Type I errors in the absence of differential expression means. There are n samples of cases and controls each and varying proportions of excess zeros s. The *MDSeq* controls type I errors well at moderate to large sample sizes. *DESeq2* and *edgeR* methods may be conservative under the presence of excess zeros $s > 0$. Results are based on 1,000 simulations without additional covariates. Reference lines (in red) are drawn at the 0.05 error rate.

Supplementary Table S1 provides computational times for all methods. The *MDSeq* can compute for 1,000 simulations in about half a minute or less in all cases. It is faster than *tweeDEseq* and *ShrinkBayes*. Although employing a more involved EM algorithm, the *MDSeq* is faster than *DESeq2* at large sample sizes, whereas it is slower than *edgeR* and *voom*.

Supplementary Figures S3 and S4 present type I errors and powers, respectively, under scenarios when additional covariates are present. Moreover, Supplementary Figures S5 and S6 present type I errors and powers, respectively, when both expression means and variability are nonconstant across cases and controls. In this scenario, *voom* has highly inflated type I errors at $s = 0$ when no excess zeros are present. Further, type I errors increase with increasing sample sizes. This suggests that the *voom*, although it performs well under equal variances between cases and controls, may be inadequate for RNA-seq count data arising from het-

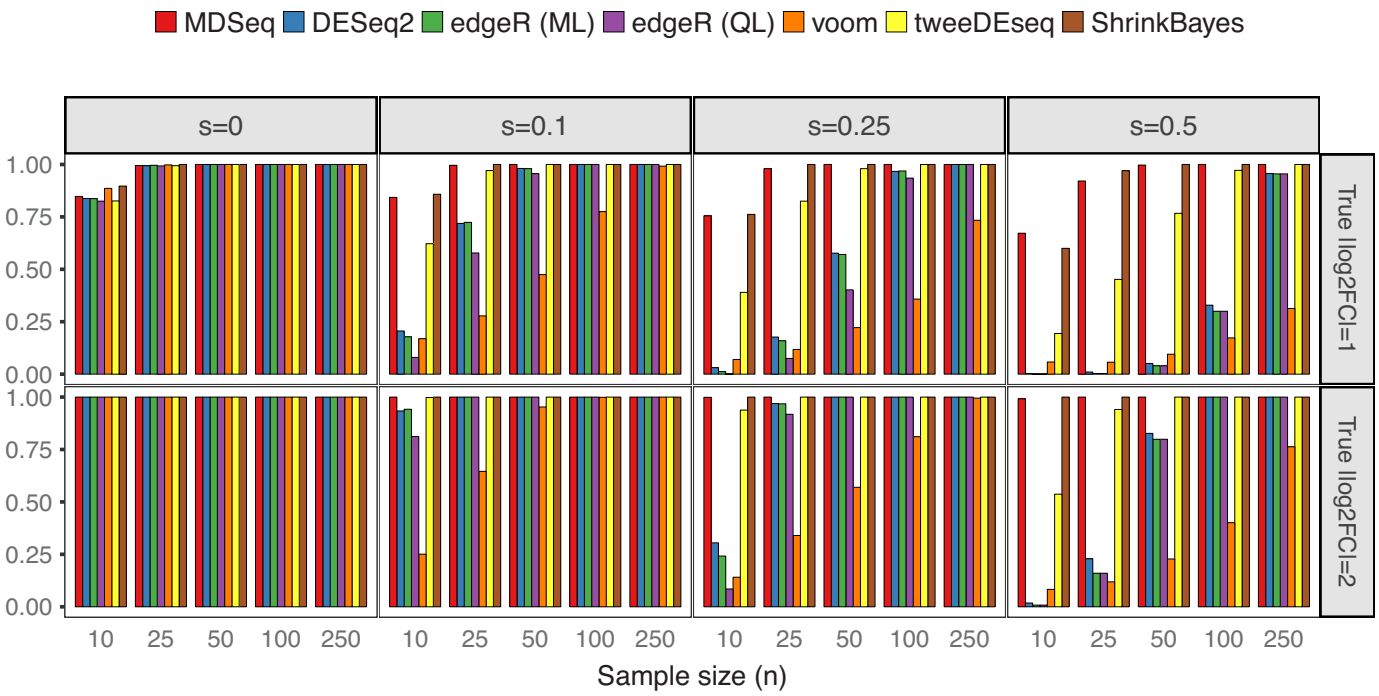

Figure 4. Powers of detecting differential expression means. There are n samples of cases and controls each and varying proportions of excess zeros s and $\log_2$ fold-change $\log_2$FC. The *MDSeq* and *ShrinkBayes* often perform the best among methods compared under the presence of excess zeros $s > 0$. Results are based on 1,000 simulations without additional covariates.

erogeneous count distributions. This may be due to the *voom* relying on a symmetric, normal distributional model (33,35,95) on log-transformed counts, which can be asymmetrically distributed (42). We note that a weighted regression approach may be helpful for mean expression analysis under heteroscedasticity with *voomWithQualityWeights* (96).

MDSeq provides valid hypothesis tests to evaluate absolute log fold changes above a given threshold

We compared hypothesis tests to evaluate absolute log FCs above given thresholds of expression means for the *MDSeq*, *DESeq2* with the *lfcThreshold* option, and *edgeR* using the *glmTreat* function. The *DESeq2* and *edgeR* do not incorporate technical zeros and were not developed for gene variability analysis. Thus, we did not generate excess zeros and focused on differential expression means in this comparison. In Figure 5, the left panel evaluates the hypothesis H_a: $|\log_2\text{FC}| > 1$, whereas the right panel considers H_a: $|\log_2\text{FC}| > 2$. We see that the *MDSeq* controls type I errors well at around or below the 0.05 theoretical level for H_a: $|\log_2\text{FC}| > 1$ (H_a: $|\log_2\text{FC}| > 2$) when the underlying log FC $|\log_2\text{FC}| \leq 1$ ($|\log_2\text{FC}| \leq 2$). On the other hand, *edgeR* has inflated type I errors when the underlying absolute log FC is at or moderately less than the given threshold. Moreover, the *MDSeq* dominates *DESeq2* in terms of power when the underlying log FC is above the given threshold. For example, at test of $|\log_2\text{FC}| > 2$ when the true $|\log_2\text{FC}| = 2.2$, the *MDSeq* has a power of 0.824 while the *DESeq2* has a power of 0.701, that represents an improvement of over 17.5% for the *MDSeq*. Nonetheless, *DESeq2* can be validly applied in most scenarios, as it does not incur inflated type I errors. We note that the *MDSeq* can further incorporate excess zeros and evaluate hypotheses involving expression variability.

Supplementary Figure S7 presents inequality hypothesis tests for differential variability analysis with the *MDSeq*. Hypothesis tests to evaluate absolute log FCs of expression variability have well controlled type I errors, while powers increase more gradually above the given threshold levels compared with those of hypothesis tests of expression means.

Computationally efficient detection of outliers influential on a set of parameters of interest

Figure 6 examines the accuracy of the one-step estimator $\hat{I}_i$ versus the leave-one-out influence measure I_i, obtained by repeatedly computing likelihood estimates with each sample i removed. The one-step estimator $\hat{I}_i$ closely reflects the true influence measure I_i when no outliers are present (Figure 6A), whereas it is slightly less accurate under the presence of outliers (Figure 6B). We see that inaccuracy due to using a one-step estimation usually does not effect the identification of outliers, which requires influence measures to have fairly large magnitudes at the extreme tails of the variance-gamma distribution. This allows the one-step estimation procedure to provide robust yet computationally efficient outlier detection in large-scale RNA-seq studies. Supplementary Figure S8 illustrates a scenario when there are no outliers influential on treatment effects of cases and controls while observations are indiscriminately identified as outliers when influence on all parameters is considered.

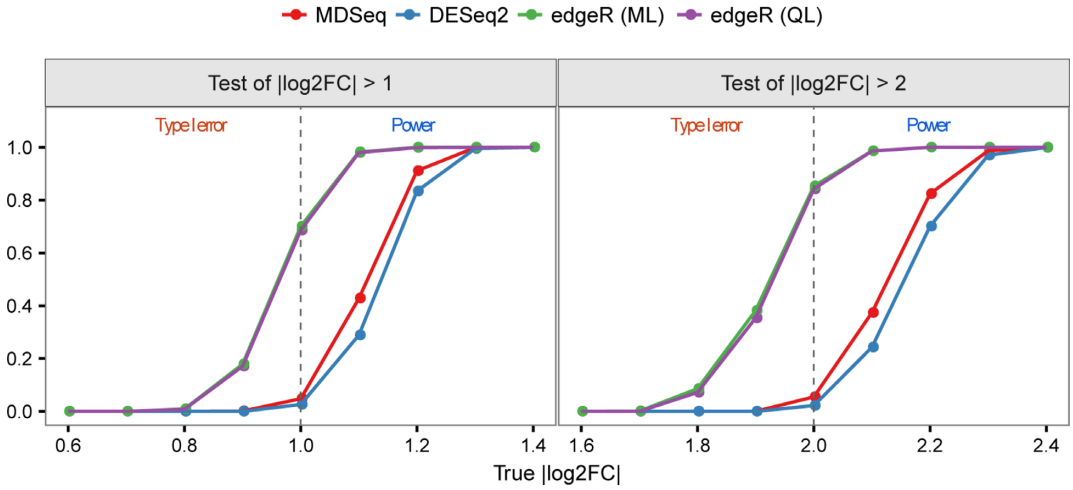

Figure 5. Hypothesis tests to evaluate absolute log fold-changes of expression means above given thresholds. Type I errors are shown at $|\log_2 FC|$ less than or equal to a given threshold, and powers are presented when $|\log_2 FC|$ is greater than the given threshold. The *MDSeq* and *DESeq2* have well controlled type I errors, whereas *edgeR* methods have highly inflated type I errors when $|\log_2 FC|$ is at or moderately less than the given thresholds. The *MDSeq* has greater power than *DESeq2* when $|\log_2 FC|$ is moderately above the given thresholds. There are 500 samples of cases and controls each. Results are based on 1,000 simulations generated from $NB(2^{\log 2FC}\mu_0, \phi_0)$ with $\mu_0 = \exp(5)$ and $\phi_0 = \exp(4)$ for varying $\log_2 FC$. No excess zeros were generated with $s = 0$. Reference lines (in gray) are drawn at the corresponding threshold levels.

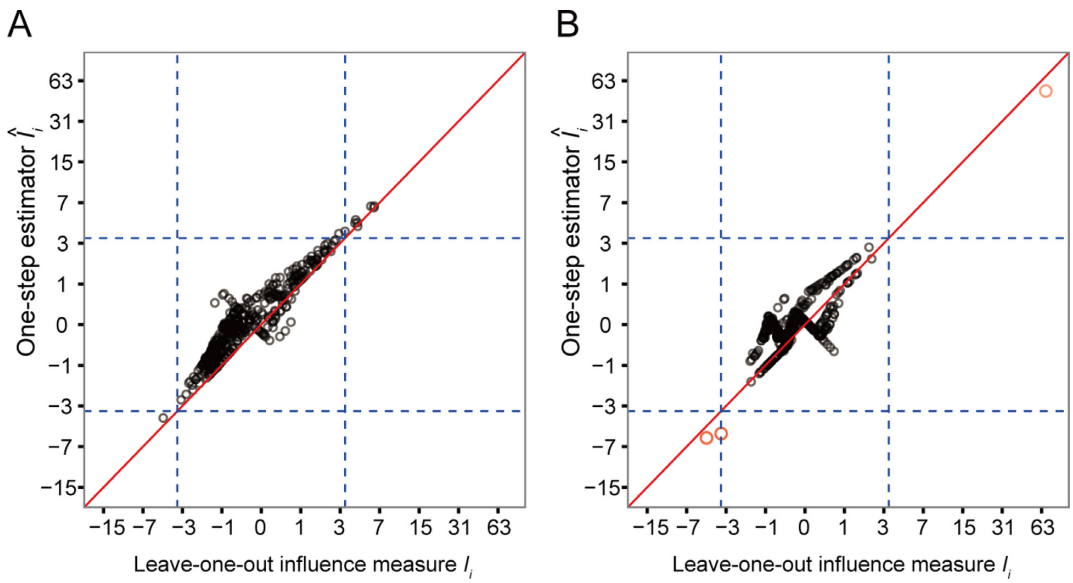

Figure 6. Accuracy of the computationally efficient one-step estimator $\hat{I}_i$. The computationally efficient one-step estimator $\hat{I}_i$ is compared with the leave-one-out influence measure I_i under scenarios when (**A**) there are no outliers and when (**B**) outliers are present. There are $n = 250$ samples each for cases and controls. Counts were generated from $NB(\mu_i, \phi_i)$ for controls and $NB(2^{\log 2FC}\mu_i, \phi_i)$ for cases with $\log_2 FC = 2$, where $\mu_i = \exp(5 + (x_{i1} + x_{i2})/2)$ and $\phi_i = \exp(4 + (x_{i1} + x_{i2})/2)$. Additional covariates x_{i1} and x_{i2} were simulated from binomial distributions $Binom(2, prob = (0.5, 0.5))$. In (**B**), five samples were randomly replaced by outliers simulated from $Pois(\exp(5)\exp(4))$. Non-outlying samples (in black) and outliers (in magenta) are plotted. Reference lines (in dashed blue) are drawn at the $(\alpha_{out}/2)$th- and $(1 - \alpha_{out}/2)$th-quantile of the variance-gamma distribution with $\alpha_{out} = 0.05$. A diagonal reference line (in solid red) is drawn at equality of $\hat{I}_i$ and I_i. In (B), all five outliers were identified by the one-step estimator $\hat{I}_i$.

Analysis of the GTEx brain and skin tissue data

The Genotype-Tissue Expression (GTEx) project provides an expansive repository of large-scale RNA-seq data across tissue types (97). Recent studies have examined tissue-specific gene expression changes at the mean level using the GTEx data (98,99). However, studies have not been conducted, as far as we know, that effectively examined and accounted for differential changes in gene expression variability across tissue types. In this section, we illustrate applications of the *MDSeq* on two large-scale RNA-seq studies that compared the expression profiles of brain tissues from the cerebral cortex (obtained from 96 subjects) against those from the cerebellum (103 subjects) and profiles of skin tissues from sun-exposed (302 subjects) against those from sun-protected (196 subjects) epidermises. RNA-seq read counts were obtained from the GTEx Portal (http://www.gtexportal.org, dbGaP Accession: phs000424.v6.p1). Read counts of 26,800 and 26,144 genes with >0.05 average

Table 1. Genes exhibiting excess zeros in the GTEx tissue data

	With excess zeros	Without excess zeros
Brain tissue	5555 (21.4%)	20396 (78.6%)
Skin tissue	3739 (14.3%)	22347 (85.7%)

Numbers and percentages, in parentheses, of genes are shown. Expressions of genes are considered to have excess zeros if significance test for presence of excess zeros has p-value <0.05.

Table 2. Genes significant for differential mean and dispersion under various hypothesis tests

	Mean only	Variability only	Both mean and dispersion	Total
Cortex versus cerebellum brain tissues				
$\|\log_2 FC\| \neq 0$	7711 (29.96%)	377 (1.46%)	11,968 (46.51%)	20,056 (77.94%)
$\|\log_2 FC\| > 1$	3945 (15.33%)	385 (1.50%)	3214 (12.49%)	7544 (29.32%)
$\|\log_2 FC\| > 2$	1489 (5.79%)	117 (0.45%)	774 (3.01%)	2380 (9.25%)
Sun-exposed versus sun-protected skin tissues				
$\|\log_2 FC\| \neq 0$	4740 (18.29%)	1896 (7.32%)	2757 (10.64%)	9393 (36.25%)
$\|\log_2 FC\| > 1$	12 (0.05%)	53 (0.20%)	49 (0.19%)	114(0.44%)
$\|\log_2 FC\| > 2$	4 (0.02%)	7 (0.03%)	8 (0.03%)	19 (0.07%)

The numbers of genes that are significant decrease with more stringent thresholds. Genes are considered significant at FDR q-value <0.05 and insignificant at FDR q-value ≥0.2. Percentages of significant genes are computed out of 25,909 and 25,734 total genes for analysis of the skin and brain tissue data, respectively.

count per million reads across all samples were retained for gene expression analysis in the brain and skin tissue data, respectively. We note that this is a very lenient filtering criterion applied in our data analyses. In practice, a more robust criterion would require a certain number of samples to be above a given count per million reads for both the case and control groups. Further details on data preprocessing strategies are provided in Supplementary Methods. Raw counts were normalized using the trimmed mean of M values (TMM) method (81). Ensuing analyses were performed using the normalized read counts (see 'Materials and Methods' section.)

Statistical tests were performed on the brain and skin tissue data to determine if counts of individual genes contain excess zeros (see 'Materials and Methods' section). About 21% of individual genes from brain tissues and 14% of genes from skin tissues were found to exhibit significant excess zeros (Table 1). Moreover, estimated proportions of technical zeros can be relatively large for these genes (Supplementary Figure S9). Thus, we see that incorporating excess zeros can be crucial toward the analysis and interpretation of RNA-seq data. Next, outliers were identified at each gene based on the influence statistics $\hat{I}_i$ with respect to coefficients of contrasts (see 'Materials and Methods' section). About 67% of genes from brain tissues and 19% of genes from skin tissues were found to contain at least one outlier at the $\alpha_{out} = 0.05$ level (Supplementary Table S2). Samples identified as outliers were removed before further analyses were performed.

Traditional significance tests for H_a: $|\log_2 FC| \neq 0$ and composite hypothesis tests with respect to a given log FC for H_a: $|\log_2 FC| > 1$ and H_a: $|\log_2 FC| > 2$ were performed. Multiple hypotheses were adjusted using the conservative Benjamini–Yekutieli false discovery rate (FDR) that accounts for arbitrary dependence (65). Due to relatively large sample sizes in these studies, classical tests for no change in expression levels are easily rejected. Hypothesis tests for H_a: $|\log_2 FC| \neq 0$ on the mean and dispersion of RNA-seq counts are significant for about 78% and 36% of genes from the brain and skin tissues, respectively, whereas composite hypothesis tests with respect to a log FC H_a: $|\log_2 FC| > 1$ are significant for about 29% and 0.44% of genes from the brain and skin tissues, respectively (Table 2). Volcano plots depicting p-values based on these hypothesis tests are shown in Supplementary Figures S10 and S11 for differential mean and variability, respectively, of the skin tissue data and Supplementary Figures S12 and S13 for differential mean and variability, respectively, of the brain tissue data. Moreover, a number of genes were found to be significant for age and gender covariates (Supplementary Table S3). The *MDSeq*, based on the GLM, allows for more interpretable results by accounting for potential biological relationships due to additional covariates. Further analyses and interpretation of results are presented as follows.

Differential variability analysis of sun-exposed and sun-protected skin tissues uncovers relevant genes overlooked by mean expression analysis

Table 3 presents genes differentially expressed in the mean or variance. Composite hypothesis tests were performed with respect to at least a two FC in either the mean or dispersion, and the Benjamini–Yekutieli false-discovery rate (FDR) was applied for multiple testing control under arbitrary dependence (65). A myriad of genes in Table 3 that are significant for differential variability but not differential expressions at the mean level are related to sun exposure of skin tissues. For example, studies have shown that ultraviolet radiations can lead to functional irregularities among genes from the histone family (*HIST1H1C*, *HIST1H1E*, *HIST1H2AE*, *HIST1H2BG*, *HIST1H3D*, *HIST1H3H*) (100,101). Keratin is an important structural material in the formation of the epidermis, and disruptions to genes of the keratin family (*KRT17P1*, *KRT39*, *KRT41P*, *KRT6B*) have been found to cause several skin disorders, including the development of carcinomas (102,103). Moreover, genes of the heat-shock protein

Table 3. Genes significant for differential mean and dispersion of sun-exposed versus sun-protected skin tissues with respect to the threshold $|\log_2 FC| > 1$

Gene	Ensembl Gene ID	Differential mean			Differential variability		
		$\log_2$FC	Statistics	FDR q-value	$\log_2$FC	Statistics	FDR q-value
Genes significant for differential mean but not differential dispersion							
C10orf99	ENSG00000188373.4	1.41	22.61	$\mathbf{4.93 \times 10^{-03}}$	1.49	7.15	1.00
FAM83A	ENSG00000147689.12	−1.66	38.80	$\mathbf{1.82 \times 10^{-06}}$	−1.56	8.67	1.00
LHFPL3-AS1	ENSG00000226869.2	−1.57	22.88	$\mathbf{4.36 \times 10^{-03}}$	−1.65	9.18	1.00
NELL2	ENSG00000184613.6	1.67	42.67	$\mathbf{2.81 \times 10^{-07}}$	1.30	2.62	1.00
RP11-252C15.1	ENSG00000254813.1	1.52	29.58	$\mathbf{1.73 \times 10^{-04}}$	1.66	12.82	3.27×10^{-01}
RP11-371I1.2	ENSG00000215808.2	1.69	32.25	$\mathbf{4.49 \times 10^{-05}}$	1.20	0.79	1.00
RP11-529A4.7	ENSG00000255305.1	1.56	19.03	$\mathbf{2.80 \times 10^{-02}}$	1.54	4.13	1.00
SIX1	ENSG00000126778.7	1.57	24.14	$\mathbf{2.40 \times 10^{-03}}$	1.33	2.25	1.00
SNORA75	ENSG00000206885.1	−1.43	21.21	$\mathbf{9.68 \times 10^{-03}}$	−1.06	0.09	1.00
STMN2	ENSG00000104435.9	1.41	18.91	$\mathbf{2.94 \times 10^{-02}}$	1.65	11.87	5.22×10^{-01}
VGLL2	ENSG00000170162.9	1.70	26.33	$\mathbf{8.33 \times 10^{-04}}$	1.35	2.58	1.00
ZNF385B	ENSG00000144331.14	−1.73	90.40	$\mathbf{1.42 \times 10^{-17}}$	−1.42	5.43	1.00
Genes significant for differential variability but not differential mean							
AC003958.2	ENSG00000234859.1	−1.04	0.06	1.00	−2.04	18.98	$\mathbf{2.01 \times 10^{-02}}$
AC018442.1	ENSG00000235683.1	1.26	6.16	1.00	1.96	21.55	$\mathbf{5.77 \times 10^{-03}}$
ACKR2	ENSG00000144648.10	1.14	3.46	1.00	1.86	22.98	$\mathbf{3.07 \times 10^{-03}}$
ACTC1	ENSG00000159251.6	1.36	7.82	1.00	1.94	18.50	$\mathbf{2.43 \times 10^{-02}}$
ALOX15B	ENSG00000179593.11	−0.54	0.00	1.00	−1.99	26.32	$\mathbf{6.50 \times 10^{-04}}$
APOC1	ENSG00000130208.5	−1.01	0.00	1.00	−1.78	17.61	$\mathbf{3.70 \times 10^{-02}}$
AWAT1	ENSG00000204195.3	−1.78	13.93	3.14×10^{-01}	−2.37	23.61	$\mathbf{2.28 \times 10^{-03}}$
CBLN2	ENSG00000141668.5	−1.76	13.09	4.75×10^{-01}	−2.30	22.55	$\mathbf{3.70 \times 10^{-03}}$
CPHL1P	ENSG00000240216.3	−1.41	5.69	1.00	−2.54	33.46	$\mathbf{1.98 \times 10^{-05}}$
CTB-36O1.7	ENSG00000244921.2	1.24	2.22	1.00	2.87	35.56	$\mathbf{7.18 \times 10^{-06}}$
EEF1A1P11	ENSG00000228502.1	0.11	0.00	1.00	1.79	20.11	$\mathbf{1.15 \times 10^{-02}}$
FADS1	ENSG00000149485.12	−0.91	0.00	1.00	−2.06	26.78	$\mathbf{5.19 \times 10^{-04}}$
FAR2	ENSG00000064763.6	−0.66	0.00	1.00	−1.86	17.94	$\mathbf{3.18 \times 10^{-02}}$
GPRC5D	ENSG00000111291.4	−1.78	10.20	1.00	−2.25	18.97	$\mathbf{2.01 \times 10^{-02}}$
HGD	ENSG00000113924.7	−1.15	1.52	1.00	−2.00	24.81	$\mathbf{1.32 \times 10^{-03}}$
HIST1H1C	ENSG00000187837.2	−0.90	0.00	1.00	−1.92	23.89	$\mathbf{2.03 \times 10^{-03}}$
HIST1H1E	ENSG00000168298.4	−1.14	1.36	1.00	−1.85	18.59	$\mathbf{2.34 \times 10^{-02}}$
HIST1H2AE	ENSG00000168274.3	−1.15	1.84	1.00	−2.50	58.83	$\mathbf{8.24 \times 10^{-11}}$
HIST1H2BG	ENSG00000187990.4	−0.99	0.00	1.00	−2.01	26.00	$\mathbf{7.43 \times 10^{-04}}$
HIST1H3D	ENSG00000197409.6	−1.27	5.81	1.00	−1.96	21.90	$\mathbf{4.92 \times 10^{-03}}$
HIST1H3H	ENSG00000203813.4	−1.10	0.71	1.00	−2.09	24.74	$\mathbf{1.34 \times 10^{-03}}$
HRK	ENSG00000135116.5	1.43	12.06	7.45×10^{-01}	2.27	39.48	$\mathbf{1.10 \times 10^{-06}}$
hsa-mir-6723	ENSG00000237973.1	0.57	0.00	1.00	2.09	32.26	$\mathbf{3.60 \times 10^{-05}}$
HSD17B2	ENSG00000086696.6	−0.99	0.00	1.00	−1.99	26.22	$\mathbf{6.73 \times 10^{-04}}$
HSPA5	ENSG00000044574.7	−0.74	0.00	1.00	−1.76	18.28	$\mathbf{2.71 \times 10^{-02}}$
HSPA6	ENSG00000173110.6	−0.91	0.00	1.00	−2.26	42.12	$\mathbf{3.06 \times 10^{-07}}$
ID1	ENSG00000125968.7	−1.10	0.90	1.00	−1.88	20.41	$\mathbf{1.00 \times 10^{-02}}$
KRT17P1	ENSG00000131885.12	−1.42	6.33	1.00	−2.14	18.97	$\mathbf{2.01 \times 10^{-02}}$
KRT39	ENSG00000196859.3	−1.71	11.89	7.93×10^{-01}	−2.12	18.85	$\mathbf{2.09 \times 10^{-02}}$
KRT41P	ENSG00000225438.1	−1.75	8.84	1.00	−2.74	22.08	$\mathbf{4.55 \times 10^{-03}}$
KRT6B	ENSG00000185479.5	−1.19	2.20	1.00	−1.89	20.09	$\mathbf{1.15 \times 10^{-02}}$
MC5R	ENSG00000176136.4	−1.51	7.97	1.00	−2.29	17.52	$\mathbf{3.81 \times 10^{-02}}$
MIR22HG	ENSG00000186594.8	−0.59	0.00	1.00	−1.81	18.87	$\mathbf{2.09 \times 10^{-02}}$
MOGAT2	ENSG00000166391.10	−1.49	12.14	7.22×10^{-01}	−2.11	27.50	$\mathbf{3.71 \times 10^{-04}}$
MTND2P28	ENSG00000225630.1	0.42	0.00	1.00	2.09	34.06	$\mathbf{1.49 \times 10^{-05}}$
MYH3	ENSG00000109063.10	0.45	0.00	1.00	2.20	46.74	$\mathbf{3.32 \times 10^{-08}}$
PDE6A	ENSG00000132915.6	−1.67	12.54	6.09×10^{-01}	−2.25	22.33	$\mathbf{4.10 \times 10^{-03}}$
PDZK1	ENSG00000174827.9	−0.93	0.00	1.00	−2.00	26.96	$\mathbf{4.81 \times 10^{-04}}$
PLIN5	ENSG00000214456.4	−0.95	0.00	1.00	−1.79	17.28	$\mathbf{4.16 \times 10^{-02}}$
RP11-206M11.7	ENSG00000244468.1	−1.04	0.03	1.00	−2.52	17.45	$\mathbf{3.88 \times 10^{-02}}$
RP11-325K4.3	ENSG00000261270.1	−1.16	2.45	1.00	−1.79	17.33	$\mathbf{4.08 \times 10^{-02}}$
RP11-325P15.2	ENSG00000230832.3	−1.55	5.79	1.00	−2.99	23.82	$\mathbf{2.08 \times 10^{-03}}$
RP11-38H17.1	ENSG00000254366.2	1.37	2.58	1.00	1.92	17.07	$\mathbf{4.56 \times 10^{-02}}$
RP11-829H16.3	ENSG00000258525.1	−1.46	12.18	7.22×10^{-01}	−2.28	40.13	$\mathbf{8.08 \times 10^{-07}}$
RP11-845M18.6	ENSG00000257829.1	−1.48	3.82	1.00	−2.25	17.59	$\mathbf{3.71 \times 10^{-02}}$
RP11-849I19.1	ENSG00000263146.2	1.04	0.08	1.00	1.74	21.42	$\mathbf{6.12 \times 10^{-03}}$
RP4-555D20.2	ENSG00000261786.1	1.21	4.52	1.00	1.85	21.00	$\mathbf{7.50 \times 10^{-03}}$
RP5-857K21.7	ENSG00000229344.1	0.71	0.00	1.00	2.66	57.96	$\mathbf{1.24 \times 10^{-10}}$
RPL29P14	ENSG00000241112.1	−1.43	12.32	6.77×10^{-01}	−2.13	28.36	$\mathbf{2.51 \times 10^{-04}}$
SLC6A1	ENSG00000157103.6	−1.29	12.99	4.96×10^{-01}	−1.78	18.75	$\mathbf{2.18 \times 10^{-02}}$
SNORD3A	ENSG00000263934.2	−1.12	0.96	1.00	−1.95	22.95	$\mathbf{3.09 \times 10^{-03}}$
SNORD3D	ENSG00000262202.3	−1.43	7.82	1.00	−2.25	27.69	$\mathbf{3.47 \times 10^{-04}}$
TRIM55	ENSG00000147573.12	−1.63	11.89	7.93×10^{-01}	−2.40	25.54	$\mathbf{9.29 \times 10^{-04}}$

Genes are considered significant at FDR q-value <0.05 and insignificant at FDR q-value ≥0.2. Significant FDR q-values are boldfaced. The $\log_2$ fold-change $\log_2$FC, test statistics, and FDR q-value with respect to the threshold $|\log_2 FC| > 1$ are shown. Positive $\log_2$FC indicates over-expression in the sun-exposed skin tissue, and negative $\log_2$FC indicates over-expression in the sun-protected skin tissue.

family A (*HSPA5*,*HSPA6*) have been shown to effect cell responses to ultraviolet irradiation in human skin tissues (104–106). Absolute log FCs tend to be relatively small for these genes at mean expression levels, which may be caused by the fact that sun-exposures on subjects are not severe enough to trigger a significant change in expressions at the mean or that expression means are conserved for these functionally important genes. Results suggest that the analysis of gene expression variability can be a useful addition to traditional differential gene expression analysis at the mean level and can provide an important component towards the interrogation of gene functionality and genetic effects of human disorders. Supplementary Figures S14 and S15 provide *p*-values from methods at subsets of decreasing sample sizes of the genes found in Table 3, according to an approach from van Wieringen and van de Wiel (107). Full results of significance tests on all genes for both the skin and brain tissues data are provided in Supplementary Data.

Gene expression variability reveals functionally important pathways in the cerebral cortex

We examine significances of gene-set pathways in terms of differential mean and variability via gene-set enrichment analysis (GSEA) (108). GSEA is performed with the software GSEA v2.2.0, obtained from http://software.broadinstitute.org/gsea, using default parameters. Genes are ranked using Wald's statistics derived from the null hypothesis H_0: $\log_2(\text{FC}) = 0$ for differential mean and dispersion. This allows us to account for lowly expressed or low variance genes with potentially significant FCs in the mean or dispersion, respectively (109). GSEA using all potentially differential genes can provide an important ancillary analysis, that complements gene-by-gene significance analysis with respect to a given threshold level.

Table 4 presents enriched pathways in terms of differential mean or variability for the brain tissue data. Normalized enrichment score (NES) accounts for both differences in gene-set sizes and correlations between gene sets, whereas the FDR *q*-value is estimated based on the NES and is adjusted for gene-set sizes and dependency (108). Out of 1,003 pathways, two exhibited enrichment in terms of expression means only, six exhibited enrichment in terms of variability only and 90 exhibited enrichment in both mean and dispersion.

Pathways enriched for gene expression differential variability often indicates functional differences between the cerebral cortex and the cerebellum. *Cell-cell adhesion* is foundational in the reorganization and assembly of neural circuits (110). Enrichment of cell-cell adhesion pathway may suggest increased variation and plasticity of the cerebral cortex relative to the cerebellum (111). *Galactosyltransferase activity* functions to catalyze the transfer of galactose to acceptor molecules. Chronic injection and build-up of D-galactose in mice have been used to model neurodegeneration and brain aging in pharmacological research (112–114). A study has noted that galactosyltransferase activity decreases progressively after birth in the cortex but not the cerebellum in mice, suggesting that galactosyltransferase activity is relatively stable in the cerebellum but variant in the cortex (115).

Moreover, pathways enriched for gene expression variability in the cortex have been found to be associated with common neurodegenerative disorders effecting the cerebral cortex, such as Alzheimer's disease and dementia. *Hemopoietic or lymphoid organ development* involves the progression of hematopoiesis or hemopoiesis by differentiation. It is an important process in the regeneration of brain tissues. Enrichment of hemopoietic development may indicate increased variation of tissue regenerations in the cerebral cortex relative to the cerebellum. The hematopoietic system has been associated with and proposed as target for treatment of Alzheimer's disease (116,117). *Hemostasis* involves the avoidance or arrest of bleeding by mediating the circulation of blood. The brain has been known to possess a sophisticated hemostasis regulatory system that protects itself from hemorrhagic injury (118). Enrichment of hemostasis variability may reflect a functional reaction to microbleeding in the cerebral cortex. Studies have found strong correlation between cerebral microbleeding and leukoaraiosis or diseases effecting cerebral white matters (116,119) and Alzheimer's disease (120). *Kinase regulator activity* and *protein kinase regulator activity* have been known to play important roles in memory and learning (121). Aberrations in the regulation of kinases have been shown to contribute towards the development of Alzheimer's disease (122–124). Therapeutic strategies targeting protein kinases of the central nervous system have been proposed (125).

On the other hand, pathways significantly enriched in terms of mean expression instead of expression variability are involved in essential biological processes that, due to evolutionary pressure, are rarely observed to contribute to common neurological disorders. *Organelle localization* involves essential processes of transportation and maintenance of organelles in regions of the cell that are important to normal functions of neurons. Moreover, *regulation of phosphate metabolic processes* is a primary means of energy regulation. For example, phosphorylation of glucose, the most significant source of energy in the brain, is needed to initiate essential pathways in the usage and storage of glucose.

Pathways significant only for differential mean or differential dispersion for the skin data are provided in Supplementary Table S4. Full GSEA results for both the skin and brain tissues data are available in Supplementary Data.

DISCUSSION

In this paper, we have presented the *MDSeq* that offers an efficient and comprehensive solution set for the analysis of gene expression means and variability in large-scale RNA-seq studies. The *MDSeq* utilizes a novel likelihood-based approach to incorporate the mean-dispersion model in a GLM framework. It introduces a zero-inflated GLM to account for technical excess zeros frequently encountered in RNA-seq data. A new approach is developed for detecting outliers influential on a user-specified set of parameters of interest, that is computationally efficient. Further, statistically rigorous hypothesis tests for gene expression differences beyond given threshold levels are provided for both differential analyses of gene expression means and variability, that allow differentially expressed genes to be identified

Table 4. Significantly enriched pathways for differential mean and dispersion of cortex versus cerebellum brain tissues

GO term	Ontology	No. genes	Differential mean		Differential variability	
			NES	FDR q-value	NES	FDR q-value
Pathways significant for differential mean but not differential dispersion						
Organelle Localization	BP	23	1.67	**0.0376**	1.28	0.207
Positive Regulation Of Phosphate Metabolic Process	BP	21	1.68	**0.0357**	1.25	0.243
Pathways significant for differential variability but not differential mean						
Cell Cell Adhesion	BP	74	1.23	0.275	1.66	**0.0324**
Galactosyltransferase Activity	MF	15	1.15	0.377	1.64	**0.037**
Hemopoietic Or Lymphoid Organ Development	BP	67	1.14	0.405	1.59	**0.0488**
Hemostasis	BP	39	1.30	0.216	1.59	**0.049**
Kinase Regulator Activity	MF	42	1.29	0.222	1.62	**0.0421**
Protein Kinase Regulator Activity	MF	36	1.30	0.216	1.65	**0.0352**

Pathways are considered significant at FDR q-value <0.05 and insignificant at FDR q-value ≥ 0.2. Significant FDR q-values are boldfaced. The normalized enrichment score (NES) and FDR q-value for NES are shown. Positive NES indicates enrichment in the cortex, and negative NES indicates enrichment in the cerebellum.

with statistical significance at biologically interesting levels. The *MDSeq* has been shown with extensive simulation studies to be advantageous for the analysis of gene expression variability on large-scale RNA-seq data and for the analysis of gene expression means of RNA-seq counts with technical excess zeros. The *MDSeq* has been shown to perform well in the simulation scenarios considered for $n \geq 25$. Applications of the *MDSeq* on the analyses of the GTEx skin and brain tissues data have identified functionally relevant genes and gene pathways. In particular, gene variability analysis of the human brain tissue data has revealed pathways associated with common neurodegenerative disorders, such as Alzheimer's disease and dementia.

The mean-dispersion model applied in the *MDSeq* is related to the quasi-Poisson, that considers the mean-variance relationship $E(Y) = \mu$ and $Var(Y) = \phi\mu$ for any positive $\phi > 0$ (56,91,126,127). The quasi-Poisson is developed based on the quasi-likelihood in order to avoid the difficulty of building a probability likelihood model for count data when ϕ can assume an arbitrary positive value (91). In this paper, we focused on modeling the mean-variance relationship $Var(Y_{ig}) = \phi_{ig}\mu_{ig}$ for over-dispersed data when $\phi > 1$. A probability likelihood model is developed using a novel approach based on a reparametrization of the negative binomial (see 'Materials and Methods' section). This allows the *MDSeq* to take advantage of an array of theoretical results and techniques from maximum likelihood theory. For example, the one-step estimator proposed for outlier detection is based on theoretical approximations of probability likelihoods. Moreover, we note that under-dispersion is rarely encountered in RNA-seq studies and can often be attributed to extreme proportions of excess zeros (128). The mean-dispersion model naturally accounts for and qualifies our model for over-dispersion $\phi_{ig} > 1$ using the negative binomial, whereas technical excess zeros are evaluated and demarcated using the zero-inflated GLM. These features allow the proposed model to be robust for the analysis of RNA-seq count data.

The *MDSeq* variance model $Var(Y_{ig}) = \phi_{ig}\mu_{ig}$ is motivated from the coefficient of dispersion $\phi_{ig} = Var(Y_{ig})/\mu_{ig}$, that has been found to be advantageous in evaluating additional variability under varying abundances (49,50). Another measure often used in evaluating additional variability is the coefficient of variation $CV_{ig} = \sqrt{Var(Y_{ig})}/\mu_{ig}$ or coefficient of variation squared $\eta_{ig} = CV_{ig}^2$ (129–131). The *MDSeq* could potentially be developed based on the mean-variance relationship $Var(Y_{ig}) = \eta_{ig}\mu_{ig}^2$ motivated from the coefficient of variation. We note that the reparametrization of the negative binomial in the *MDSeq* can be analogously developed to attain a probability likelihood model for the alternative variance relationship when $\eta_{ig} > 1$. However, investigators are usually interested in the analysis of gene expression variability in order to interrogate additional information beyond those already acquired in the standard analysis of gene expression means. Thus, differential variability analysis is often most informative when mean expression levels are consistent across treatments. In this scenario, as μ_{ig} is undifferentiated, the analysis of gene expression variability would be unaffected by the choice of the mean-variance relationship.

The *MDSeq* utilizes a zero-inflated GLM to account for excess zeros in Equation (4). We note that the probabilistic framework is agnostic to the sources of excess zeros. It only requires estimations of the overall probability of excess zeros at a given gene in order to provide robust inference on biological variations at the random negative binomial state. For example, the *MDSeq* is expected to remain relatively stable when proportions of excess zeros are different across cases and controls but the overall proportions of excess zeros are the same, especially at moderate to large sample sizes (Supplementary Tables S5 and S6).

In this paper, we focused on differential analyses of gene expression means and variability due to the prevalence of case-control data in large-scale RNA-seq studies and their importance towards the identification of functional impacts of genes. The analysis of variance quantitative trait loci (QTLs) that associates genetic variants with quantitative traits has drawn much attention in recent literature (22–26,87,132–135). Variance QTLs can be a source of gene expression variability and play an important role in the genetic regulation of complex traits. The identification of variance QTLs can also help to uncover interactions among genetic variants due to the increased variability of traits influenced

by genetic interactions. The *MDSeq* can be directly applied for the analysis of variance QTLs with RNA-seq data by associating genetic variances with discrete quantitative traits.

Moreover, the *MDSeq* can be applied for other types of high-throughput count data, such as Chromatin Immunoprecipitation (ChIP) sequencing (136,137), CRISPR/Cas assay (138), etc. Standard RNA-Seq expressions are often profiled by averaging over a large number of individual cells. Recent developments have led to the availability of single-cell RNA sequencing (scRNA-seq) data, that characterize gene expressions at each individual cell (139–142). Gene expression variability analysis of scRNA-seq (143–145) will allow the evaluation and interpretation, at unprecedented resolution, of biological variations among individual cells, that can lead to new insights on cell populations effected by tumor mutations (146,147), infectious diseases (148), etc. In future works, we plan to extend the *MDSeq* for the analysis of gene expression variability in these studies.

CONCLUSION

The *MDSeq* is available in an efficient and user-friendly R package at https://github.com/zjdaye/MDSeq. Outlier detection and differential analyses for around 20,000 genes and 200 samples took ∼20 min using four parallel processes on a Windows machine with 3.6-GHz i7-4790 CPUs and 8-GB RAM. With rapidly decreasing cost of NGS, large-scale RNA-seq studies will soon become routinely available. In this paper, we presented the *MDSeq* to fulfill the need for a comprehensive toolset to interrogate both gene expression means and variability in large-scale RNA-seq studies.

A large number of simulation scenarios has been considered in this article and its accompanying Supplementary Materials. We hope that our results by encompassing a wide spectrum of data scenarios will help to guide practitioners in designing their own experiments. The *MDSeq* software also contains a *sim.ZIMD* function that can provide simulated data for type I error and power analysis in additional scenarios.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Shanshan Zhang, Ian Lian and Paul (Chiu-Hsieh) Hsu for helpful comments on the manuscript. In addition, we thank the editor and two anonymous reviewers for constructive suggestions that have led to a much improved manuscript.
Conflict of interest statement. None declared.

REFERENCES

1. Markert,J.M., Fuller,C.M., Gillespie,G.Y., Bubien,J.K., McLean,L.A., Hong,R.L., Lee,K., Gullans,S.R., Mapstone,T.B. and Benos,D.J. (2001) Differential gene expression profiling in human brain tumors. *Physiol. Genomics*, **5**, 21–33.
2. Jiang,Y., Harlocker,S.L., Molesh,D.A., Dillon,D.C., Stolk,J.A., Houghton,R.L., Repasky,E.A., Badaro,R., Reed,S.G. and Xu,J. (2002) Discovery of differentially expressed genes in human breast cancer using subtracted cDNA libraries and cDNA microarrays. *Oncogene*, **21**, 2270–2282.
3. Richer,J.K., Jacobsen,B.M., Manning,N.G., Abel,M.G., Wolf,D.M. and Horwitz,K.B. (2002) Differential gene regulation by the two progesterone receptor isoforms in human breast cancer cells. *J. Biol. Chem.*, **277**, 5209–5218.
4. Gur-Dedeoglu,B., Konu,O., Kir,S., Ozturk,A.R., Bozkurt,B., Ergul,G. and Yulug,I.G. (2008) A resampling-based meta-analysis for detection of differential gene expression in breast cancer. *BMC Cancer*, **8**, 396.
5. Howell,B.G., Solish,N., Lu,C., Watanabe,H., Mamelak,A.J., Freed,I., Wang,B. and Sauder,D.N. (2005) Microarray profiles of human basal cell carcinoma: insights into tumor growth and behavior. *J. Dermatol. Sci.*, **39**, 39–51.
6. Glanzer,J.G., Haydon,P.G. and Eberwine,J.H. (2004) Expression profile analysis of neurodegenerative disease: advances in specificity and resolution. *Neurochem. Res.*, **29**, 1161–1168.
7. Liang,W.S., Dunckley,T., Beach,T.G., Grover,A., Mastroeni,D., Ramsey,K., Caselli,R.J., Kukull,W.A., McKeel,D., Morris,J.C. *et al.* (2008) Altered neuronal gene expression in brain regions differentially affected by Alzheimer's disease: a reference data set. *Physiol. Genomics*, **33**, 240–256.
8. Altar,C.A., Vawter,M.P. and Ginsberg,S.D. (2009) Target identification for CNS diseases by transcriptional profiling. *Neuropsychopharmacology*, **34**, 18–54.
9. Handley,D., Serban,N., Peters,D., O'Doherty,R., Field,M., Wasserman,L., Spirtes,P., Scheines,R. and Glymour,C. (2003) Evidence of cross-hybridization artifact in expressed sequence tags (ESTs) on cDNA microarrays. *Genetics*, http://www.phil.cmu.edu/projects/genegroup/papers/handley2002a.pdf.
10. Yue,H., Eastman,P.S., Wang,B.B., Minor,J., Doctolero,M.H., Nuttall,R.L., Stack,R., Becker,J.W., Montgomery,J.R., Vainer,M. *et al.* (2001) An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. *Nucleic Acids Res.*, **29**, e41.
11. Scott,C.P., VanWye,J., McDonald,M.D. and Crawford,D.L. (2009) Technical analysis of cDNA microarrays. *PLoS One*, **4**, e4486.
12. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
13. Consortium,E.P. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
14. Lonsdale,J., Thomas,J., Salvatore,M., Phillips,R., Lo,E., Shad,S., Hasz,R., Walters,G., Garcia,F., Young,N. *et al.* (2013) The genotype-tissue expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.
15. Lappalainen,T., Sammeth,M., Friedländer,M.R., t'Hoen,P.A., Monlong,J., Rivas,M.A., Gonzàlez-Porta,M., Kurbatova,N., Griebel,T., Ferreira,P.G. *et al.* (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**, 506–511.
16. Hasegawa,Y., Taylor,D., Ovchinnikov,D.A., Wolvetang,E.J., de Torrenté,L. and Mar,J.C. (2015) Variability of gene expression identifies transcriptional regulators of early human embryonic development. *PLoS Genet.*, **11**, e1005428.
17. Raser,J.M. and O'Shea,E.K. (2004) Control of stochasticity in eukaryotic gene expression. *Science*, **304**, 1811–1814.
18. Raser,J.M. and O'Shea,E.K. (2005) Noise in gene expression: origins, consequences, and control. *Science*, **309**, 2010–2013.
19. Zhang,F., Shugart,Y.Y., Yue,W., Cheng,Z., Wang,G., Zhou,Z., Jin,C., Yuan,J., Liu,S. and Xu,Y. (2015) Increased variability of genomic transcription in Schizophrenia. *Scientific Rep.*, **5**, 17995.
20. Ecker,S., Pancaldi,V., Rico,D. and Valencia,A. (2015) Higher gene expression variability in the more aggressive subtype of chronic lymphocytic leukemia. *Genome Med.*, **7**, 1.
21. Ho,J.W.K., Stefani,M., dos Remedios,C.G. and Charleston,M.A. (2008) Differential variability analysis of gene expression and its application to human diseases. *Bioinformatics*, **24**, i390–i398.
22. Pare,G., Cook,N.R., Ridker,P.M. and Chasman,D.I. (2010) On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study. *PLoS Genet.*, **6**, e1000981.
23. Struchalin,M.V., Dehghan,A., Witteman,J.C., van Duijn,C. and Aulchenko,Y.S. (2010) Variance heterogeneity analysis for detection of potentially interacting genetic loci: method and its limitations. *BMC Genet.*, **11**, 92.

24. Ronnegard,L. and Valdar,W. (2011) Detecting major genetic loci controlling phenotypic variability in experimental crosses. *Genetics*, **188**, 435–447.
25. Daye,Z.J., Chen,J. and Li,H. (2012) High-dimensional heteroscedastic regression with an application to eQTL data analysis. *Biometrics*, **68**, 316–326.
26. Hulse,A.M. and Cai,J.J. (2013) Genetic variants contribute to gene expression variability in humans. *Genetics*, **193**, 95–108.
27. Deng,W.Q., Asma,S. and Paré,G. (2014) Meta-analysis of SNPs involved in variance heterogeneity using Levene's test for equal variances. *Eur. J. Hum. Genet.*, **22**, 427–430.
28. Lu,J., Tomfohr,J.K. and Kepler,T.B. (2005) Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach. *BMC Bioinformatics*, **6**, 1.
29. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.
30. 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.
31. Lund,S.P., Nettleton,D., McCarthy,D.J. and Smyth,G.K. (2012) Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Stat. Appl. Genet. Mol. Biol.*, **11**, doi:10.1515/1544-6115.1826.
32. 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**, 519–536.
33. 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.
34. Zhou,X., Lindsay,H. and Robinson,M.D. (2014) Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Res.*, **42**, e91.
35. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.
36. Bhargava,V., Head,S.R., Ordoukhanian,P., Mercola,M. and Subramaniam,S. (2014) Technical variations in low-input RNA-seq methodologies. *Scientific Rep.*, **4**, 3678.
37. van de Wiel,M.A., Leday,G. G.R., Pardo,L., Rue,H., van der Vaart,A.W. and van Wieringen,W.N. (2012) Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics*, **14**, 113–128.
38. George,N.I., Bowyer,J.F., Crabtree,N.M. and Chang,C.W. (2015) An iterative leave-one-out approach to outlier detection in RNA-Seq data. *PLoS One*, **10**, e0125224.
39. Peart,M.J., Smyth,G.K., van Laar,R.K., Bowtell,D.D., Richon,V.M., Marks,P.A., Holloway,A.J. and Johnstone,R.W. (2005) Identification and functional significance of genes regulated by structurally different histone deacetylase inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 3697–3702.
40. Raouf,A., Zhao,Y., To,K., Stingl,J., Delaney,A., Barbara,M., Iscove,N., Jones,S., McKinney,S., Emerman,J., Aparicio,S. *et al.* (2008) Transcriptome analysis of the normal human mammary cell commitment and differentiation process. *Cell Stem Cell.*, **3**, 109–118.
41. Hoyle,D.C., Rattray,M., Jupp,R. and Brass,A. (2002) Making sense of microarray data distributions. *Bioinformatics*, **18**, 576–584.
42. O'hara,R.B. and Kotze,D.J. (2010) Do not log-transform count data. *Methods Ecol. Evol.*, **1**, 118–122.
43. Audic,S. and Claverie,J.-M. (1997) The significance of digital gene expression profiles. *Genome Res.*, **7**, 986–995.
44. Li,J., Witten,D.M., Johnstone,I.M. and Tibshirani,R. (2011) Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, kxr031.
45. Marioni,J.C., Mason,C.E., Mane,S.M., Stephens,M. and Gilad,Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.
46. Bullard,J.H., Purdom,E., Hansen,K.D. and Dudoit,S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 1.
47. Bishay,K., Ory,K., Olivier,M.-F., Lebeau,J., Levalois,C. and Chevillard,S. (2001) DNA damage-related RNA expression to assess individual sensitivity to ionizing radiation. *Carcinogenesis*, **22**, 1179–1183.
48. Hu,M., Zhu,Y., Taylor,J.M., Liu,J.S. and Qin,Z.S. (2012) Using Poisson mixed-effects model to quantify transcript-level gene expression in RNA-Seq. *Bioinformatics*, **28**, 63–68.
49. Thattai,M. and Van Oudenaarden,A. (2001) Intrinsic noise in gene regulatory networks. *Proc. Natll. Acad. Sci. U.S.A.*, **98**, 8614–8619.
50. Kærn,M., Elston,T.C., Blake,W.J. and Collins,J.J. (2005) Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.*, **6**, 451–464.
51. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
52. Cook,R.D. and Weisberg,S. (1982) *Residuals and Influence in Regression*. Chapman and Hall/CRC, NY.
53. Robinson,M.D. and Smyth,G.K. (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**, 2881–2887.
54. Robinson,M.D. and Smyth,G.K. (2008) Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, **9**, 321–332.
55. 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**, 4288–4297.
56. McCullagh,P. and Nelder,J.A. (1989) *Generalized Linear Models*. Chapman and Hall/CRC, London.
57. Nocedal,J. and Wright,S.J. (2006) *Numerical Optimization*. Springer, NY.
58. Lange,K. (2010) *Numerical Analysis for Statisticians*. Springer, NY.
59. Lambert,D. (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.
60. Hall,D.B. (2000) Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, **56**, 1030–1039.
61. Ridout,M., Hinde,J. and Demetrio,C. G.B. (2001) A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, **57**, 219–223.
62. Yau,K. K.W., Wang,K. and Lee,A.H. (2003) Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical J.*, **4**, 437–452.
63. Dempster,A.P., Laird,N.M. and Rubin,D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B*, **39**, 1–38.
64. Efron,B. and Hinkley,D.V. (1978) Assessing the accuracy of the maximum likelihood estimator: observed versus expected fisher information. *Biometrika*, **65**, 457–482.
65. Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1188.
66. Kodde,D.A. and Palm,F.C. (1986) Wald criteria for jointly testing equality and inequality restrictions. *Econometrica*, **54**, 1243–1248.
67. Piegorsch,W.W. (1990) One-sided significance tests for generalized linear models under dichotomous response. *Biometrics*, **46**, 309–316.
68. Fahrmeir,L. and Klinger,J. (1994) Estimating and testing generalized linear models under inequality restrictions. *Stat. Pap.*, **35**, 211–229.
69. Roy,S.N. (1953) On a heuristic method of test construction and its use in multivariate analysis. *Ann. Math. Stat.*, **24**, 220–238.
70. Casella,G. and Berger,R.L. (2002) *Statistical Inference*. Duxbury Press, Pacific Grove.
71. McCarthy,D.J., Chen,Y. and Smyth,G.K. (2009) Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics*, **26**, 765–771.
72. Kudo,A. (1963) A multivariate analogue of the one-sided test. *Biometrika*, **50**, 403–418.
73. Perlman,M.D. (1969) One-sided testing problems in multivariate analysis. *Ann. Math. Stat.*, **40**, 549–567.
74. Gourieroux,C., Holly,A. and Monfort,A. (1982) Likelihood ratio test, Wald test, and Kuhn-Tucker test in linear models with inequality constraints on the regression parameters. *Econometrica*, **50**, 63–80.
75. Wolak,F.A. (1989) Testing inequality constraints in linear econometric models. *J. Econometrics*, **41**, 205–235.

76. Belsley,D.A., Kuh,K. and Welsch,R.E. (1980) *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons, NY.
77. Pregibon,D. (1981) Logistic regression diagnostics. *Ann. Stat.*, **9**, 705–724.
78. Williams,D.A. (1987) Generalized linear model diagnosis using the deviance and single case deletions. *Appl. Stat.*, **36**, 181–191.
79. Seneta,E. (2004) Fitting the variance-gamma model to financial data. *J. Appl. Probab.*, **41**, 177–187.
80. Kotz,S., Kozubowski,T.J. and Podgorski,K. (2001) *The Laplace Distribution and Generalizations*. Birkhauser, Boston.
81. Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, **11**, 1.
82. Hansen,K.D., Irizarry,R.A. and Zhijin,W. (2012) Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*, **13**, 204–216.
83. Bartlett,M.S. (1937) Properties of sufficiency and statistical tests. *Proc. R. Soc. Lond. A*, **160**, 268–282.
84. Levene,H. (1960) Robust Tests for Equality of Variances. In: Olkin,I (ed). *In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, Stanford University Press, Palo Alto, pp. 278–292.
85. Shen,X., Pettersson,M., Ronnegard,L. and Carlborg,O. (2012) Inheritance beyond plain heritability: variance-controlling genes in Arabidopsis thaliana. *PLoS Genet.*, **8**, e1002839.
86. Phipson,B. and Oshlack,A. (2014) DiffVar: a new method for detecting differential variability with application to methylation in cancer and aging. *Genome Biol.*, **15**, 1.
87. Cao,Y., Wei,P., Bailey,M., Kauwe,J.S. and Maxwell,T.J. (2014) A versatile omnibus test for detecting mean and variance heterogeneity. *Genet. Epidemiol.*, **38**, 51–59.
88. Brown,M.B. and Forsythe,A.B. (1974) Robust tests for equality of variances. *J. Am. Stat. Assoc.*, **69**, 364–367.
89. Rousseeuw,P.J. and Croux,C. (1993) Alternatives to the median absolute deviation. *J. Am. Stat. Assoc.*, **88**, 1273–1283.
90. Conover,W.J., Johnson,M.E. and Johnson,M.M. (1981) A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, **23**, 351–361.
91. McCullagh,P. (1983) Quasi-likelihood functions. *Ann. Stat.*, **11**, 59–67.
92. El-Shaarawi,A.H., Zhu,R. and Joe,H. (2011) Modelling species abundance using the Poisson-Tweedie family. *Environmetrics*, **22**, 152–164.
93. Esnaola,M., Puig,P., Gonzalez,D., Castelo,R. and Gonzalez,J.R. (2013) A flexible count data model to fit the wide diversity of expression profiles arising from extensively replicated RNA-seq experiments. *BMC Bioinformatics*, **14**, 254.
94. van de Wiel,M.A., Neerincx,M., Buffart,T.E., Sie,D. and Verheul,H.M. (2014) ShrinkBayes: a versatile R-package for analysis of count-based sequencing data in complex study designs. *BMC Bioinformatics*, **15**, 116.
95. Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Applic. Genet. Mol. Biol.*, **3**, doi:10.2202/1544-6115.1027.
96. Liu,R., Holik,A.Z., Su,S., Jansz,N., Chen,K., San Leong,H., Blewitt,M.E., Asselin-Labat,M.-L., Smyth,G.K. and Ritchie,M.E. (2015) Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses. *Nucleic Acids Res.*, **43**, e97.
97. GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.
98. Baran,Y., Subramaniam,M., Biton,A., Tukiainen,T., Tsang,E.K., Rivas,M.A., Pirinen,M., Gutierrez-Arcelus,M., Smith,K.S., Kukurba,K.R. *et al.* (2015) The landscape of genomic imprinting across diverse adult human tissues. *Genome Res.*, **25**, 927–936.
99. Melé,M., Ferreira,P.G., Reverter,F., DeLuca,D.S., Monlong,J., Sammeth,M., Young,T.R., Goldmann,J.M., Pervouchine,D.D., Sullivan,T.J. *et al.* (2015) The human transcriptome across tissues and individuals. *Science*, **348**, 660–665.
100. Zhang,X., Kluz,T., Gesumaria,L., Matsui,M.S., Costa,M. and Sun,H. (2016) Solar simulated ultraviolet radiation induces global histone hypoacetylation in human keratinocytes. *PLoS One*, **11**, e0150175.
101. Goymer,P. (2006) The DNA's fixed, but what about the histones?. *Nat. Rev. Genet.*, **7**, 904–905.
102. Tan,T.S., Ng,Y.Z., Badowski,C., Dang,T., Common,J.E., Lacina,L., Szeverenyi,I. and Lane,E.B. (2016) Assays to study consequences of cytoplasmic intermediate filament mutations: the case of epidermal keratins. *Methods Enzymol.*, **568**, 219–253.
103. Santos,M., Ballestín,C., Garcia-Martín,R. and Jorcano,J.L. (1997) Delays in malignant tumor development in transgenic mice by forced epidermal keratin 10 expression in mouse skin carcinomas. *Mol Carcinog.*, **20**, 3–9.
104. Ritossa,F. (1962) A new puffing pattern induced by temperature shock and DNP in Drosophila. *Experientia*, **18**, 571–573.
105. Simon,M.M., Reikerstorfer,A., Schwarz,A., Krone,C., Luger,T.A., Jaattela,M. and Schwarz,T. (1995) Heat shock protein 70 overexpression affects the response to ultraviolet light in murine fibroblasts. Evidence for increased cell viability and suppression of cytokine release. *J. Clin. Invest.*, **95**, 926–933.
106. Cao,Y., Ohwatari,N., Matsumoto,T., Kosaka,M., Ohtsuru,A. and Yamashita,S. (1999) TGF-beta1 mediates 70-kDa heat shock protein induction due to ultraviolet irradiation in human skin fibroblasts. *Pflugers Arch.*, **438**, 239–244.
107. van Wieringen,W.N. and van de Wiel,M.A. (2009) Nonparametric testing for DNA copy number induced differential mRNA gene expression. *Biometrics*, **65**, 19–29.
108. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–50.
109. Plaisier,S.B., Taschereau,R., Wong,J.A. and Graeber,T.G. (2010) Rank-rank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures. *Nucleic Acids Res.*, **38**, e169.
110. Chao,D.L., Ma,L. and Shen,K. (2009) Transient cell-cell interactions in neural circuit formation. *Nat. Rev. Neurosci.*, **10**, 262–271.
111. Pascual-Leone,A., Amedi,A., Fregni,F. and Merabet,L.B. (2005) The plastic human brain cortex. *Annu. Rev. Neurosci.*, **28**, 377–401.
112. Xu,F.B. (1985) Sub-acute toxicity of D-galactose. *Proceedings of the Second National Conference on Aging Research*. Herbin.
113. Wei,H., Li,L., Song,Q., Ai,H., Chu,J. and Li,W. (2005) Behavioural study of the D-galactoses induced aging model in C57BL/6J mice. *Behav. Brain Res.*, **157**, 245–251.
114. Cui,X., Zuo,P., Zhang,Q., Li,X., Hu,Y., Long,J., Packer,L. and Liu,J. (2006) Chronic systemic D-galactose exposure induces memory loss, neurodegeration, and oxidative damage in mice: protective effects of R-alpha-lipoic adic. *J. Neurosci.*, **84**, 647–654.
115. Braulke,T. and Biesold,D. (1981) Developmental patterns of galactosyltransferase activity in various regions of rat brain. *J. Neurochem.*, **36**, 1289–1291.
116. Maia,L.F., Vasconcelos,C., Seixas,S., Magalhaes,R. and M,M.C. (2006) Lobar brain hemorrhages and white matter changes: Clinical, radiological and laboratorial profiles. *Cerebrovasc. Dis.*, **22**, 155–161.
117. Lampron,A., Gosselin,D. and Rivest,S. (2011) Targeting the hematopoietic system for the treatment of Alzheimer's disease. *Brain Behav. Immun.*, **25**(Suppl. 1), S71–S79.
118. Fisher,M.J. (2013) Brain regulation of thrombosis and hemostasis: from theory to practice. *Stroke*, **44**, 3275–3285.
119. Yamada,S., Saiki,M., Satow,T., Fukuda,A., Ito,M., Minami,S. and Miyamoto,S. (2012) Periventricular and deep white matter leukoaraiosis have a closer association with cerebral microbleeds than age. *Eur. J. Neurol.*, **19**, 98–104.
120. Pettersen,J.A., Sathiyamoorthy,G., Gao,F.Q., Szilagyi,G., Nadkarni,N.K., St George-Hyslop,P., Rogaeva,E. and Black,S.E. (2008) Microbleed topography, leukoaraiosis, and cognition in probable Alzheimer disease from the Sunnybrook dementia study. *Arch. Neurol.*, **65**, 790–795.
121. Giese,K.P. and Mizuno,K. (2013) The roles of protein kinases in learning and memory. *Learn. Mem.*, **20**, 540–552.
122. Kawamata,T., Taniguchi,T., Mukai,H., Kitagawa,M., Hashimoto,T., Maeda,K., Ono,Y. and Tanaka,C. (1998) A protein kinase, PKN, accumulates in Alzheimer neurofibrillary tangles and

associated endoplasmic reticulum-derived vesicles and phosphorylates tau protein. *J. Neurosci.*, **18**, 7402–7410.
123. Cai,Z., Yan,L.J., Li,K., Quazi,S.H. and Zhao,B. (2012) Roles of AMP-activated protein kinase in Alzheimer's disease. *Neuromol. Med.*, **14**, 1–14.
124. Martin,L., Latypova,X., Wilson,C.M., Magnaudeix,A., Perrin,M.L., Yardin,C. and Terro,F. (2013) Tau protein kinases: involvement in Alzheimer's disease. *Ageing Res. Rev.*, **12**, 289–309.
125. Chico,L.K., van Eldik,L.J. and Watterson,D.M. (2010) Targeting protein kinases in central nervous system disorders. *Nat. Rev. Drug Discov.*, **8**, 892–909.
126. Wedderburn,R.W. (1974) Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika*, **61**, 439–447.
127. Smyth,G.K. (1989) Generalized linear models with varying dispersion. *J. R. Stat. Soc. Ser. B*, **51**, 47–60.
128. Famoye,F. and Singh,K.P. (2006) Zero-inflated generalized Poisson regression model with an application to domestic violence data. *J. Data Sci.*, **4**, 117–130.
129. Elowitz,M.B., Levine,A.J., Siggia,E.D. and Swain,P.S. (2002) Stochastic gene expression in a single cell. *Science*, **297**, 1183–1186.
130. Li,J., Liu,Y., Kim,T., Min,R. and Zhang,Z. (2010) Gene expression variability within and between human populations and implications toward disease susceptibility. *PLoS Comput. Biol.*, **6**, e1000910.
131. Jimenez-Gomez,J.M., Corwin,J.A., Joseph,B., Maloof,J.N. and Kliebenstein,D.J. (2011) Genomic analysis of QTLs and genes altering natural variation in stochastic noise. *PLoS Genet.*, **7**, e1002295.
132. Yang,J., Loos,R.J., Powell,J.E., Medland,S.E., Speliotes,E.K., Chasman,D.I., Rose,L.M., Thorleifsson,G., Steinthorsdottir,V., Mägi,R. *et al.* (2012) FTO genotype is associated with phenotypic variability of body mass index. *Nature*, **490**, 267–272.
133. Brown,A.A., Buil,A., Viñuela,A., Lappalainen,T., Zheng,H.-F., Richards,J.B., Small,K.S., Spector,T.D., Dermitzakis,E.T. and Durbin,R. (2014) Genetic interactions affecting human gene expression identified by variance association mapping. *Elife*, **3**, e01381.
134. Ayroles,J.F., Buchanan,S.M., O'Leary,C., Skutt-Kakaria,K., Grenier,J.K., Clark,A.G., Hartl,D.L. and de Bivort,B.L. (2015) Behavioral idiosyncrasy reveals genetic control of phenotypic variability. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 6706–6711.
135. Metzger,B.P., Yuan,D.C., Gruber,J.D., Duveau,F. and Wittkopp,P.J. (2015) Selection on noise constrains variation in a eukaryotic promoter. *Nature*, **521**, 344–347.
136. Collas,P. (2010) The current state of chromatin immunoprecipitation. *Mol. Biotechnol.*, **45**, 87–100.
137. Niu,W., Lu,Z.J., Zhong,M., Sarov,M., Murray,J.I., Brdlik,C.M., Janette,J., Chen,C., Alves,P., Preston,E. *et al.* (2011) Diverse transcription factor binding features revealed by genome-wide ChIP-seq in C. elegans. *Genome Res.*, **21**, 245–254.
138. Zhou,Y., Zhu,S., Cai,C., Yuan,P., Li,C., Huang,Y. and Wei,W. (2014) High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. *Nature*, **509**, 487–491.
139. Saliba,A.-E., Westermann,A.J., Gorski,S.A. and Vogel,J. (2014) Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res.*, **42**, 8845–8860.
140. Grün,D. and van Oudenaarden,A. (2015) Design and analysis of single-cell sequencing experiments. *Cell*, **163**, 799–810.
141. Kolodziejczyk,A.A., Kim,J.K., Svensson,V., Marioni,J.C. and Teichmann,S.A. (2015) The technology and biology of single-cell RNA sequencing. *Mol. Cell*, **58**, 610–620.
142. Kowalczyk,M.S., Tirosh,I., Heckl,D., Rao,T.N., Dixit,A., Haas,B.J., Schneider,R.K., Wagers,A.J., Ebert,B.L. and Regev,A. (2015) Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res.*, **25**, 1860–1872.
143. Munsky,B., Neuert,G. and van Oudenaarden,A. (2012) Using gene expression noise to understand gene regulation. *Science*, **336**, 183–187.
144. Dueck,H., Khaladkar,M., Kim,T.K., Spaethling,J.M., Francis,C., Suresh,S., Fisher,S.A., Seale,P., Beck,S.G., Bartfai,T. *et al.* (2015) Deep sequencing reveals cell-type-specific patterns of single-cell transcriptome variation. *Genome Biol.*, **16**, 1.
145. Lv,D., Wang,X., Dong,J., Zhuang,Y., Huang,S., Ma,B., Chen,P., Li,X., Zhang,B., Li,Z. *et al.* (2016) Systematic characterization of lncRNAs' cell-to-cell expression heterogeneity in glioblastoma cells. *Oncotarget*, **7**, 18403–18414.
146. Olmos,D., Arkenau,H.-T., Ang,J., Ledaki,I., Attard,G., Carden,C., Reid,A., A'Hern,R., Fong,P., Oomen,N. *et al.* (2009) Circulating tumour cell (CTC) counts as intermediate end points in castration-resistant prostate cancer (CRPC): a single-centre experience. *Ann. Oncol.*, **20**, 27–33.
147. Kim,K.-T., Lee,H.W., Lee,H.-O., Kim,S.C., Seo,Y.J., Chung,W., Eum,H.H., Nam,D.-H., Kim,J., Joo,K.M. *et al.* (2015) Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. *Genome Biol.*, **16**, 1.
148. Avraham,R., Haseley,N., Brown,D., Penaranda,C., Jijon,H.B., Trombetta,J.J., Satija,R., Shalek,A.K., Xavier,R.J., Regev,A. *et al.* (2015) Pathogen cell-to-cell variability drives heterogeneity in host immune responses. *Cell*, **162**, 1309–1321.