1 The penalized likelihood

Assume the read counts for gene g and sample i, Y_{gi}, follows negative binomial distribution with mean ν_{gi} and dispersion ϕ_g. Under this parametrization the probability mass function for Y_{gi} can be derived as:

$$P(Y_{gi} = y|\nu_{gi}, \phi_g) = \frac{\Gamma(y + \phi_g^{-1})}{y!\Gamma(\phi_g^{-1})} \left(\frac{1}{1 + \nu_{gi}\phi_g}\right)^{\phi_g^{-1}} \left(\frac{\nu_{gi}\phi_g}{1 + \nu_{gi}\phi_g}\right)^{y}$$

We assume a log-normal prior on ϕ_g, e.g., $\phi_g \sim$ log-normal(m_0, τ^2), with density function:

$$f(\phi_g) = \frac{1}{\phi_g\sqrt{2\pi\tau^2}} e^{-\frac{(\log(\phi_g) - m_0)^2}{2\tau^2}}$$

Then the conditional posterior distribution for ϕ_g given all counts and means is

$$p(\phi_g|Y_{gi}, \nu_{gi}, i = 1, \ldots, n) \propto f(\phi_g) \prod_i P(Y_{gi}|\nu_{gi}, \phi_g).$$

The logarithm of the posterior distribution is proportional to

$$\begin{aligned} & \sum_i \psi(\phi_g^{-1} + Y_{gi}) - n\psi(\phi_g^{-1}) - \phi_g^{-1} \sum_i \log(1 + \nu_{gi}\phi_g) \\ + & \sum_i Y_{gi}[\log(\nu_{gi}\phi_g) - \log(1 + \nu_{gi}\phi_g)] \\ - & \frac{[\log(\phi_g) - m_0]^2}{2\tau^2} - \log(\phi_g) - \log(\tau), \end{aligned}$$

as shown in Equation 4.1 in the paper.

2 Moment estimator of dispersion parameter

Given $Y_{gi} \sim NB(s_i\mu_{g,k(i)}, \phi_g)$, we have

$$E[Y_{gi}] = s_i\mu_{g,k(i)},\ E[Y_{gi}^2] = s_i\mu_{g,k(i)} + (\phi_g + 1)s_i^2\mu_{g,k(i)}^2,\ Var(Y_{gi}) = s_i\mu_{g,k(i)} + \phi_g s_i^2\mu_{g,k(i)}^2.$$

Define a new random variable $z_{gi} \equiv \frac{Y_{gi}^2 - Y_{gi}}{s_i^2}$, we have

$$E[z_{gi}] = (E[Y_{gi}]^2 - E[Y_{gi}])/s_i^2 = \mu_{g,k(i)}^2(\phi_g + 1).$$

Equate the observed and expected first moment, we have:

$$\sum_i z_{gi} = \sum_i [\mu_{g,k(i)}^2(\phi_g + 1)] = (\phi_g + 1) \sum_i \mu_{g,k(i)}^2,$$

which leads to an estimator for ϕ_g as: $\hat{\phi}_g = \frac{\sum_i z_{gi}}{\sum_i \hat{\mu}_{g,k(i)}^2} - 1$, where $\hat{\mu}_{g,k(i)} = \frac{\sum_{j:k(j)=k(i)} Y_{gj}/s_j}{n_{k(i)}}$ is the estimate of $\mu_{g,k(i)}$ defined in Section 4 of the paper.

3 Derivation of the Wald statistic

The variance of the estimated group 1 mean, $\hat{\mu}_{g,1}$, can be derived as:

$$Var(\hat{\mu}_{g,1}) = \frac{1}{n_1^2} \sum_{j:k(j)=1} \left(\frac{\mu_{g,1}}{s_j} + \phi_g \mu_{g,1}^2 \right) = \frac{1}{n_1^2} \left[\mu_{g,1} \left(\sum_{j:k(j)=1} \frac{1}{s_j} \right) + n_1 \mu_{g,1}^2 \phi_g \right],$$

which can be estimated by

$$\frac{1}{n_1^2} \left[\hat{\mu}_{g,1} \left(\sum_{j:k(j)=1} \frac{1}{s_j} \right) + n_1 \hat{\mu}_{g,1}^2 \tilde{\phi}_g \right].$$

With this result, the Wald test statistics can be constructed in the typical way.

4 Simulation settings

Extensive simulations were conducted to evaluate the performance of DE detection of the proposed method. To make the simulations more realistic, simulation parameters are mainly derived from real data (Gilad and Cheung data). In all simulations, the library sizes and mean expressions were randomly sampled from those calculated from real data. The dispersions were generated from parametric distributions with parameters computed from real data.

To be specific, we first estimate the dispersions from Gilad or Cheung data. Then under parametric assumptions that these estimated dispersions are from log-normal or Gamma distributions, the distribution parameters were calculated based on logarithm dispersions. Under log-normal assumption, we have $\phi_g \sim logN(-2.70, 1.44^2)$ from Gilad, and $\phi_g \sim logN(-1.72, 1.07^2)$ from Cheung data. Under Gamma distribution assumption, the distributions for ϕ_g are $Gamma(0.48, 0.31)$ from Gilad and $Gamma(0.21, 1.73)$ from Cheung data. These distributions were used to generate ϕ_g's in the simulations. Using different parametric model to generate ϕ_g demonstrates the robustness of the proposed method. Expressions of 20000 genes were generated in most simulations except for Figure 3, where 2000 genes were considered to make the simulation scenario identical to that in Robinson and Smyth (2007). All simulations were performed for two treatment groups with 5% of the genes assumed to be differentially expressed. Data were generated for 4 replicates in each treatment group.

5 Validation of simulated data

Figure S1 compares the estimated dispersions from the Cheung data and the simulated data based on the Cheung data and under log-normal dispersion. The two histograms appear very similar, indicating that the simulated data mimic the real data well in the distribution of ϕ.

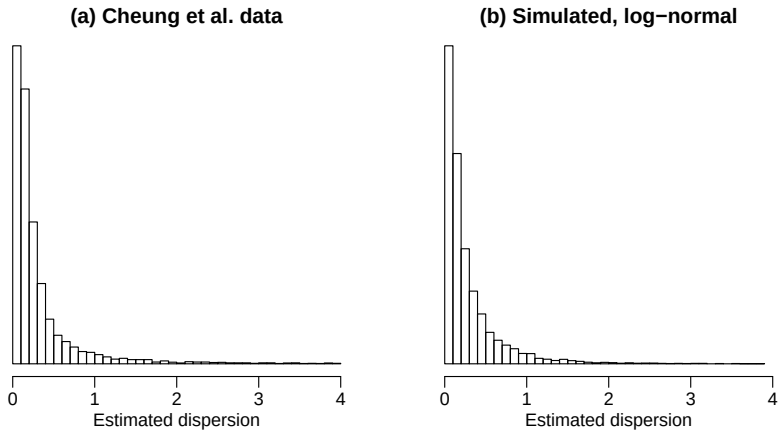

Figure S1: Histograms of estimated dispersions from (a) the Cheung data set; and (b) simulated data using log normal dispersion, with parameters estimated from Cheung data.

6 Dispersion estimates from different methods

Figure 1 in the paper compares the dispersion estimates from different methods, stratified by mean counts. Figure S2 below shows the same results, but different strata were plotted separately in different panels.

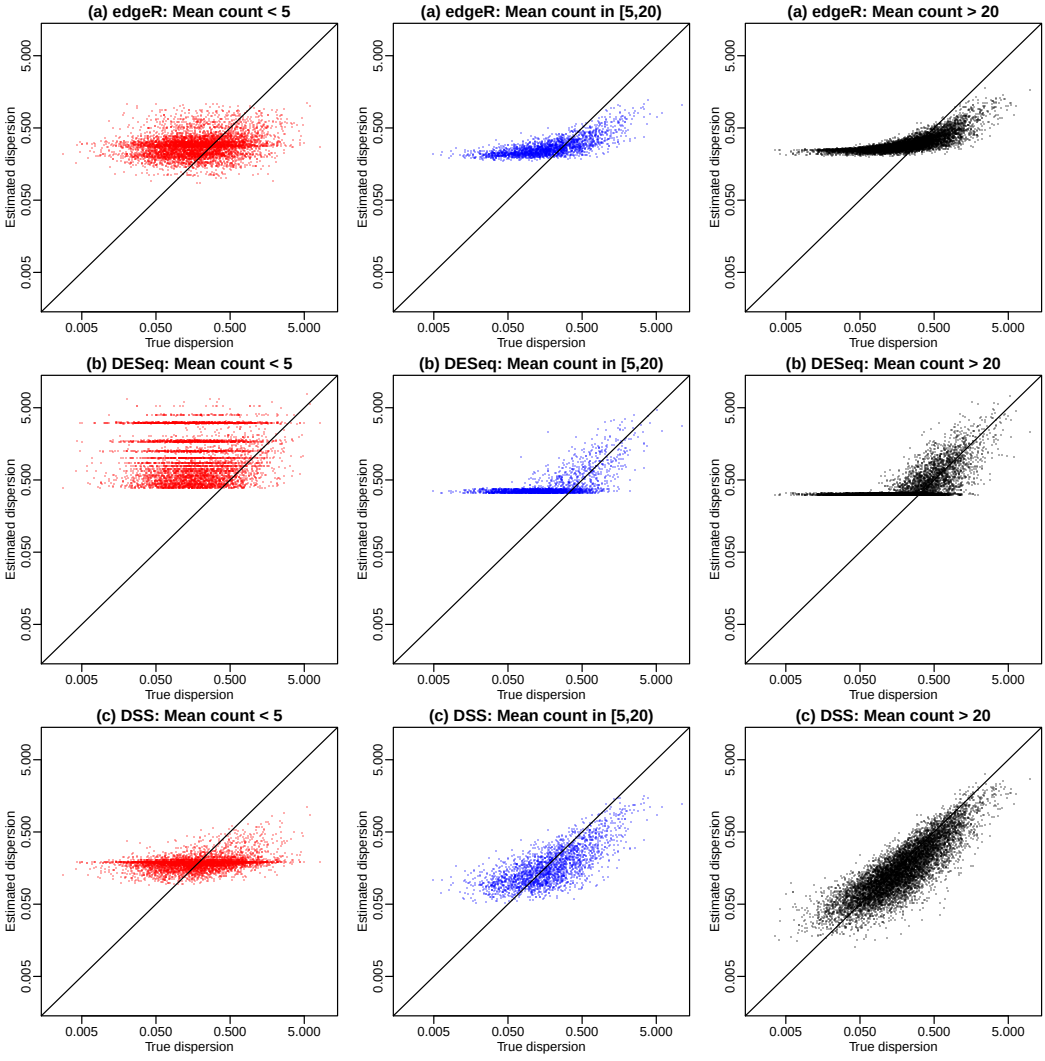

Figure S2: Comparison of dispersion estimates from different methods, stratified by mean counts.

7 Comparison of MSEs of dispersion estimates

Figure 3 in the paper shows the comparison of distribution of MSE for dispersion estimations from edgeR, DESeq and DSS over 50 simulations. Below Figure S3 shows the similar comparison, but genes are stratified by the true means under three strata: $\mu < 5, 5 \leq \mu < 50, \mu \geq 50$. DSS provides the best performance in all strata.

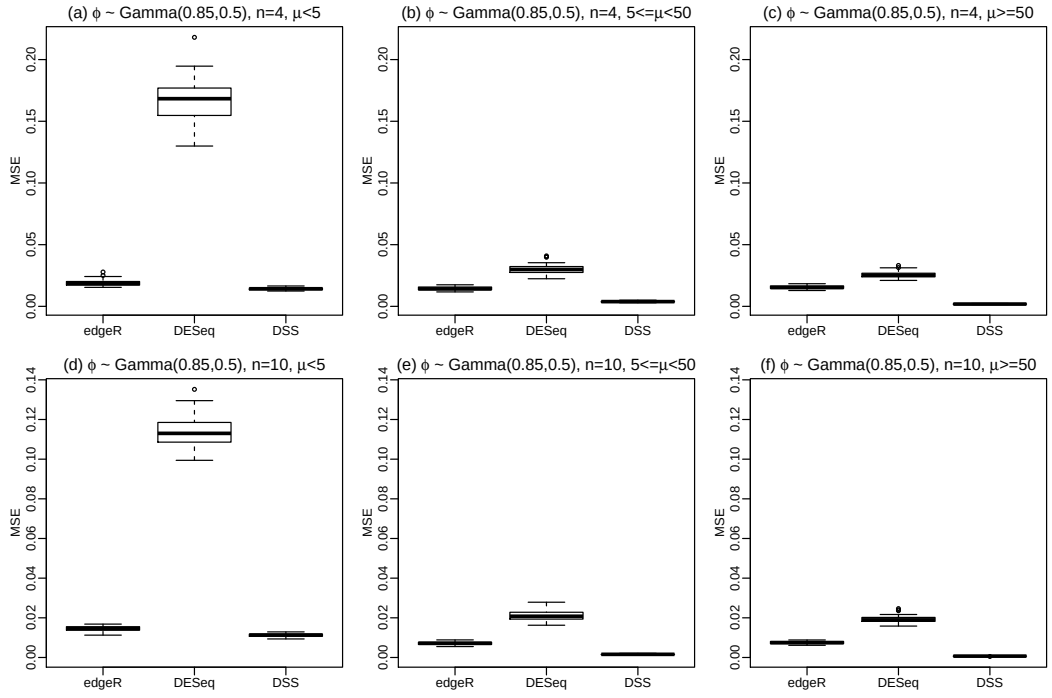

Figure S3: Boxplots comparing distribution of MSE for dispersion estimations from edgeR, DESeq and DSS over 50 simulations. Simulation settings are the same as for Figure 3 in the paper. Genes are stratified by mean, and boxplots are generated for all strata.

8 Dispersion estimates from edgeR using different options

There are several parameters in the function `estimateTagwiseDisp` from edgeR to estimate the dispersions from data, some with a finite number of options (for example, `trend` has three options, `method` has two options) and some are tuning parameters (for example, `prior.n, prop.used, grid.length`) (The default settings are `trend="movingave", prop.used=0.3, method="grid"`). We tested on a few flavors in addition to the default setting. Figure S4 shows the estimated versus true dispersions under different settings.

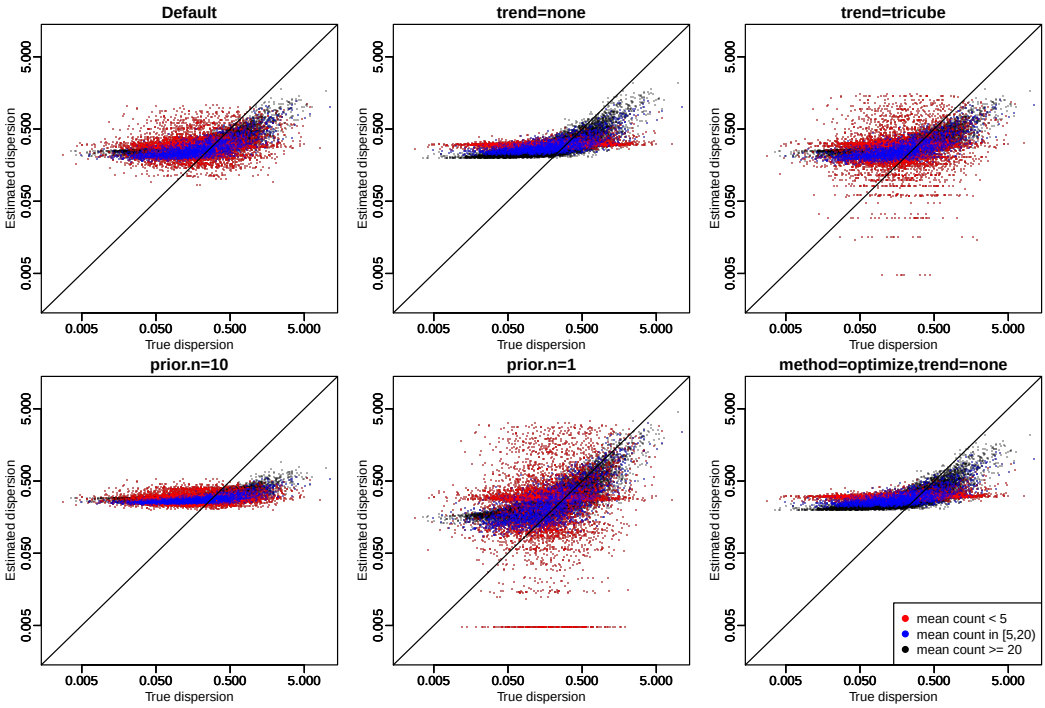

Figure S4: Dispersion estimates from edgeR under different options. The title of each sub-panel indicates the deviation from the default setting.

9 Comparison of DE detection using ROC curves

Figure 4 in the paper compares DE detection accuracies from different methods. ROC curves from the same simulations are shown in Figure S5.

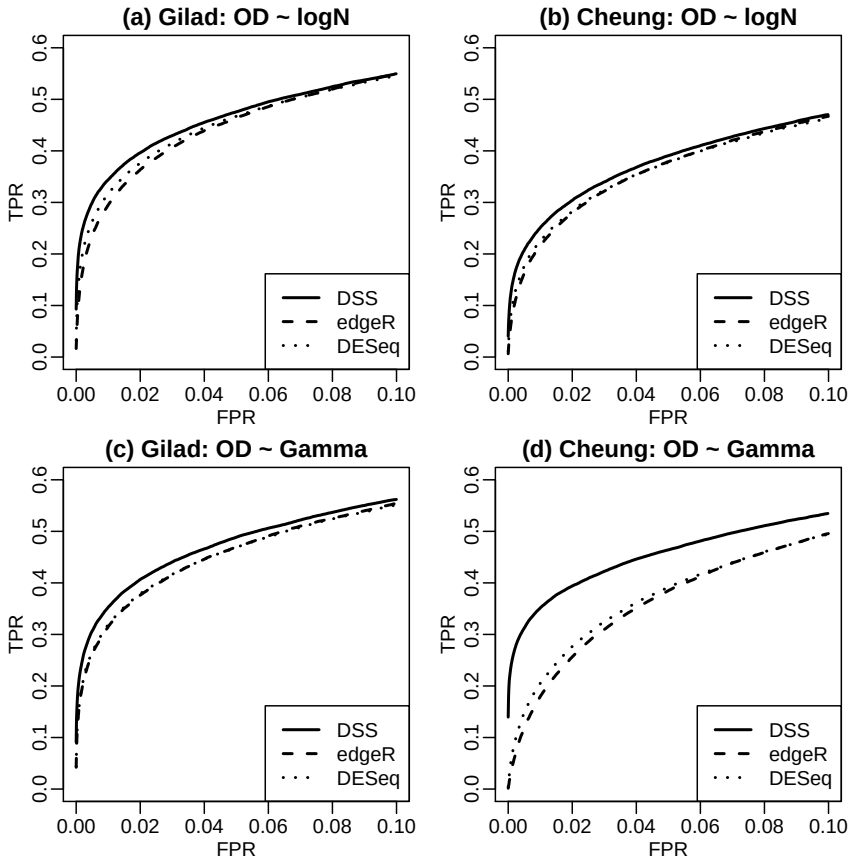

Figure S5: ROC curves for comparing DE detection results from DSS, edgeR and DESeq. The curves are averaged ROC from 50 simulations. All simulations are based on 20000 genes, and 5% of the genes are true DE. Dispersion are generated from log-normal distributions (Figures a and b) or Gamma (Figures c and d). Parameters for dispersion distributions were estimated from real data (Gilad for Figures a and c, Cheung for Figures b and d).

10 Additional simulation for DE detection

We performed additional simulations to compare the performances of DSS, DESeq and edgeR. The general simulation settings are as described in previous section. The only differences in simulations are the way to generate dispersions. We presented four additional simulations here. The gene-specific dispersions were generated from log-normal or Gamma distributions with different parameters. In all four simulations the average dispersions levels are roughly the same, but the variances of dispersions are different. The distributions used for generating dispersions are listed in Table S1.

	ϕ_g distribution	$E[\phi_g]$	$Var[\phi_g]$
Sim 1	$\log N(-1.5, 0.5^2)$	0.25	0.018
Sim 2	$\log N(-2.5, 1.5^2)$	0.25	0.51
Sim 3	$Gamma(5, 0.05)$	0.25	0.013
Sim 4	$Gamma(0.25, 1)$	0.25	0.25

Table S1: Distributions for ϕ_g in additional simulations.

The DE detection results for these simulations are shown in Figures S6 and S7. When the variation in dispersion is large, DSS provides better DE results than other methods.

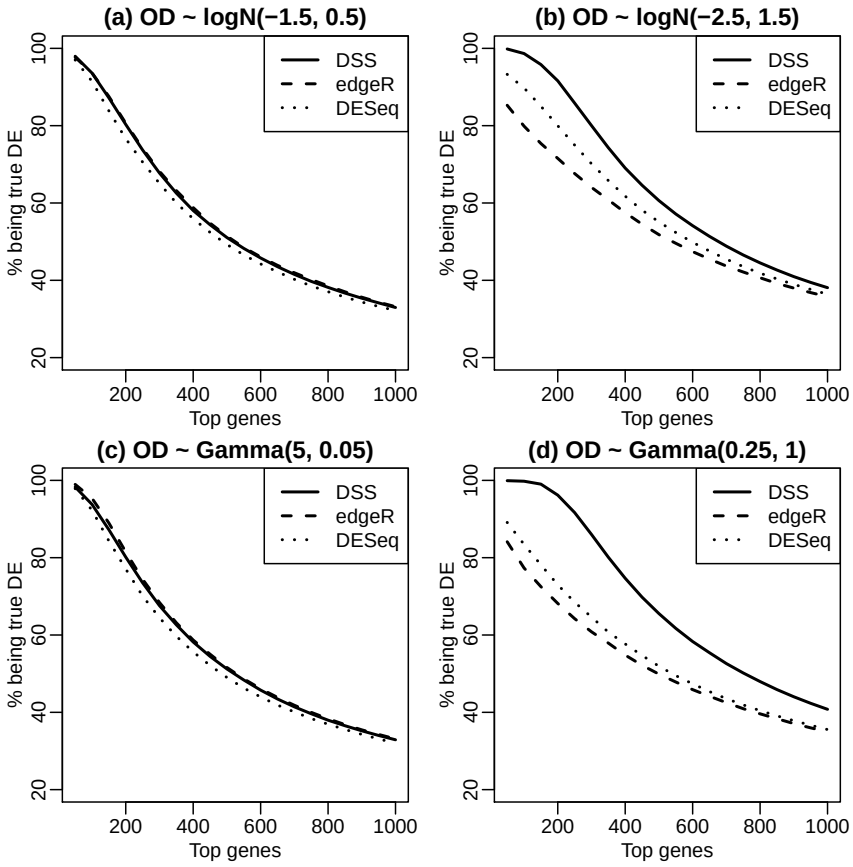

Figure S6: DE detection accuracies from additional simulations. ϕ_g were generated from log-normal (a and b) or Gamma distributions (c and d) with different parameters. The mean dispersions are roughly identical from all simulations, whereas the variance of dispersions are different. It can be seen that when the variation in dispersion is large DSS provides better performance. The curves are average from 50 simulations.

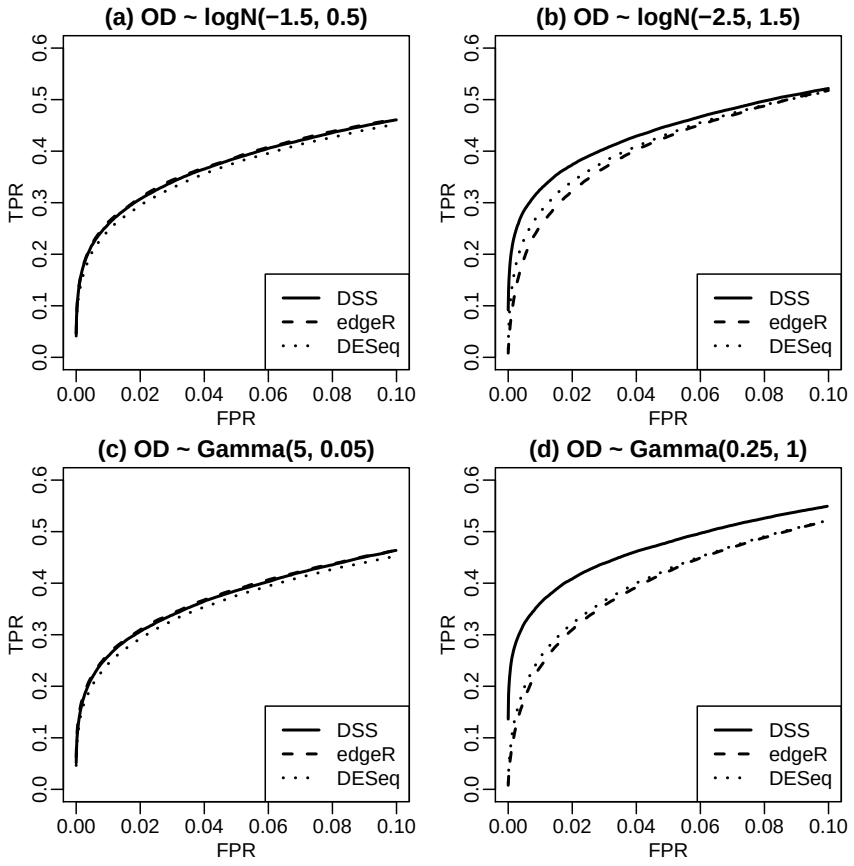

Figure S7: ROC curves from the same additional simulations as in Figure S6.

11 FDR estimation from other simulation settings

Figure S8 shows the comparison of true and reported FDR curves from simulations presented in the paper. In general DSS provides reasonable FDR estimations.

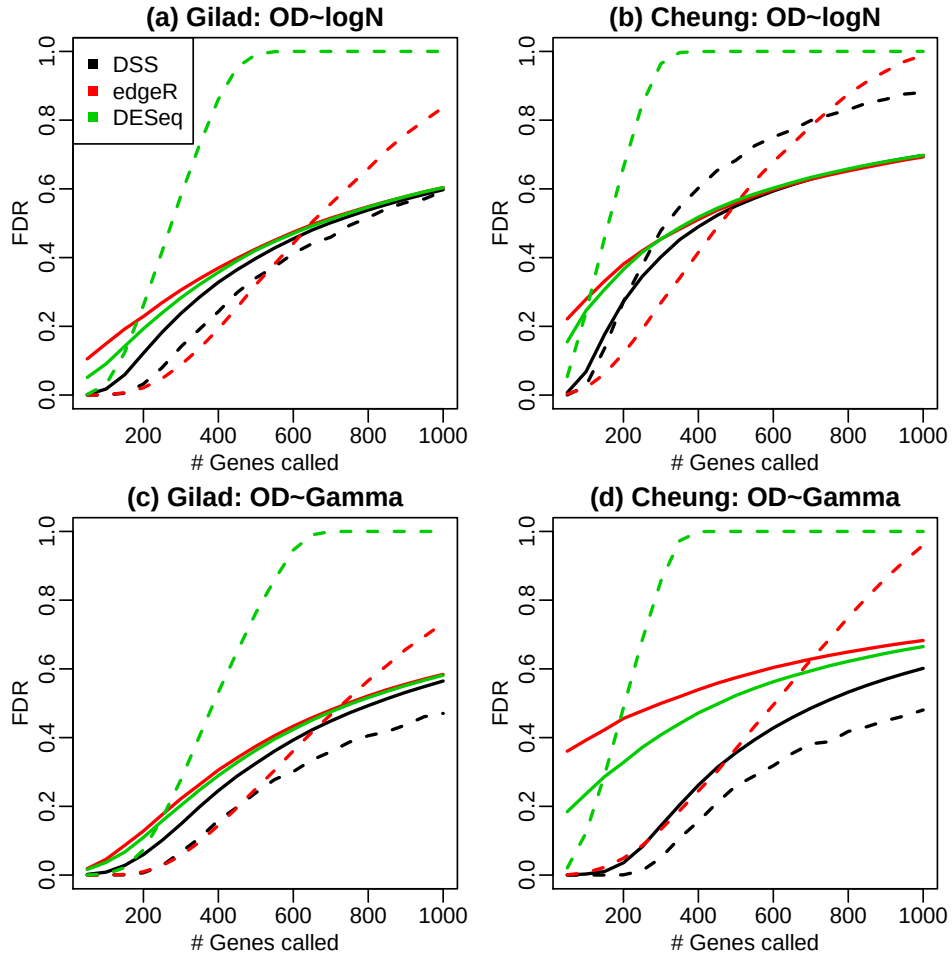

Figure S8: FDR estimation from different simulation settings. In all figures the solid curves represent the true FDR. The dashed curves represent the reported FDR. All curves are averaged from 50 simulations. Simulation settings are the same as described before. Dispersion are generated from log-normal distributions (Figures a and b) or Gamma (Figures c and d). Parameters for dispersion distributions were estimated from real data (Gilad for Figures a and c, Cheung for Figures b and d).

12 Comparison of Wald and exact tests

For two class comparison, DESeq and edgeR both provide the exact test functions: `nbinomTest (DESeq)` and `exactTest(edgeR)`. DSS uses Wald test. We compared the performance of Wald and the exact test under our simulation settings, and found that both tests provide comparable results. Figure S9 shows the comparisons of DE detection accuracies from the simulation based on the Cheung data, with dispersions from Gamma distribution. The left panel shows DE detection accuracies from each method, and the right panel presents the results using DSS's dispersion estimates in all tests. It shows that both Wald and exact tests provide almost identical results, indicating that most of the improvement in DE detection is due to better dispersion estimation.

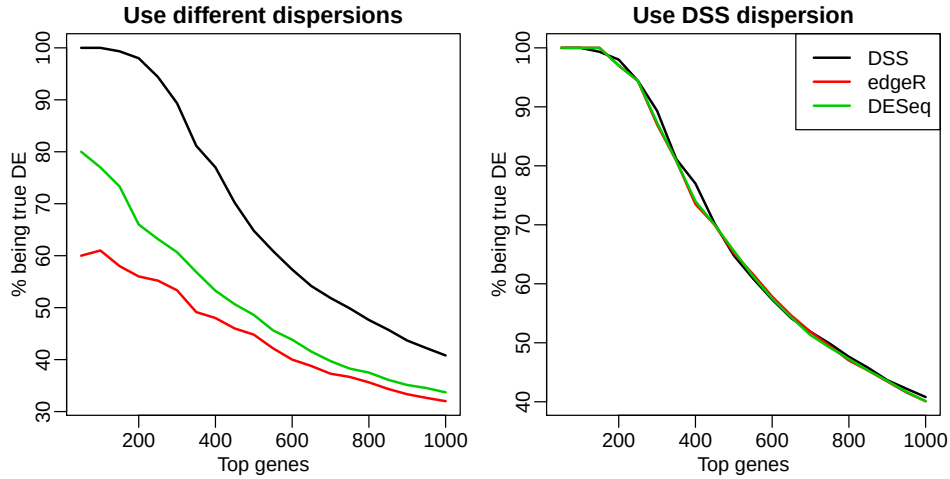

Figure S9: Comparison of Wald and exact tests from simulation. Data are simulated based on Cheung data, with dispersions from Gamma distribution. Left panel shows the DE detection accuracies from each method using their own dispersion estimates. Right panel shows results from using DSS's dispersion estimates.

13 Information of ϕ depends on the mean in NB distribution

To understand how much information about the dispersion parameter ϕ is available in the data with a small number of replicates, we simulated NB random variables with expectation ranging from 0.5 to 30 and ϕ at three levels. We use the ratio between sample variance and sample mean as an indicator of "observable overdispersion". The sampling distribution of this ratio is shown in boxplots. As expected, when $\phi = 0$ the ratio is centered around 1 for all mean levels. For NB random variable with mean as high as 30, the impact of ϕ on the ratio is obvious. For NB random variables with means less than 5, the ratio appear very similar whether there is overdispersion or not. In other words, the data provide little information about ϕ.

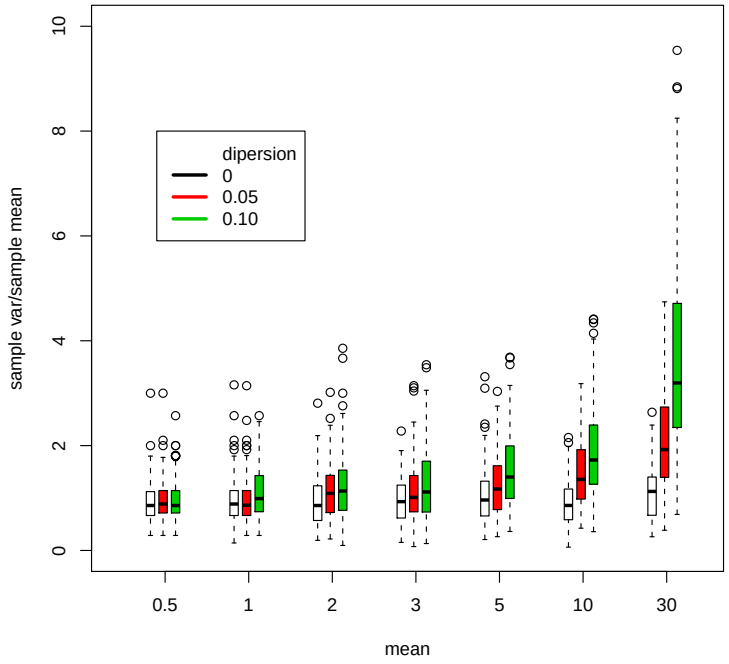

Figure S10: Boxplot of the ratio $\frac{\text{sample variance}}{\text{sample mean}}$ for NB random variables at various expectations and dispersions. Sample size is 8 and 100 samples taken at each parameter setting. When the dispersion parameter $\phi = 0.05$, the ratio $\frac{\text{sample variance}}{\text{sample mean}}$ in NB random variables may appear similar to Poisson random variables. Even when the dispersion parameter is $\phi = 0.10$, the NB random variable is not easily distinguishable from Poisson distribution with mean up to 5.

14 Simulation results based on MAQC data

We obtained another RNA-seq dataset generated by the MicroArray Quality Control Project [1] phase III, also known as Sequencing Quality Control (SEQC). In this experiment two biological samples, human brain and universal human reference sample (UHR), are assayed using seven lanes each. The MAQC data use samples from the same library preparation protocols so that there is little biological variation and the data can be considered as technical replicates.

We conducted simulations based on the MAQC data. The procedures are the same as in simulations based on other (Gilad and Cheung) data. Figure S11 compares the DE detection results. It shows that the three methods provide comparable results.

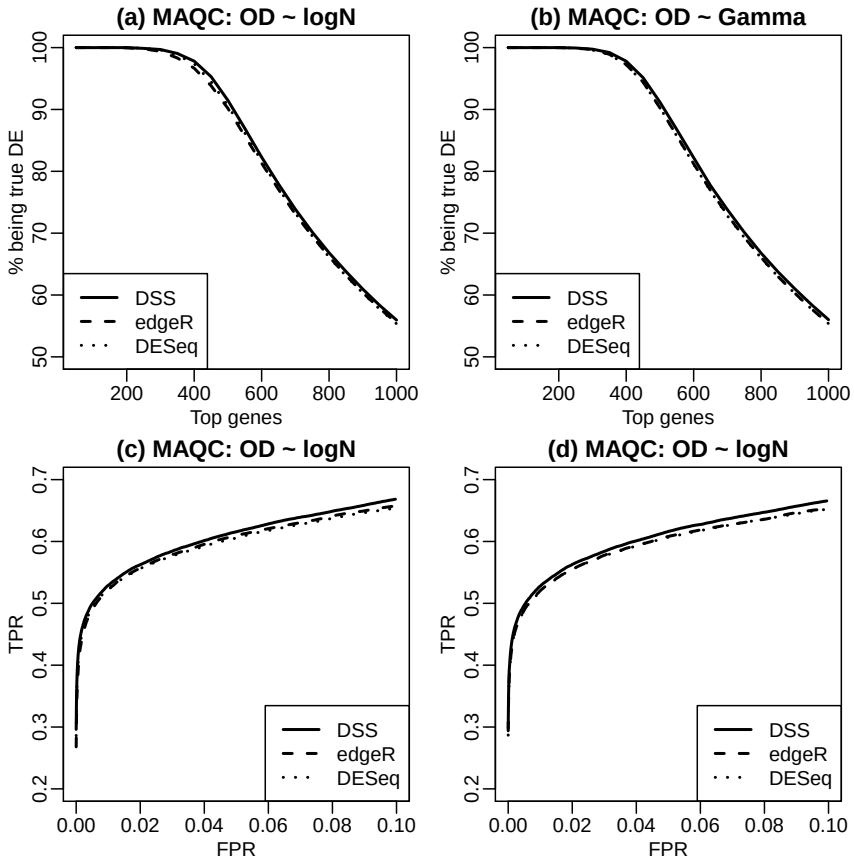

Figure S11: DE detection results from MAQC data. (a) and (b) show the percent of top ranked genes being true DE. (c) and (d) show the ROC curves. (a) and (c) assume the dispersions are from log-normal distribution. (b) and (d) assume the dispersions are from Gamma distribution.

15 DE detection from MAQC using qRT-PCR as gold standard

Quantitative real-time polymerase chain reaction (qRT-PCR) data are available for 992 genes from the MAQC experiment. These data were used as gold standard for comparing the DE detection results from different methods. To be specific, we first compute the log fold changes between the average expression measures for UHR and brain. Then a genes is called "DE" if the absolute value of log fold changes is greater than 2, or "non-DE" is the absolute value of log fold changes is less than 0.2. Under these criteria there are 312 DE genes and 159 non-DE genes. Using these as gold standard, the comparison of DE detections from DSS, edgeR and DESeq are shown in Figure S12. It can be seen that DSS provides better results.

However a deeper study of the results reveals that from all three methods, the estimated dispersions ($\tilde{\phi}_g$) are at the lower boundary for most genes. This is not surprising since the samples in MAQC data are technical replicates and possess very little biological variations. Thus the difference in DE detection results are in fact caused by different lower bound of $\tilde{\phi}_g$ used in different algorithm, (the lower bounds are 0.01 in DSS, 0.001 in edgeR and 10^{-8} in DESeq). Although these results cannot be interpreted as in favor of DSS, it does show that assuming a not-too-small minimum dispersion improves DE detection.

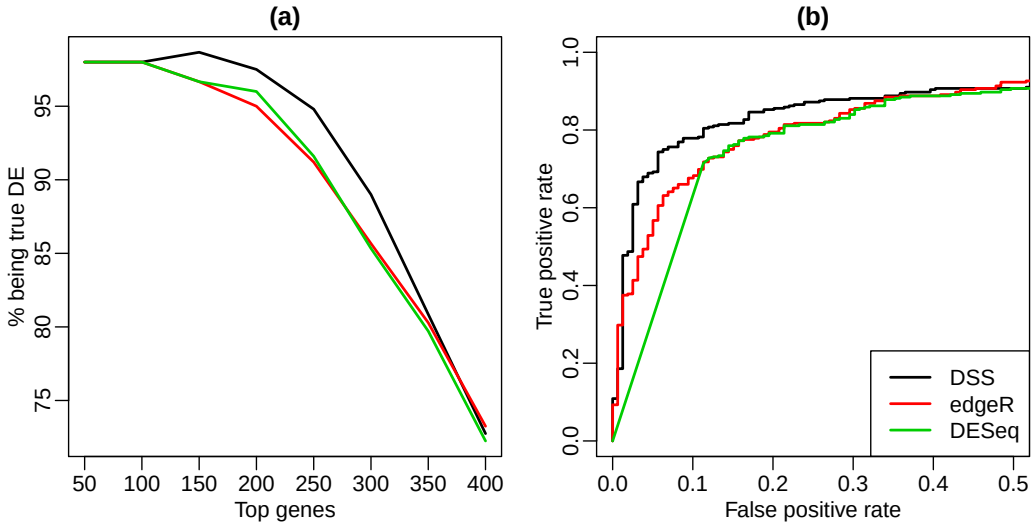

Figure S12: Comparison of DE detection results from MAQC data using Taqman qRT-PCR as gold standard. (a) ROC curve. (b) Percentage of top ranked genes being true DE.

16 Estimating Trended dispersion

There have been some reports that the dispersion may depend on the mean expression level. DESeq[2] models the variance as a smooth function of the mean. EdgeR [3, 4, 5] provides several methods to estimate the trend, using either a moving average smoother or locally weighted approximate conditional likelihood. The common feature is to treat the dispersion as a smooth function of the average expression. We account for the possible dependence between ϕ and mean expression by allowing the hyper-parameter m_0 to be a smooth function of μ_g. First we divide the genes into strata of average expression level and estimate a $\hat{m}_{0i}$ for the $i-th$ stratum non-parametrically. We then use a smoothing spline to estimate the relationship $\hat{m}_0(\mu_g)$ and let each gene have its own prior expectation of dispersion.

We illustrate the trended estimation in Figure S13 in a simulation of 5 vs 5 comparison A linear trend between $\log(\phi)$ and log mean expression is imposed in the simulation, as shown in red. The edgeR estimates are shrunk heavily towards the mean (Figure S13B) and the DEseq essentially sets a lower bound for ϕ (Figure S13C). The DSS estimates reflect the trend and shrink genes with low counts more heavily. Figure S14 show the comparison between estimated and true dispersion in the simulation. As seen in Figure 1 in the main text, the estimates from DSS show better concordance with the truth.

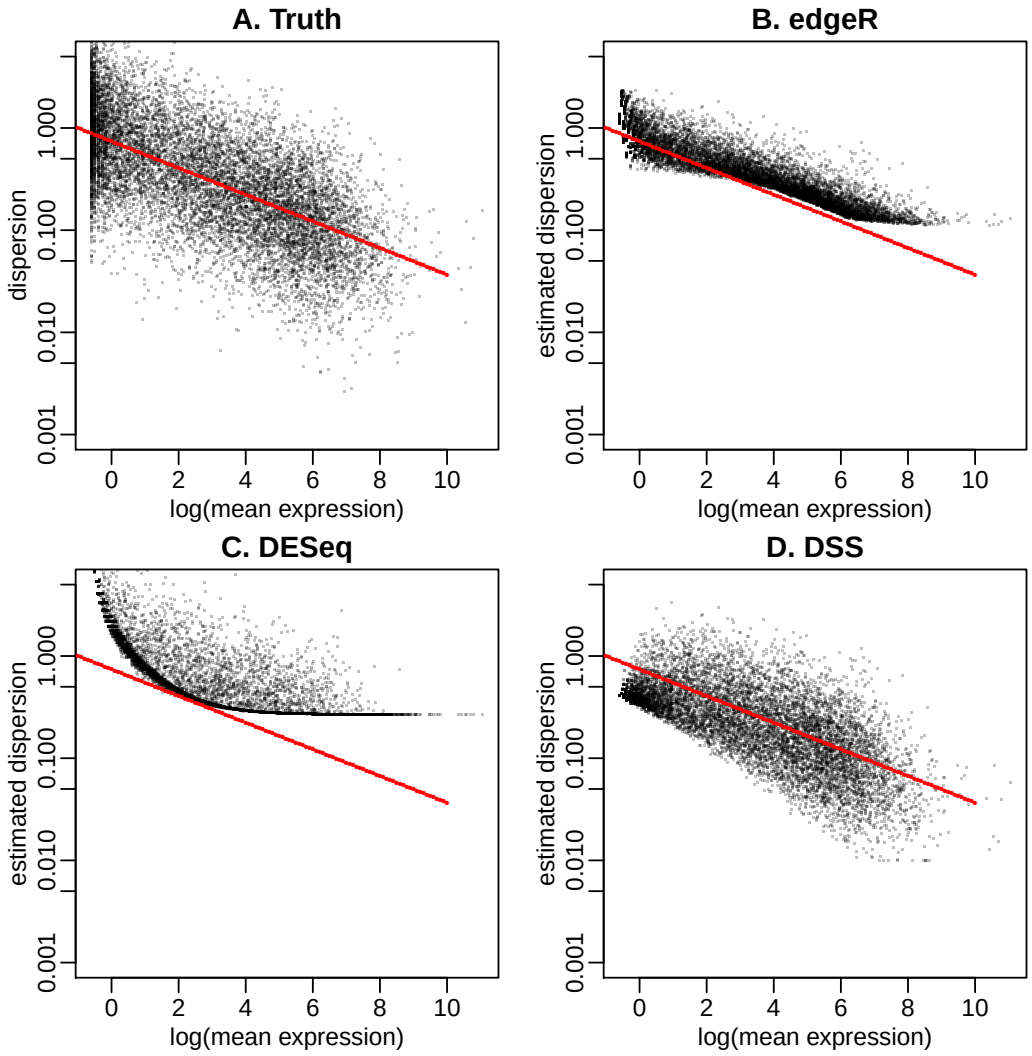

Figure S13: Mean-dependent dispersion. A. True dispersion ϕ plotted against true log expression used in simulation. The simulation includes 52580 genes (as seen in reCount human RNAseq data) and 10 samples, and mean expression level are simulated based on the Cheung data. A linear trend between $\log(\phi)$ and log mean expression μ is imposed in the simulation, shown as the red line in all panels. Specifically, $\log(\phi)|\mu \sim N(-0.3\mu - 0.3, 1)$. Genes with all zero counts are excluded in analysis. B. Estimated dispersion from edgeR versus mean expression. C. Estimated dispersion from DESeq versus mean expression. D. Estimated dispersion from DSS versus mean expression.

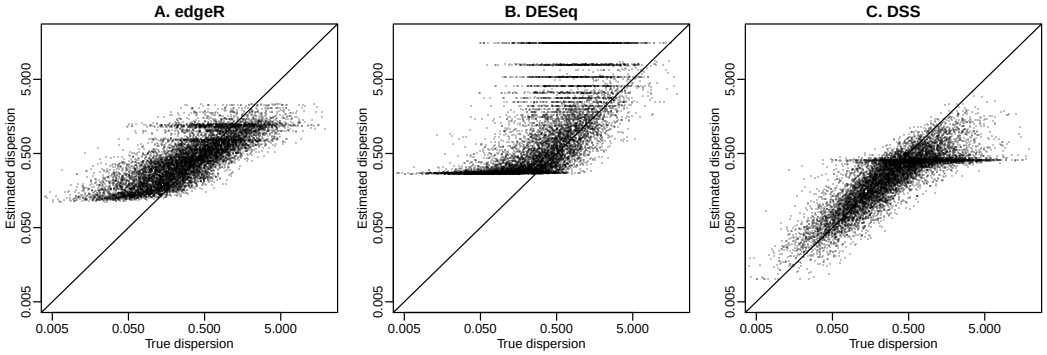

Figure S14: Mean-dependent dispersion estimation. Estimated dispersion plotted against true dispersion for A. edgeR, B. DESeq and C. DSS. The results are from the simulation as described in Figure S13.

References

[1] Shi, L., Reid, L.H., Jones, W.D., Shippy, R., Warrington, J.A., Baker, S.C., Collins, P.J., De Longueville, F., Kawasaki, E.S., Lee, K.Y. *and others*. (2006). The microarray quality control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature biotechnology* **24**(9), 1151–1161.

[2] Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol* **11**(10), R106.

[3] Robinson, Mark D, McCarthy, Davis J and Smyth, Gordon K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression dataSMotn. *Bioinformatics* **26**(1), 139–140.

[4] Robinson, Mark D and Smyth, Gordon K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* **9**(2), 321–332.

[5] 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 Research* **doi:10.1093/nar/gks042**.