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

Supplementary Materials

Danni Yu\textsuperscript{1,2}, Wolfgang Huber\textsuperscript{1}, Olga Vitek\textsuperscript{2,3}
\textsuperscript{1}European Molecular Biology Laboratory, Mayerhofstraße 1, Heidelberg, Germany
\textsuperscript{2}Department of Statistics, Purdue University, West Lafayette, IN USA
\textsuperscript{3}Department of Computer Science, Purdue University, West Lafayette, IN USA

February 24, 2013

Contents

1 Background  2
  1.1 The Negative Binomial distribution  2
  1.2 Sum of Negative Binomial random variables  2
  1.3 Connection to other distributions  2

2 Methods  4
  2.1 The model  4
  2.2 Exact test for a two-group comparison  5
  2.3 Exact test for complex experiments  5

3 Datasets  7

4 Results  9
  4.1 Use of the existing approaches  9
  4.2 sSeq yields accurate estimates of variation  10
  4.3 sSeq accurately detects differential expression  15
  4.4 Effect of sample size  19
  4.5 Effect of size factors  20
  4.6 Representative R scripts (Sultan dataset)  23

1
1 Background

1.1 The Negative Binomial distribution

A Negative Binomial random variable $X \sim NB(p, r)$ counts the number of failures before the $r$th success in a series of independent and identical Bernoulli trials with probability of success $p$. Its probability distribution is

\[
P\{X = x \mid p, r\} = \left(\frac{r + x - 1}{x}\right)(1 - p)^x p^r \text{ for integer } x \geq 0 \text{ and } p \in (0, 1),
\]

\[
E\{X\} = r \frac{1 - p}{p}, \text{ and } Var\{X\} = r \frac{1 - p}{p^2}
\]

An alternative parametrization used in this manuscript is $X \sim NB(\mu, \phi)$ such that

\[
P\{X = x \mid \mu, \phi\} = \left(\frac{1}{\phi} + x - 1\right) \left(\frac{\mu\phi}{\mu\phi + 1}\right)^x \left(\frac{1}{\mu\phi + 1}\right)^\frac{1}{\phi},
\]

\[
E\{X\} = \mu, \text{ and } Var\{X\} = \mu + \mu^2 \phi
\]

In this parametrization

\[
\mu = r \frac{1 - p}{p} \text{ and } \phi = \frac{1}{r}; \text{ or equivalently } p = \frac{1}{\mu\phi + 1}; \text{ } r = \frac{1}{\phi}
\]

1.2 Sum of Negative Binomial random variables

A property of the Negative Binomial distribution is that if $X_j \sim NB(p, r_j), j = 1, \ldots, n$ and are independent, then $\sum_{j=1}^{n} X_j \sim NB(p, \sum_{j=1}^{n} r_j)$.

1.3 Connection to other distributions

The Negative Binomial distribution generalizes the Poisson distribution for count data. If we assume that $X \sim Poisson(\lambda)$, and the expected value of the Poisson distribution is itself a random variable $\lambda \sim Gamma(\phi^{-1}, \mu_\phi)$, $\mu_\phi, \phi_\phi > 0$, then the marginal distribution of $X$ (on average over all possible expected values $\lambda$) is

\[
P\{X = x \mid \mu, \phi\} = \int_{0}^{\infty} P(X \mid \lambda) \cdot f(\lambda \mid \mu, \phi) \, d\lambda = \int_{0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} \cdot \frac{\lambda^{\phi^{-1} - 1} e^{-\lambda} \mu^{-1} \phi^{-1}}{(\mu\phi)^{\phi^{-1} - 1} \Gamma(\phi^{-1})} \, d\lambda
\]

\[
= \frac{\Gamma(x + \frac{1}{\phi})}{x! \Gamma(\frac{1}{\phi})} \left(\frac{\mu\phi}{\mu\phi + 1}\right)^x \left(\frac{1}{\mu\phi + 1}\right)^\frac{1}{\phi}
\]

This connection motivates the use of the Negative Binomial distribution in RNA-seq experiments. Marioni et al. [8] demonstrated that the Poisson distribution adequately represents the technical variation of the counts in replicated libraries of a same biological sample. The Negative Binomial distribution allows us to explicitly model the combined effect of the biological and technical variation, as it inflates the total variation beyond what is specified by the Poisson distribution.
The Poisson-Gamma model is not the only motivation for modeling counts of RNA-seq reads with the Negative Binomial distribution. The Negative Binomial distribution can also arise as the sum of Geometric distributions, i.e. if \( X_j \overset{iid}{\sim} \text{Geometric}([1 + (\mu \phi)^{-1}]^{-1}), j = 1, \ldots, \phi^{-1}, \) then \( \sum_{j=1}^{\phi^{-1}} X_j \sim \text{NB}(\mu, \phi). \) Alternatively, the Negative Binomial distribution can be represented as a compound Poisson distribution, i.e. if \( X_j \overset{iid}{\sim} \text{Logarithmic}\left(\frac{1}{1+(\mu \phi)^{-1}}\right), j = 1, \ldots, J, \) and \( J \sim \text{Poisson}\left(-\phi^{-1}\log\frac{1}{1+\mu \phi}\right), \) then \( \sum_{j=1}^{J} X_j \sim \text{NB}(\mu, \phi). \)
2 Methods

2.1 The model

Denote \( X_{gij} \) the counts of reads of gene \( g = 1, \ldots, G \), replicate \( j = 1, \ldots, n_i \) and condition \( i = A, B \). Denote \( s_{ij} \) the size factor of the replicate \( j \) in the condition \( i \). The probability model is

\[
X_{gij} \sim \text{NB}(\mu_{gi}, s_{ij}/\phi_g), \quad \text{where} \quad \mu_{gi} \geq 0, \quad \phi_g \geq 0, \quad s_{ij} > 0 \quad \text{such that}
\]

\[
E\{X_{gij}\} = \mu_{gi} s_{ij}, \quad \text{and} \quad \text{Var}\{X_{gij}\} = \mu_{gi} s_{ij} + \mu_{gi}^2 \phi_g s_{ij} \tag{5}
\]

This model is not the only possibility to represent RNA-seq experiments, however it is quite flexible. First, the free per-gene dispersion parameter \( \phi_g \) accommodates arbitrary dependencies of dispersion on the expected value, and is particularly useful in experiments with a small sample size where the true relationship may be obscured by the noise.

Second, the assumption regarding the size factors can be meaningful from the experimental viewpoint, and also for technical modeling reasons. From the experimental viewpoint, Eq. (5) shows that size factors \( s_{ij} \) linearly scale both the expected value of the counts of reads and their variance. Since the differences in library size are technical artifacts, and since the technical variation in RNA-seq experiments can be characterized with the Poisson distribution, the linear scaling of the variance with the library size is consistent with the Poisson\(^1\). Moreover, this assumption enables us to use the Negative Binomial distribution to directly model both the counts of reads in one replicate library (necessary to introduce the size factors), and the sum of the counts in the replicate libraries of a condition (necessary to conduct the exact test). As detailed in Sec. 1.2, a traditional interpretation of a Negative Binomial random variable is ‘the number of failures before the \( r \)th success in Bernouilli trials with probability of success \( p' \). In this traditional parametrization

\[
X_{gij} \sim \text{NB}(p_{gi}, r_g), \quad j = 1, \ldots, n_i, \quad \text{such that} \quad p_{gi} = \frac{1}{\mu_{gi} \phi_g + 1} \quad \text{and} \quad r_g = \frac{1}{\phi_g}
\]

Introducing the size factor as in Eq. (5) implies a constant \( p \), i.e.

\[
p_{gi} = \frac{1}{s_{ij} \mu_{gi} \phi_g + 1} = \frac{1}{\mu_{gi} \phi_g + 1}, \quad r_{ij} = \frac{s_{ij}}{\phi_g} \tag{6}
\]

Therefore, the size factors can be interpreted as scaling the required number of successes in the Bernouilli trials while fixing the probability of success. Then the sum of the counts

\[
\sum_{j=1}^{n_i} X_{gij} \sim \text{NB} \left( p_{gi}, \frac{\sum_{j=1}^{n_i} s_{ij}}{\phi_g} \right), \quad E\{\sum_{j=1}^{n_i} X_{gij}\} = \mu_{gi} \sum_{j=1}^{n_i} s_{ij}, \quad \text{Var}\{\sum_{j=1}^{n_i} X_{gij}\} = \mu_{gi} \sum_{j=1}^{n_i} s_{ij} + \mu_{gi}^2 \sum_{j=1}^{n_i} s_{ij} \phi_g \tag{7}
\]

\(^1\)Alternative models of the scaling factors, such as \( X_{gij} \sim \text{NB}(\mu_{gi}, s_{ij}, \phi_g) \), assume extra-Poisson scaling of the variance, i.e. \( \text{Var}\{X_{gij}\} = \mu_{gi} s_{ij} + \mu_{gi}^2 \phi_g s_{ij}^2 \). In experiments with small sample size it may be difficult to evaluate which model fits best, and the two approaches are similar when the size factors are close to 1.
2.2 Exact test for a two-group comparison

The null hypothesis is \( H_0 : \mu_{gA} = \mu_{gB} \). The test statistics are the combined counts of reads in each condition \( X_{gA} = \sum_{j=1}^{n_A} X_{gij} \) and \( X_{gB} = \sum_{j=1}^{n_B} X_{gij} \). According to Eq. (7),

\[
X_{gA} \sim NB(s_A \cdot \mu_{gA}, \phi_g/s_A), \quad \text{and} \quad X_{gB} \sim NB(s_B \cdot \mu_{gB}, \phi_g/s_B)
\]

where \( s_A = \sum_{j=1}^{n_A} s_{Aj} \) and \( s_B = \sum_{j=1}^{n_B} s_{Bj} \). Under \( H_0 \),

\[
X_{gA} \sim NB(s_A \cdot \mu_g, \phi_g/s_A), \quad \text{and} \quad X_{gB} \sim NB(s_B \cdot \mu_g, \phi_g/s_B)
\]

Since the counts from the two conditions are independent, the joint probability distribution of \( (X_{gA}, X_{gB}) \) is

\[
P\{X_{gA} = x_{gA}, X_{gB} = x_{gB} | H_0 \} = P\{X_{gA} = x_{gA} | H_0 \} \cdot P\{X_{gB} = x_{gB} | H_0 \}
\]

The p-value of the exact test is the combined probability of all the read counts per group, such that under \( H_0 \) they have a same or a lower probability than the counts observed, conditional on the total counts equal to the observed. Mathematically,

\[
p\text{-value}_g = P_{2,g}/P_{1,g}
\]

where

\[
P_{1,g} = \sum_{x_{gA}, x_{gB} \in \text{set1}} P\{X_{gA} = x_{gA} | H_0 \} \cdot P\{X_{gB} = x_{gB} | H_0 \}
\]

\[
P_{2,g} = \sum_{x_{gA}, x_{gB} \in \text{set2}} P\{X_{gA} = x_{gA} | H_0 \} \cdot P\{X_{gB} = x_{gB} | H_0 \}
\]

\[
\text{set1} = \{X_{gA}, X_{gB} | X_{gA} + X_{gB} = x_{gA} + x_{gB} \}
\]

\[
\text{set2} = \{X_{gA}, X_{gB} | X_{gA} + X_{gB} = x_{gA} + x_{gB} \ \cap \ \sum_{j=1}^{n_A} \hat{s}_{Aj} \leq \sum_{j=1}^{n_B} \hat{s}_{Bj} \}
\]

In practice the probabilities in Eq. (11) are calculated by substituting \( \hat{\mu}_g, \hat{\phi}_g, \hat{s}_{Seq}, \hat{s}_A = \sum_{j=1}^{n_A} \hat{s}_{Aj} \) and \( \hat{s}_B = \sum_{j=1}^{n_B} \hat{s}_{Bj} \) into the probability distributions in Eq. (9). The p-values are adjusted by method such as in [2] to control the False Discovery Rate.

2.3 Exact test for complex experiments

We consider two types of experimental designs that are more complex than a two-group comparison, and focus on pairwise comparisons of conditions.

**Factorial experiments:** The experiment in Hammer et al. [5] had a factorial design. It considered two factors (rat strains Sprague Dawley and L5 SNL Sprague Dawley 2, and time points 2 weeks and 2 months), and considered distinct biological replicates for each combination of strain and time.

Pairwise comparisons of conditions in such experiments is straightforward. We created a new condition with fours levels (Sprague Dawley, 2 weeks; Sprague Dawley, 2 months; L5 SNL Sprague Dawley 2, 2 weeks’ L5 SNL Sprague Dawley 2, 2 months). The null hypotheses comparing pairs of conditions \( H_0 : \mu_{gi} = \mu_{gi'}, i, i' \in \{1, 2, 3, 4\}, i \neq i' \), were then tested as in Sec. 2.2 while the dispersion parameter were estimated using read counts from all the conditions.
**Repeated measurements:**  The experiment in Tuch *et al.* [9] had a paired design, in that pairs of normal and tumor samples were obtained from each of the three patients. This is a special case of the repeated measurements design. The design is advantageous because it eliminates the between-subject biological variation from consideration when comparing the conditions.

To analyze such experiments we propose to view each subject as a separate independent unreplicated experiment. Since sSeq can handle unreplicated experiments, we derived the estimates of dispersion $\hat{\phi}_{sSeq}^g$ separately for each subject, thus reflecting the within-subject variation. Next we tested $H_0$ separately for each subject, obtaining separate p-values for each subject and each gene. Finally, assuming that the subjects are independent, we combined the p-values for each gene using Fisher’s method [7] and obtained the consensus p-values.

There are at least two ways to obtain $\hat{\phi}_{sSeq}^g$. The first is to shrink the method of moment estimates $\hat{\phi}_{MM}^g$ separately for each subject across conditions. The second is to average the per-subject method of moments estimates $\hat{\phi}_{MM}^g$, and proceed with a single shrinkage step of the averaged estimates. We compared the sensitivity of these two approach for the Tuch dataset. Fig. 9 in Sec. 4.3 shows that shrinking averaged estimates of dispersions resulted in a higher sensitivity. At the same time, both approaches are less sensitive than the analysis that ignores the paired nature of the design. The loss of sensitivity is possibly due to inefficiencies of the Fisher’s methods, and several alternatives can potentially be considered [7].

Fig. 2 in the main manuscript is based on the shrinkage of the averaged method of moment estimates. The figure shows that despite the loss of efficiency the method has a same or a better accuracy of detecting differentially expressed genes as compared to the GLM-based approaches.
3 Datasets

Simulation1, Simulation2 and Simulation3 each generated \( G = 20,000 \) genes in conditions \( A \) and \( B \), \( n_A = n_B = 2 \). Parameters \( \mu_gA \) and \( \phi_g \) were simulation-specific (see below). 30\% of the genes were simulated as differentially expressed, and for these genes \( \mu_gB = \mu_g / \exp(\epsilon) \) where \( \epsilon \sim N(0.5, 0.25^2) \). Size factors were sampled from the Uniform distribution \( s_{ij} \sim \text{Uniform}(0.5, 1.7) \), and rounded to the first decimal place. The simulated size factors are reported in Supplementary Section Sec. 4.5. Read counts for gene \( g \) in sample \( j \) (\( j = 1, 2 \)) and condition \( i \) (\( i = A, B \)) were randomly generated from \( NB(\mu_{gi} \cdot s_{ij}, \phi_g / s_{ij}) \).

Simulation1 The expected values are randomly sampled from \( \mu_g \sim \text{Exponential}(\lambda = 250) \), where \( \lambda \) is the expected value. The dispersion parameter is considered as a constant across genes, \( \phi_g = 0.1 \).

Simulation2 As above, the expected values are randomly sampled from \( \mu_g \sim \text{Exponential}(\lambda = 250) \), where \( \lambda \) is the expected value. The dispersion parameters are functions of the expected values \( \phi_g = 1/(100 + \mu_g) \). This setting is the same as in [1].

Simulation3 From the dataset by Bottomly et al [3, 4], the largest experimental dataset in this manuscript, we selected a subset of non-differentially expressed genes (as determined by a consensus of sSeq, edgeR and DESeq), and sampled pairs \((\hat{\mu}_gA, \hat{\phi}_g)\) from this subset as the true parameters.

MAQC The read counts for the RNA-Seq experiment were downloaded from \texttt{http://www.ncbi.nlm.nih.gov} accession number SRA010153. The human genome \texttt{hg19} was downloaded from \texttt{http://genome.ucsc.edu}. The reads from each library were mapped to the human genome using Bowtie \texttt{bowtie -q -v 2 -a -m 1 -p 8 --quiet hg19 input.fastq output.map} This resulted in 19,580 mapped genes that were sequenced in at least one of the 5 libraries.

To obtain the external ‘gold standard’ of differential expression, the qRT-PCR quantifications were downloaded from Gene Expression Omnibus (GEO, accession GSE5350). We compared the quantifications in the two conditions with the t-test. A gene was termed as differentially expressed if its p-value was less than 0.00001 (statistical significance), and the absolute fold change exceeded 2.1 (practical significance). A gene was termed non-differentially expressed if its p-value exceeded 0.2 (statistical significance), and absolute fold change was less than 1.5 (practical significance). This produced 323 differential genes and 85 non-differentially expressed genes.

Griffith The counts of aligned RNA-seq reads were downloaded from GEO (accession GSE23776). The qPCR data were downloaded from \texttt{http://www.alexaplatform.org/alexa_seq/index.html}. As above, we compared the conditions with the t-test. For this dataset a gene was considered as truly differentially expressed if its p-value was less than 0.00001 (statistical significance) and the absolute fold change exceeded 3 (practical significance), and truly non-differentially expressed if its p-value exceeded 0.2 (statistical significance) and absolute fold change was less than 0.9 (practical significance). This produced 12 differentially expressed genes and 19 non-differentially expressed genes.
Brooks  The table of read counts was downloaded from the R package pasilla.

Sultan, Bottomly and Hammer  The read counts for these dataset were downloaded from [http://bowtie-bio.sourceforge.net/recount](http://bowtie-bio.sourceforge.net/recount) [4].

Dataset by Tuch et al.  The table of read counts was downloaded from GEO (accession GSE20116).
4 Results

4.1 Use of the existing approaches

**edgeR** v3.0.8 (January 2013) under R v2.15.2, Mac OSX. The default parameters were used for the replicated experiments. Since this version discourages the analysis of unreplicated experiments, the `estimateCommonDisp` function in an older version of edgeR package (v2.4.6) was used to automatically analyze the datasets with $n_i = 1$. Pairwise comparisons of conditions were conducted with the exact test for the datasets with a two-group design. GLM was used for the datasets with more complex designs.

**DESeq** v1.10.1 (October 2012) under R v2.15.2, Mac OSX. For the datasets with multiple replicates per-condition dispersion was assumed. For the unreplicated datasets the dispersion was assumed common between conditions. As with edgeR, exact test was used for the datasets with a two-group design, and GLM was used for the datasets with more complex designs.

**baySeq** v1.12.0 (October 2012) under R v2.15.2, Mac OSX. The default parameters were used.

**BBSeq** v1.0 (March 2011) under R v2.12.0, Windows. The package is not available in Mac OSX. It could not be installed under more recent versions of R due to a missing “NAMESPACE” file. The default parameters were used.

**SAMSeq** is part of the R package **samr** v2.0 (June 2011) under R v2.15.2, Mac OSX. The default parameters were used. The method is not applicable to unreplicated experiments.
4.2 sSeq yields accurate estimates of variation

Fig. 1 illustrates the estimation of variances and of dispersions for the experimental dataset by Hammer et al. Although the true values of the dispersions are unknown, qualitatively the patterns are similar to the patterns in Fig. 1 of the main manuscript.

(a) Hammer, \( n_i = 2 \)  
(b) Hammer, \( n_i = 2 \)

(c) Hammer, \( n_i = 2 \): Proposed  
(d) Hammer, \( n_i = 2 \): Proposed

(e) Hammer, \( n_i = 2 \): edgeR  
(f) Hammer, \( n_i = 2 \): edgeR

(g) Hammer, \( n_i = 2 \): DESeq  
(h) Hammer, \( n_i = 2 \): DESeq

Figure 1: Dispersion and variance estimation for the Hammer dataset with a factorial design. The figures compare two strains at the time point 2 months. The dispersion is estimated from all the four conditions (i.e. two strains and two time points). (a) Average squared difference (ASD) versus shrinkage target \( \xi \). ASD is maximized at \( \xi = \hat{\phi}^{MM} \) (solid horizontal line). The dashed lines are the selected target \( \hat{\xi} \) and its ASD. (b) The proposed shrinkage estimator is a linear transformation of \( \hat{\phi}^{MM} \), with the slope \( 1 - \delta = 0.78 \) and the fixed point \( \hat{\xi} = 0.605 \). All \( \hat{\phi}^{MM} \leq 0 \) are transformed to \( \delta \xi = 0.13 \). (c),(e) and (g) Dispersion estimates by sSeq, edgeR and DESeq, versus the per-gene mean read counts across the two conditions. Gray smooth scatter indicates the estimates \( \hat{\phi}^{MM} \) (same on all the plots). Black dots are the estimates by each method. Gray lines indicate the true dispersion parameters. (d),(f) and (h) Same as above, but for the variances of the read counts.
Fig. 2 illustrates the estimation of variances and of dispersions for Simulation2. As discussed in the main manuscript, \( \hat{\phi}_{sSeq} \) are biased towards larger values. Since the proposed approach starts by pooling the read counts across conditions to obtain the initial method of moment estimates of dispersion \( \hat{\phi}_{MM} \), the estimates for differentially expressed genes with large counts tend to be high (and form a separate upper cloud in the gray scatter).

Table 1 in the main manuscript shows that the bias does not have a negative effect on accurate detection of differentially expressed genes in this simulation. This is due to the fact that for genes with low mean counts of reads all the methods yield similar estimates of variation. For genes with high read counts, the change in mean counts dominates the near-Poisson variation. The somewhat sub-optimal performance of edgeR for this simulation (reported in Table 2 of the main manuscript) is possibly due to approximations at the pseudocount calculation step.

(a) Simulation2, \( n_i=2 \)  
(b) Simulation2, \( n_i=2 \)  
(c) Simulation2, \( n_i=2 \): Proposed  
(d) Simulation2, \( n_i=2 \): Proposed  
(e) Simulation2, \( n_i=2 \): edgeR  
(f) Simulation2, \( n_i=2 \): edgeR  
(g) Simulation2, \( n_i=2 \): DESeq  
(h) Simulation2, \( n_i=2 \): DESeq

Figure 2: Dispersion and variance estimation for the Simulation2. The true dispersion is a function of the true mean \( \phi_g = 1/(100 + \mu_g) \). (a) Average squared difference (ASD) versus shrinkage target \( \xi \). ASD is maximized at \( \xi = \hat{\phi}_{MM} \) (solid horizontal line). The dashed lines are the selected target \( \hat{\xi} \) and its ASD. (b) The proposed shrinkage estimator is a linear transformation of \( \hat{\phi}_{MM} \), with the slope \( (1-\delta) = 0.69 \) and the fixed point \( \hat{\xi} = 0.345 \). All \( \hat{\phi}_{MM} \leq 0 \) are transformed to \( \delta \xi = 0.11 \). (c),(e) and (g) Dispersion estimates by sSeq, edgeR and DESeq, versus the per-gene mean read counts across the two conditions. Gray smooth scatter indicates the estimates \( \hat{\phi}_{MM} \) (same on all the plots). Black dots are the estimates by each method. Gray lines indicate the true dispersion parameters. (d),(f) and (h) Same as above, but for the variances of the read counts.

11
Fig. 3 illustrates the estimation of variances and of dispersions for Simulation3. Similar conclusions can be made from this figure.

For the remaining experimental datasets, the patterns of the estimated dispersion parameters as function of the means are illustrated in Fig. 4 and the patterns of the estimated variances as function of the means in Fig. 5.

(a) Simulation3, \( n_i = 2 \)  
(b) Simulation3, \( n_i = 2 \)

(c) Simulation3, \( n_i = 2 \): Proposed  
(d) Simulation3, \( n_i = 2 \): Proposed

(e) Simulation3, \( n_i = 2 \): edgeR  
(f) Simulation3, \( n_i = 2 \): edgeR

(g) Simulation3, \( n_i = 2 \): DESeq  
(h) Simulation3, \( n_i = 2 \): DESeq

Figure 3: Dispersion and variance estimation for the Simulation3 dataset that is close to the real-world data. The pairs of true means and true dispersions are generated from the non-differential genes in Bottomly et al. data. (a) Average squared difference (ASD) versus shrinkage target \( \xi \). ASD is maximized at \( \xi = \hat{\phi}^{MM} \) (solid horizontal line). The dashed lines are the selected target \( \hat{\xi} \) and its ASD. (b) The proposed shrinkage estimator is a linear transformation of \( \hat{\phi}^{MM} \). (c),(e) and (g) Dispersion estimates by sSeq, edgeR and DESeq, versus the per-gene mean read counts across the two conditions. Gray smooth scatter indicates the estimates \( \hat{\phi}^{MM} \) (same on all the plots). Black dots are the estimates by each method. Blue dots are the true dispersion parameters. (d),(f) and (h) Same as above, but for the variances of the read counts.
Figure 4: Dispersion estimates by sSeq, edgeR and DESeq, versus the per-gene mean read counts across the two conditions. Gray smooth scatter indicates the estimates $\hat{\phi}^{MM}$ (same on all the plots). Black dots are the estimates by each method. The dataset by Hammer et al. is in Fig. 1.
(a) Griffith
\[ n_A = n_B = 1 \]

(b) Bullard
\[ n_A = 3, \ n_B = 2 \]

(c) Brook
\[ n_A = n_B = 2 \]

(d) Sultan
\[ n_A = n_B = 2 \]

(e) Bottomly
\[ n_A = 10, \ n_B = 11 \]

(f) Tuch
\[ n_A = n_B = 3 \]

Figure 5: Variance estimates by sSeq, edgeR and DESeq, versus the per-gene mean read counts across the two conditions. Gray smooth scatter indicates the estimates \( \hat{\phi}_{MM} \) (same on all the plots). Black dots are the estimates by each method. The dataset by Hammer et al. is in Fig. [1].
4.3 sSeq accurately detects differential expression

Figure 6: Areas under the ROC curves of detecting differentially expressed genes for the simulated datasets in Table 2 of the main manuscript.
Figure 7: Areas under the ROC curves of detecting differentially expressed genes for the experimental datasets with an external ‘gold standard’ in Table 2 of the main manuscript.
Table I and Fig. 8 show the details of detecting differentially expressed genes while shrinking the method of moments estimates of variance, as opposed to the proposed shrinkage of dispersion. They illustrate that shrinking the variance undermines the accuracy of the results.

| Simulation1 | Simulation2 | Simulation3 | MAQC Project | Griffith et al. |
|-------------|-------------|-------------|--------------|----------------|
| n = 1       | n = 2       | n = 1       | n = 2        | n = 1          |
| 0.605       | 0.863       | 0.654       | 0.885        | 0.602          |

Table 1: Areas under the ROC curves of detecting differentially expressed genes for the datasets with external 'gold standard', while varying the FDR-adjusted p-value or posterior probability cutoff, obtained with the shrinkage of variance as opposed to the proposed shrinkage of dispersion. Sub-columns are subsets of the data with one randomly selected replicate per condition, and the full available datasets. Values closer to 1 indicate higher sensitivity and specificity. The areas under the ROC curves are consistently smaller than the values in the first row of Table 1 in the main manuscript.

Figure 8: The empirical cumulative distribution function (ECDF) curves of detecting differentially expressed genes for the five datasets with no external 'gold standard' when shrinking the variance estimates. Y-axis: ECDF, function of the gene rank. X-axis: p-value. Solid line: unreplicated comparison AvsA. Dotted line: unreplicated comparison AvsB. Dashed line: AvsB on the full dataset for two-group designs. Dotted-dashed line: AvsB on the full dataset for more complex designs. Gray line: 45 degree. The the curves are less consistent with the expected patterns than the curves in the first column of Fig. 2 in the main manuscript.
Fig. 9 compares the sensitivity of three analysis strategies of the Tuch dataset with paired experimental design as described in Sec. 2.3. Shrinking averaged estimates of dispersions resulted in a higher sensitivity. At the same time, both approaches are less sensitive than the analysis that ignores the paired nature of the design. The loss of sensitivity is possibly due to inefficiencies of the Fisher’s methods, and several alternatives can be considered.

Figure 9: The number of differentially expressed genes in the Tuch dataset with paired experimental design. ‘perPairDisp’: separate dispersion estimation and shrinkage for each subject. ‘poolDisp’: averaged per-subject method of moments estimates of dispersion, and a single shrinkage step of the averaged estimates. ‘factor2’: analysis that ignores the paired nature of the design, and treats it as a two-group factorial experiment.
4.4 Effect of sample size

To study the effect of sample size, we repeated simulation3 for \( n_i = 1, 2, 3, 4, 6, 8, 10 \). Table 2 summarizes the performance of the proposed method, as well as of DESeq and edgeR with both the exact test and the generalized linear model-based approach. The results indicate that sSeq is particularly advantageous for experiments with \( n_i \leq 4 \).

| sSeq | 0.856 | 0.888 | 0.903 | 0.913 | 0.917 | 0.926 | 0.929 |
|------|-------|-------|-------|-------|-------|-------|-------|
| edgeR | 0.840 | 0.833 | 0.844 | 0.852 | 0.860 | 0.875 | 0.871 |
| DESeq | 0.842 | 0.815 | 0.885 | 0.894 | 0.907 | 0.914 | 0.915 |
| baySeq | 0.558 | 0.628 | 0.628 | 0.620 | 0.616 | 0.609 | 0.603 |
| BBSseq | 0.578 | 0.619 | 0.601 | 0.613 | 0.610 | 0.591 | 0.607 |
| SAMseq | - | 0.882 | 0.897 | 0.903 | 0.916 | 0.925 | 0.926 |
| GLM edgeR | - | 0.799 | 0.898 | 0.878 | 0.943 | 0.946 | 0.956 |
| GLM DESeq | - | 0.828 | 0.832 | 0.831 | 0.835 | 0.834 | 0.831 |

Table 2: Areas under the ROC curves of detecting differentially expressed genes for the Simulation3 when the number of samples increases. Values closer to 1 indicate higher sensitivity and specificity.
4.5 Effect of size factors

Table 3 the true and the estimated size factors for the ten datasets in this manuscript. The estimates are obtained with the proposed approach (i.e. are identical to the estimates by DESeq).

| Datasets            | $s_{A_j}$  |         | $s_{B_j}$  |         |
|---------------------|------------|---------|------------|---------|
|                     | True       | Estimated | True       | Estimated |
| Simulation1         | 0.6, 1.5   | 0.684, 1.729 | 0.8, 0.7   | 0.987, 0.866 |
| Simulation2         | 0.6, 1.5   | 0.686, 1.732 | 0.8, 0.7   | 0.984, 0.863 |
| Simulation3         | 1.3, 1.6   | 1.006, 1.196 | 1.4, 0.9   | 1.126, 0.749 |
| MAQC                | 0.966, 1.202, 1.023 | 0.516, 1.706 |
| Griffith et al.     | 1.317      |          | 0.760      |         |
| Brooks              | 1.297, 1.042 |          | 0.819, 0.911 |         |
| Sultan              | 0.917, 0.862 |          | 1.160, 1.132 |         |
| Bottomly            | between 0.578 and 1.524 | between 0.756 and 1.616 |         |
| Hammer              | 1.038, 0.897 |          | 1.027, 1.065 |         |
| Tuch                | 0.719, 0.831, 1.753 |          | 0.627, 1.084, 1.424 |         |

Table 3: The true and estimated size factors for the ten datasets. The estimates are obtained with the proposed approach (i.e. equivalent to DESeq). The true values of size factors are only available for simulated datasets.
To investigate the effect of the estimates of size factors on the accuracy of the results, we conducted an additional evaluation for the three simulated datasets with \( n_A = n_B = 2 \), and for the methods that assume a Negative Binomial distribution. Table 4 shows the results of the original implementation of each method. Note that edgeR and baySeq estimate size factors on the total count scale, while the other methods use a relative scale. edgeR multiplies the total library size by the output of \texttt{calcNormFactors}. baySeq uses the total library size directly.

**Table 4:** The estimates of the size factors by each method, and the corresponding areas under the ROC curves for the three simulated datasets.

|                | Sim1          | Sim2          | Sim3          | AUROC |
|----------------|---------------|---------------|---------------|-------|
| \( \hat{s}_{ij} \) by each method |              |               |               |       |
| sSeq           | 0.684, 1.729, | 0.686, 1.732, | 1.006, 1.196, | 0.962 |
|                | 0.987, 0.866  | 0.984, 0.863  | 1.126, 0.749  | 0.967 |
|                | 0.987, 0.866  | 0.984, 0.863  | 1.126, 0.749  | 0.888 |
| edgeR          | 2674477×1.119 | 2651592×1.129 | 7720510×1.132 | 0.948 |
|                | 6677525×1.136 | 6625427×1.130 | 9168128×1.128 | 0.951 |
|                | 4832029×0.887 | 4818957×0.885 | 10830341×0.885 | 0.833 |
|                | 4230815×0.887 | 4215156×0.886 | 7196265×0.885 |       |
| DESeq          | 0.684, 1.729, | 0.686, 1.732, | 1.006, 1.196, | 0.940 |
|                | 0.987, 0.866  | 0.984, 0.863  | 1.126, 0.749  | 0.949 |
|                | 0.987, 0.866  | 0.984, 0.863  | 1.126, 0.749  | 0.816 |
| baySeq         | 2674477, 6677525, 4832029, 4230815 | 2651592, 6625427, 4818957, 4215156 | 7720510, 9168128, 10830341, 7196265 | 0.711 |
|                | 7720510, 9168128, 10830341, 7196265 | 0.714 | 0.628 |

Table 5 uses the size factors estimated by sSeq (equivalently, by DESeq) with the two other methods. The size factors were converted to the appropriate scale by multiplying the total library sizes by the size factor estimates of sSeq and DESeq.

**Table 5:** Areas under the ROC curves for edgeR and baySeq for the three simulated datasets, while using the size factors estimated by sSeq (equivalently, by DESeq).
Table 6 uses the true size factors with all the methods. The size factors were converted to the appropriate scale for edgeR and baySeq by multiplying the total library sizes by the true simulated size factors.

| True size factors $s_{ij}$ | AUROC |
|-----------------------------|-------|
|                            | Sim1  | Sim2  | Sim3  | Sim1 | Sim2 | Sim3 |
| sSeq 0.6, 1.5, 0.8, 0.7     | 0.974 | 0.977 | 0.896 |
| edgeR 2674477×0.6, 6677525×1.5, 4832029×0.8, 4230815×0.7 | 0.912 | 0.916 | 0.872 |
| DESeq 0.6, 1.5, 0.8, 0.7    | 0.967 | 0.971 | 0.860 |
| baySeq 2674477×0.6, 6677525×1.5, 4832029×0.8, 4230815×0.7 | 0.890 | 0.899 | 0.923 |

Table 6: Areas under the ROC curves for edgeR and baySeq for the three simulated datasets, while using the true values of size factors used for the simulations.

To summarize, the results showed that size factors do indeed play a role in the accuracy of the results. Importantly, the changes did not affect the conclusions of the manuscript. The proposed method sSeq consistently showed a strong performance, except in Simulation3 where baySeq combined with a better size factor had a higher area under the ROC curve.
4.6 Representative R scripts (Sultan dataset)

```r
library(sSeq);
library(edgeR);
library(DESeq);
library(baySeq);
library(BBSseq);
library(samr);

data(Sultan); # included in the sSeq package
head(countsTable); conds; # included in the sSeq package

# sSeq
res.sSeq = nbTestSH( countsTable, conds, "A", "B");
head(res.sSeq); # A table includes the p-values;

# edgeR
dgl = DGEList( counts=countsTable, group=conds)
dgl = calcNormFactors(dgl);
dgl = estimateCommonDisp(dgl);
dgl = estimateTagwiseDisp(dgl);
res.edgeR = exactTest( dgl, dispersion="auto");
head(res.edgeR$table); # A table includes the p-values;

# DESeq
cds = newCountDataSet( countsTable, conds )
cds = estimateSizeFactors( cds );
cds = estimateDispersions(cds , fitType="local", method="per-condition"); # with replication in each condition
# When there are not replicates in each condition
# cds = estimateDispersions( cds , fitType="local", method="blind", sharingMode="fit-only");
res.DESeq = nbinomTest( cds, "A", "B" );
head(res.DESeq); # A table includes p-values;

# baySeq
ngenes = nrow(countsTable);
replicates = conds=="B"; replicates <- replicates +1;
groups = list(NDE=rep(1, length(conds)), DE=replicates);
libsizes = colSums(countsTable);
CD = new("countData", data=as.matrix(countsTable), replicates=replicates, libsizes=libsizes, groups=groups);
CD@annotation = data.frame(names=paste("count", 1:ngenes, sep="_"));
CD.nbml = getPriors.NB(CD, samplesize=ngenes, estimation="QL", cl=NULL); # QuasiLikelihood
CDpost.nbml = getLikelihoods.NB(CD.nbml, pET="BIC", cl=NULL);
res.baySeq = topCounts(CDpost.nbml, group="B", number=ngenes)[rownames(countsTable),];
head(res.baySeq); # The Likelihood in the table is the posterior probability;

# BBSseq
groups = cbind(1, conds=="B");
ofree = free.estimate(countsTable, groups);
res.BBSseq = data.frame(id=rownames(countsTable), p.free=ofree$p.free , flag=outlier.flag(countsTable));
head(res.BBSseq); # A table includes p-values;

# SAMseq
groups = conds=="B" +1 ;
samfit = SAMseq(x=countsTable, y=groups, resp.type="Two class unpaired");
res.SAMseq= samr.pvalues.from.perms(samfit$samr.obj$tt, samfit$samr.obj$ttstar);
head(res.SAMseq); # p-values
```
References

[1] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biol.*, 11:R106, 2010.

[2] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. *JRSS(B)*, 57:289–300, 1995.

[3] D. Bottomly et al. Evaluating Gene Expression in C57BL/6J and DBA/2J Mouse Striatum Using RNA-seq and Microarrays. *PloS one*, 6(3):e17820, 2011.

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

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

[6] B. Langmead et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10(3):R25, 2009.

[7] T. M Loughin. A systematic comparison of methods for combining p-values from independent tests. *Computational Statistics and Data Analysis*, 47:467–485, 2004.

[8] J.C. Marioni et al. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, 18:1509–1517, 2008.

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