

MODERATED ESTIMATION OF FOLD CHANGE AND DISPERSION FOR RNA-SEQ DATA WITH DESEQ2

LOVE ET AL.

PRESENTED BY KEONVIN PARK

Published in 2014 in Genome Biology

METHOD-DESEQ2



TABLE OF CONTENTS

1. Background
2. Results and Discussion
3. Conclusions
4. Implementation (using colab notebook)

I. BACKGROUND

Aim: finding genes that are differentially expressed across groups of samples.

Problem: non-normality, dependence of variance on the means, small number of samples

Solution: Pooling information across genes

Solution: DESeq method

DEG Analysis

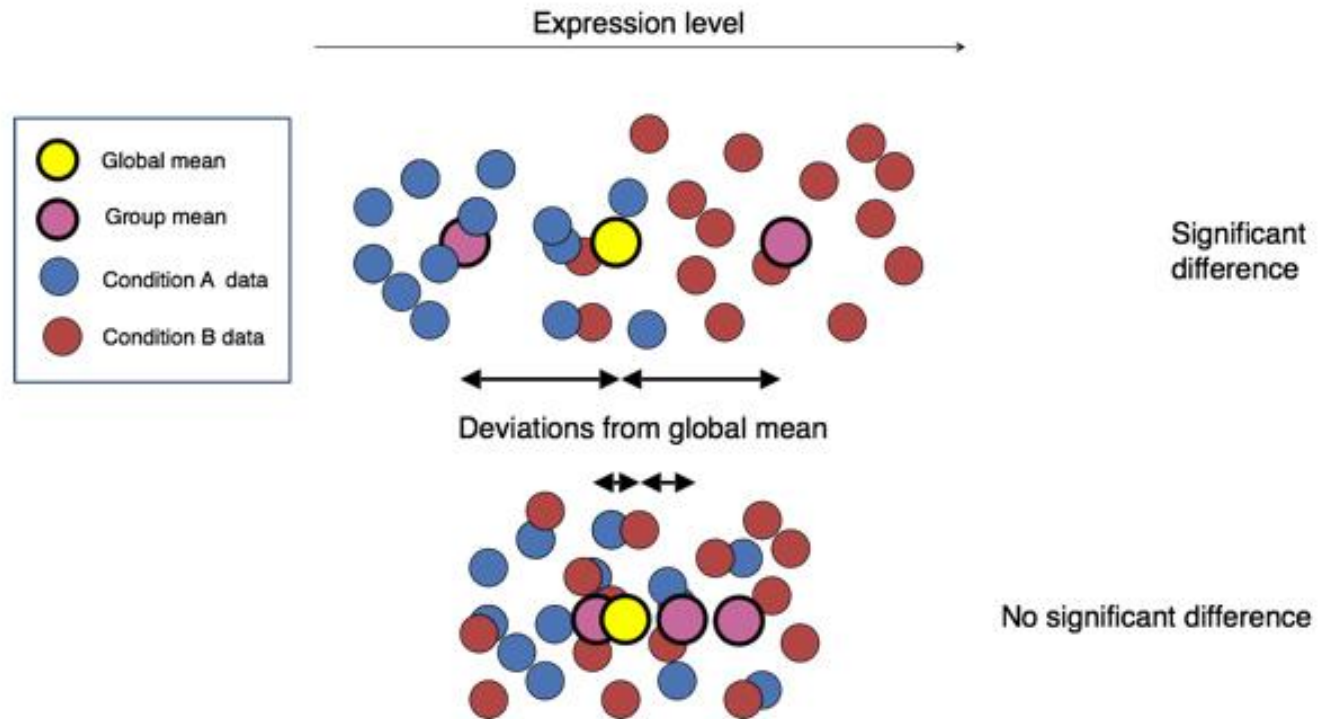


Image credit: Paul Pavlidis, UBC

I. BACKGROUND

Aim : test LFC between treatment and control for a gene expression



DESeq2: statistical framework to estimate LFC, also testing of differential expression



DESeq2: using shrinkage estimator for dispersion and fold change



DESeq2 : compared to existing methods, higher sensitivity and precision while controlling false positive rate.

2. RESULTS AND DISCUSSION

➤ Model and normalization

Model and normalization

The read count K_{ij} for gene i in sample j is described with a GLM of the negative binomial family with a log link

$$K_{ij} \sim \text{NB}(\text{mean} = \mu_{ij}, \text{dispersion} = \alpha_i)$$

$\mu_{ij} = \underbrace{\phi_{ij}}_{\text{scaling factor}} g_{ij}$

$$\log g_{ij} = \sum_r x_{ijr} \beta_{ir}$$

Where pdf of $\text{NB}(r, p)$ is as follows

$$p = \frac{\mu}{\sigma^2} \text{ and } r = \frac{\mu^2}{\sigma^2 - \mu}$$

$$Pr(K=k) = \binom{k+r-1}{r-1} p^r (1-p)^k$$

2. RESULTS AND DISCUSSION

➤ Model and normalization

$$\log_2 g_{ij} = \sum_{r=1}^R x_{ir} B_{ir}$$

if $R=2$ (two-group)

$$\log_2 g_{ij} = x_{i0} B_{i0} + x_{i1} B_{i1}$$

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{array}{l} \text{Control} \rightarrow \log_2 g_{ij} = B_{i0} \\ \text{Case} \rightarrow \log_2 g_{ij}' = B_{i0} + B_{i1} \end{array}$$

design matrix

Log-fold change.

$$\log_2 g_{ij}' - \log_2 g_{ij} = \log_2 \frac{g_{ij}'}{g_{ij}} = B_{i1}$$

ex) Sample (j), $j = 1, 2, 3, \dots, m$

when $j=1$

gene	1	2	3	4	K_1^R
(i)	5	6	7	8	K_2^R
	9	10	11	12	K_3^R
$i=1, 2, 3, 4$	13	14	15	16	K_4^R

$$s_j = \text{median}_{i: K_i^R \neq 0} \frac{K_{ij}}{K_i^R} \text{ with } K_i^R = \left(\prod_{j=1}^m K_{ij} \right)^{1/m}$$

For example, to get s_1 ($j=1$)

$$s_1 = \text{median}_{i: K_i^R \neq 0} \frac{K_{i1}}{K_i^R} = \text{median} \left(\frac{1}{K_1^R}, \dots, \frac{13}{K_4^R} \right)$$

$$\text{Where } K_1^R = \left(\prod_{j=1}^4 K_{1j} \right)^{1/4} = (1 \cdot 2 \cdot 3 \cdot 4)^{1/4}$$

$$K_4^R = \left(\prod_{j=1}^4 K_{4j} \right)^{1/4} = (13 \cdot 14 \cdot 15 \cdot 16)^{1/4}$$

2. RESULTS AND DISCUSSION

➤ Empirical Bayes shrinkage for dispersion estimation

$$\text{Var}(K_{ij}) = \mu_{ij} + \alpha_i \mu_{ij}^2$$

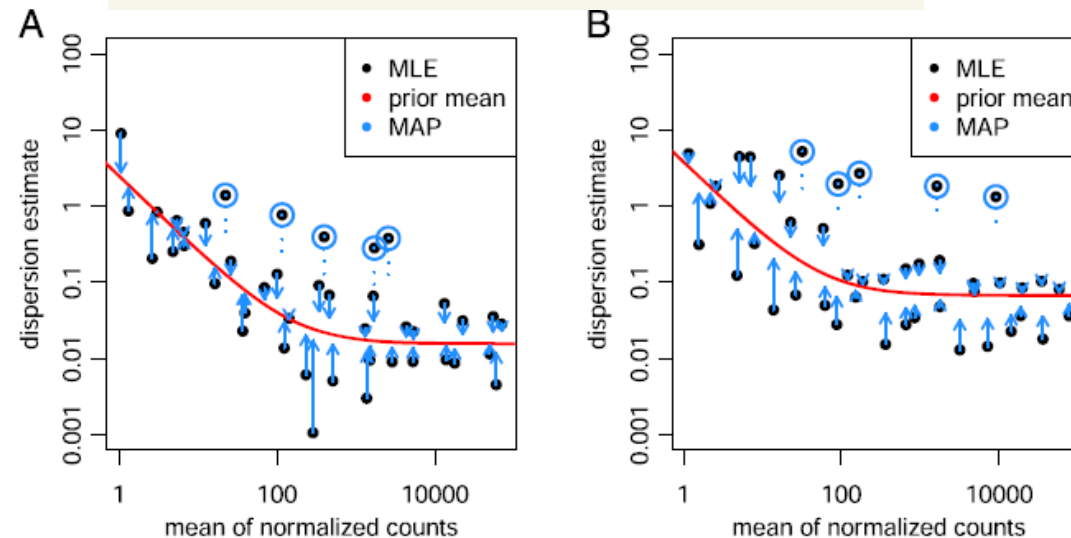
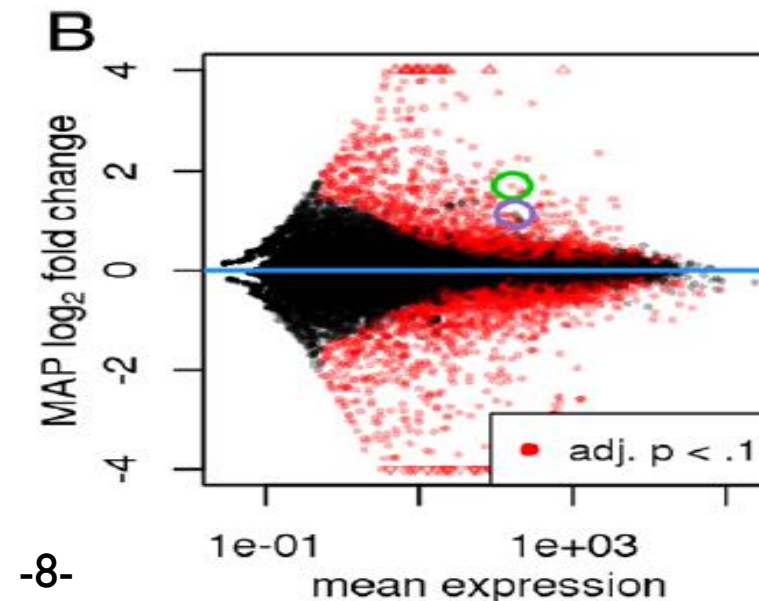
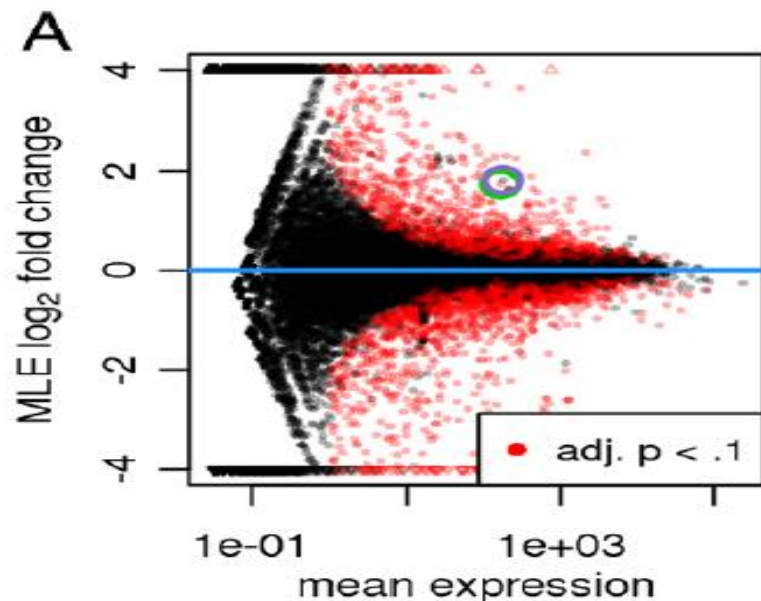


Figure 1 Shrinkage estimation of dispersion. Plot of dispersion estimates over the average expression strength **(A)** for the Bottomly *et al.* [16] dataset with six samples across two groups and **(B)** for five samples from the Pickrell *et al.* [17] dataset, fitting only an intercept term. First, gene-wise MLEs are obtained using only the respective gene's data (black dots). Then, a curve (red) is fit to the MLEs to capture the overall trend of dispersion-mean dependence. This fit is used as a prior mean for a second estimation round, which results in the final MAP estimates of dispersion (arrow heads). This can be understood as a shrinkage (along the blue arrows) of the noisy gene-wise estimates toward the consensus represented by the red line. The black points circled in blue are detected as dispersion outliers and not shrunk toward the prior (shrinkage would follow the dotted line). For clarity, only a subset of genes is shown, which is enriched for dispersion outliers. Additional file 1: Figure S1 displays the same data but with dispersions of all genes shown. MAP, maximum *a posteriori*; MLE, maximum-likelihood estimate.

2. RESULTS AND DISCUSSION

➤ Empirical Bayes shrinkage for fold-change estimation

- A common difficulty in the analysis of HTS data is the **strong variance of LFC estimates** for genes with low read count.
- DESeq2 overcomes this issue by **shrinking LFC estimates toward zero** in a manner such that shrinkage is stronger when the available information for gene is low, which may be because counts are low, dispersion is high or there are few degrees of freedom.



2. RESULTS AND DISCUSSION

➤ Hypothesis tests for differential expression

Aim: one may test whether each model coefficient differs significantly from zero.



Solution: Wald test resulting in a z-statistic, which is compared to a standard normal distribution.



Solution2: p-values are adjusted for multiple testing using the procedure of **Benjamini and Hochberg [21]**.

Benjamini-Hochberg procedure

Background: The setting for many procedures is such that we have H_1, \dots, H_m null hypotheses tested and p_1, \dots, p_m their corresponding p-values. We list these p-values in ascending order: $p_{(1)}, \dots, p_{(m)}$.

↗ expected proportion of discoveries that are false.

Procedure: B-H controls **FDR** at level α . It works as follows.

1. For given α , find the largest K s.t. $p_{(K)} \leq \frac{K}{m} \alpha$

2. Reject the null hypothesis for all $H_{(i)}$, $i=1, \dots, K$

↗ prob. of at least one false discovery.
Difference with Bonferroni Correction: B-F correction controls **FWER** (familywise error rate) which control the prob. of at least one Type I error. Thus, FDR-controlling B-H have greater power, at the cost of increased Type-I error.

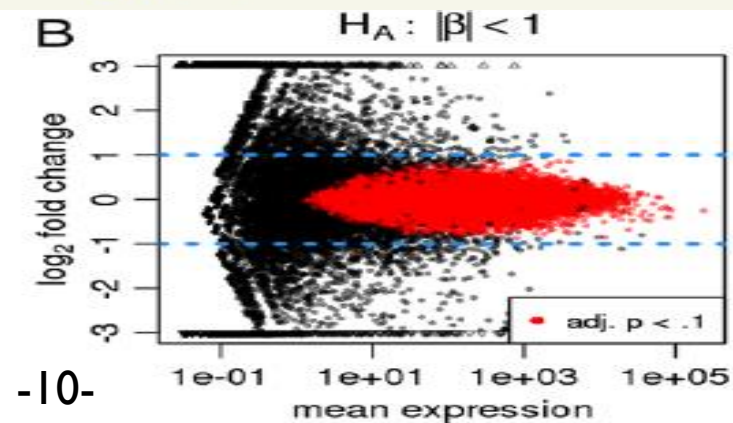
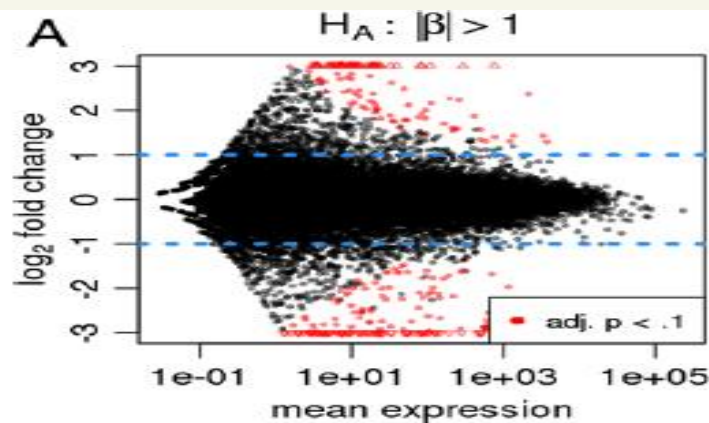
2. RESULTS AND DISCUSSION

- Hypothesis tests for thresholds on effect size (specifying minimum effect size)

Disregard genes whose estimated LFC β_{ir} is below some threshold $|\beta_{ir}| \leq \theta$

- Hypothesis tests for thresholds on effect size (specifying maximum effect size)

Disregard genes whose estimated LFC β_{ir} is above some threshold $|\beta_{ir}| \geq \theta$



2. RESULTS AND DISCUSSION

➤ Detection of count outliers

Cook's distance

The MLE of $\vec{\beta}_i = (\beta_{i0}, \beta_{i1})^T$ is used for calculating Cook's distance.

Considering a gene i and sample j , Cook's distance for GLMs is given by

$$D_{ij} = \frac{R_{ij}^2}{\text{DF}(1-h_{ij})^2}$$

\rightarrow feature residual.
 h_{ij} \rightarrow j -th diagonal element of the hat matrix
 $H = W^{1/2} X (X^T W X)^{-1} X^T W^{1/2}$
 DF \rightarrow # of parameters.

$$, \text{ where } R_{ij} = \frac{(k_{ij} - \mu_{ij})}{\sqrt{V(\mu_{ij})}}$$

Aim: Detection of count outliers



Solution: Calculate cook's distance



Solution: Delete sample points which value is over 1 or $n/4$

2. RESULTS AND DISCUSSION

➤ Comparative benchmarks (in simulation setting)

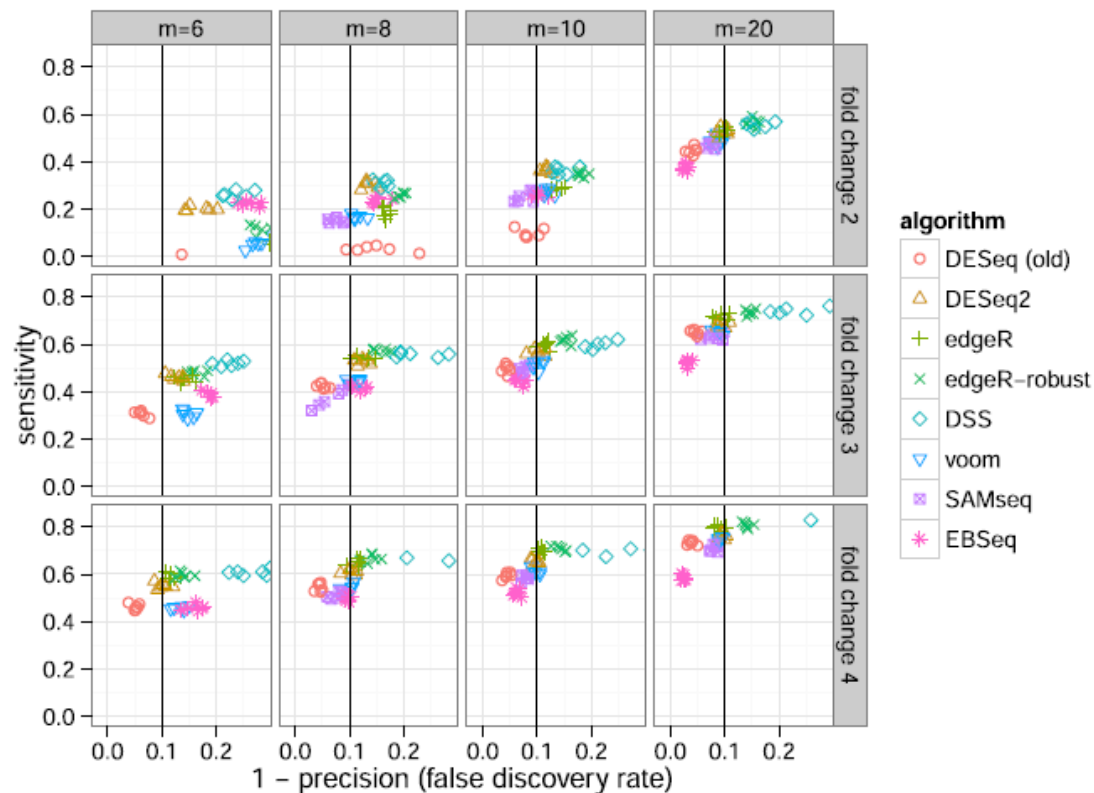


Figure 6 Sensitivity and precision of algorithms across combinations of sample size and effect size. *DESeq2* and *edgeR* often have sensitivity of those algorithms that controlled the FDR, i.e., those algorithms which fall on or to the left of the vertical black line. For a plot of sensitivity against false positive rate, rather than FDR, see Additional file 1: Figure S8, and for the dependence of sensitivity on the mean see Additional file 1: Figure S9. Note that *EBSeq* filters low-count genes (see main text for details).

These datasets were of varying total sample size ($m \in \{6, 8, 10, 20\}$), and the samples were split into two equal-sized groups; 80% of the simulated genes had no true differential expression, while for 20% of the genes, true fold changes of 2, 3 and 4 were used to generate counts across the two groups, with the direction of fold change chosen randomly. The simulated differentially expressed genes were chosen uniformly at random among all the

P value < 0.1 . The sensitivity is plotted over 1 - precision, or the FDR, in Figure 6. *DESeq2*, and also *edgeR*, often had the highest sensitivity of the algorithms that controlled type-I error in the sense that the actual FDR was at or below 0.1, the threshold for adjusted P values used for calling differentially expressed genes. *DESeq2* had higher sensitivity compared to the other algorithms, particularly for small fold change (2 or 3), as was also found in benchmarks performed by Zhou *et al.* [34]. For larger sample sizes and larger fold changes the performance of the various algorithms was more consistent.

The overly conservative calling of the old *DESeq* tool can be observed, with reduced sensitivity compared to the other algorithms and an actual FDR less than the nominal value of 0.1. We note that *EBSeq* version 1.4.0 by default

2. RESULTS AND DISCUSSION

➤ Comparative benchmarks (in real data setting)

Metric: 1-specificity

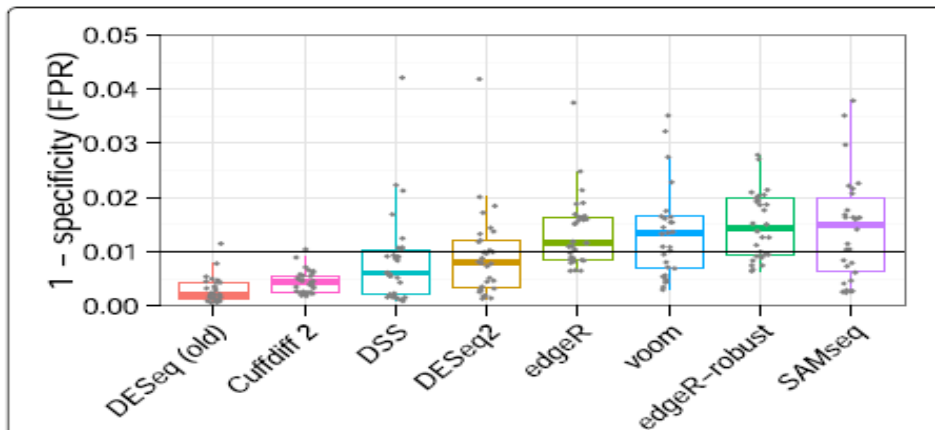


Figure 7 Benchmark of false positive calling. Shown are estimates of $P(P\text{value} < 0.01)$ under the null hypothesis. The FPR is the number of P values less than 0.01 divided by the total number of tests, from randomly selected comparisons of five vs five samples from the Pickrell *et al.* [17] dataset, with no known condition dividing the samples. Type-I error control requires that the tool does not substantially exceed the nominal value of 0.01 (black line). *EBSeq* results were not included in this plot as it returns posterior probabilities, which unlike P values are not expected to be uniformly distributed under the null hypothesis. FPR, false positive rate.

Aim: Validation of algorithms in real data



Metric: 1-specificity, sensitivity, preicision



Method: true means differences is set by $P < 0.1$ in larger validation set

2. RESULTS AND DISCUSSION

➤ Comparative benchmarks (in real data setting)

Metric: sensitivity

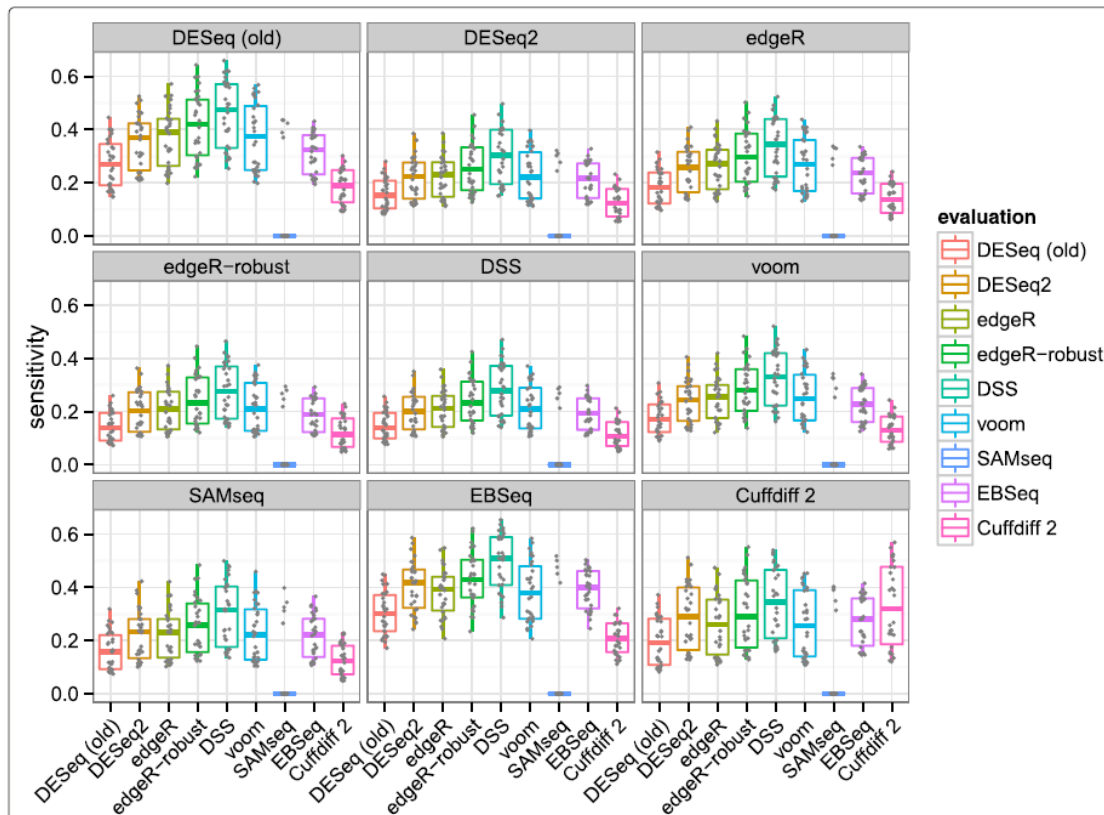


Figure 8 Sensitivity estimated from experimental reproducibility. Each algorithm's sensitivity in the evaluation set (box plots) is evaluated using the calls of each other algorithm in the verification set (panels with grey label).

Aim: Validation of algorithms in real data



Metric: I-specificity, sensitivity, preicision



Method: true means differences is set by $P < 0.1$ in larger validation set

2. RESULTS AND DISCUSSION

➤ Comparative benchmarks (in real data setting)

Metric: precision

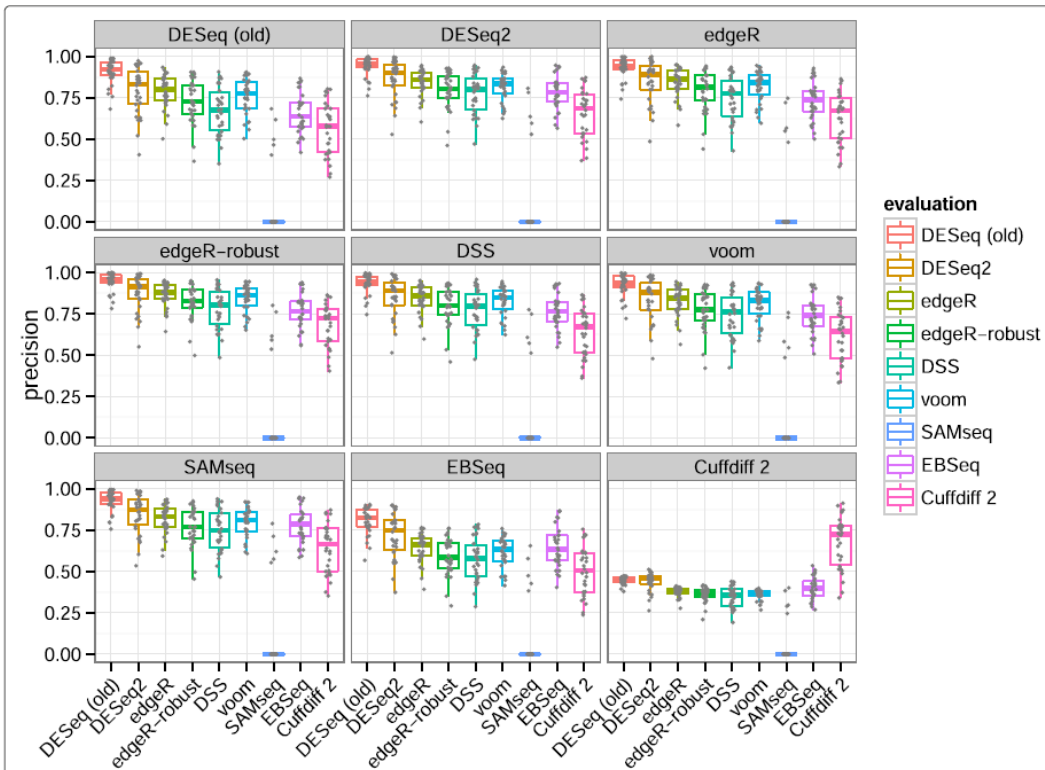


Figure 9 Precision estimated from experimental reproducibility. Each algorithm's precision in the evaluation set (box plots) is evaluated using the calls of each other algorithm in the verification set (panels with grey label).

Aim: Validation of algorithms in real data



Metric: I-specificity, sensitivity, preicision



Method: true means differences is set by $P < 0.1$ in larger validation set

3. CONCLUSIONS

➤ Conclusions

- In summary, the benchmarking tests showed that DESeq2 effectively controlled type-I errors, maintaining a median false positive rate just below the chosen critical value.
- For both simulation and analysis of real data, DESeq2 often achieved the highest sensitivity of those algorithms that controlled the FDR.
- DESeq2 offers a comprehensive and general solution for gene-level analysis of RNA-seq data.
- Shrinkage estimator substantially improve the stability and reproducibility of analysis results compared to maximum-likelihood-based solutions.
- Finally, the DESeq2 package is integrated well in the Bioconductor infrastructure and comes with extensive documentation.

4. IMPLEMENTATION (USING COLAB NOTEBOOK)

➤ Implementation (using R in colab)

First) connect to <https://colab.research.google.com/notebook#create=true&language=r>



Second) upload .ipynb files



4. IMPLEMENTATION (USING COLAB NOTEBOOK)

➤ Implementation (using R in colab)

00. Connect to below site for running R in colab

```
[ ] https://colab.research.google.com/notebook#create=true&language=r
```

01. Installation

This section should be accompanied by a code with instructions on how to install it.

```
[ ] if (!requireNamespace("BiocManager", quietly = TRUE))  
    install.packages("BiocManager")
```


Installing package into '/usr/local/lib/R/site-library'
(as 'lib' is unspecified)

```
▶ BiocManager::install("DESeq2") # 10mins
```

↳ 'getOption("repos")' replaces Bioconductor standard repositories, see
'?repositories' for details

replacement repositories:
CRAN: <https://cran.rstudio.com>

Bioconductor version 3.13 (BiocManager 1.30.16), R 4.1.1 (2021-08-10)



Q&A