Dispersion Estimation and Its Effect on Test Performance in RNA-seq Data Analysis

Will Landau and Dr. Peng Liu

Iowa State University, Ames, IA, USA

Abstract

A central goal of RNA sequencing (RNA-seq) experiments is to detect differentially expressed genes. In the ubiquitous negative binomial model for RNA-seq data, each gene is given a dispersion parameter, and correctly estimating these dispersion parameters is vital to detecting differential expression. Since the dispersions help control the variances of the gene counts, underestimation may lead to false discovery, while overestimation may lower the rate of true detection. The simulation study here compares several existing dispersion estimation methods in terms of point estimation and performance in tests for differential expression. The methods that maximize the test performance are the ones that use a moderate degree of dispersion shrinkage: the DSS, Tagwise wqCML, and Tagwise APL methods. In practical RNA-seq data analysis, we recommend using one of these moderate-shrinkage methods with the QLShrink test in `QuasiSeq` R package.

Modeling RNA-seq data

- Negative Binomial (NB) model for the expression level of gene g in sample i:

$$y_{g,i} \sim \text{NB}(\text{mean} = \mu_{g,i},\ \text{dispersion} = \phi_g)$$
$$\mu_{g,i} = s_i \cdot \nu_{g,k(i)}$$
$$k(i) = \text{Treatment group of sample } i$$
$$s_i = \text{Normalization factor of sample } i$$

- Dispersion ϕ_g controls the variance.

$$Var(y_{g,i}) = \mu_{g,i} + \mu_{g,i}^2 \cdot \phi_g$$

- Overestimating dispersions may lead to false negatives, while underestimating dispersions may lead to false positives.

Existing methods for estimating dispersions

Method	Description	Authors
QL	Quasi-Likelihood	cited by Robinson and Smyth (2008)
DSS	Dispersion Shrinkage for Sequencing	Wu, Wang, and Wu (2012)
wqCML	Weighted Quantile-Adjusted Conditional Maximum Likelihood	Robinson and Smyth (2008)
APL	Cox-Reid Adjusted Profile Likelihood	McCarthy, Chen, and Smyth (2012)
DESeq	Differential Expression for Sequence Count Data	Anders and Huber (2010)

Most methods shrink dispersions towards a common value, trend, or prior distribution.

Method	Dispersion shrinkage options
QL	None
DSS	Empirical Bayes model with a shared log-normal prior for the ϕ_g's.
wqCML	Common: maximize shared log likelihood, $l_S(\phi_g)$ Tagwise: maximize weighted log likelihood, $l_g(\phi_g) + \alpha \cdot l_S(\phi_g)$
APL	Common: maximize $APL_S(\phi_g) = \frac{1}{G}\sum_{g'=1}^{G} APL_{g'}(\phi_g)$ Trended: Restrict ϕ_g's to a trend and maximize $APL_g(\phi_g)$. Tagwise: Use local mean APL (APL_{S_g}) and maximize $APL_g(\phi_g) + \alpha \cdot APL_{S_g}(\phi_g)$
DESeq	None Trended: fit ϕ_g's to trend. Maximum: take ϕ_g to be the maximum of the raw estimate and trend.

The simulation study

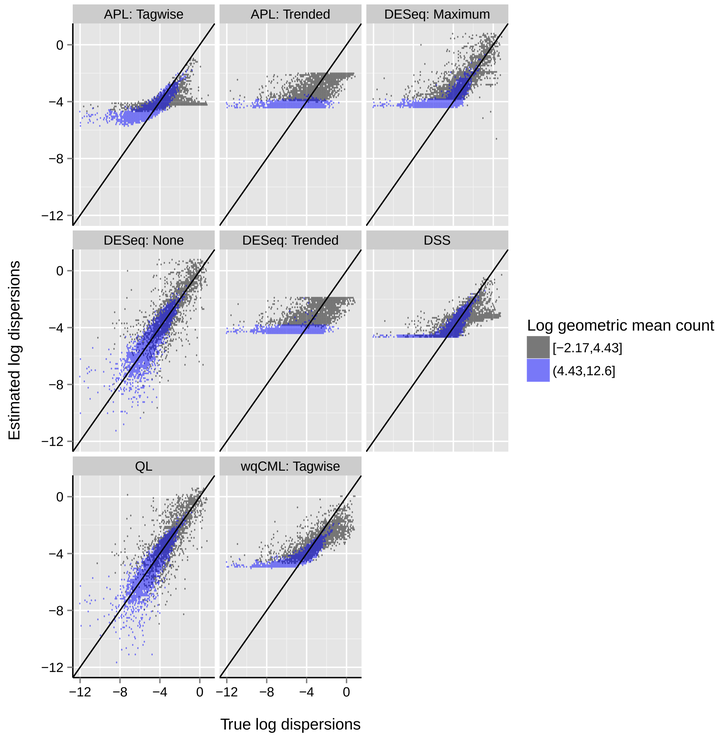
- 30 datasets (2 treatment groups and 10,000 genes each) were simulated from the negative binomial model under each of the following settings:

Setting	Dataset	Group 1 samples	Group 2 samples
I	Pickrell	3	3
II	Pickrell	3	15
III	Pickrell	9	9
IV	Hammer	3	3
V	Hammer	3	16
VI	Hammer	9	9

- In the simulated datasets, the true overall per-gene expression levels and true dispersions were taken from real datasets:

Pickrell: GSE19480 human
Hammer: GSE20895 mouse

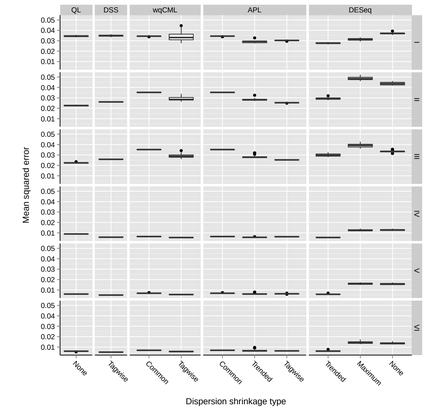
Estimation error: simulation setting VI



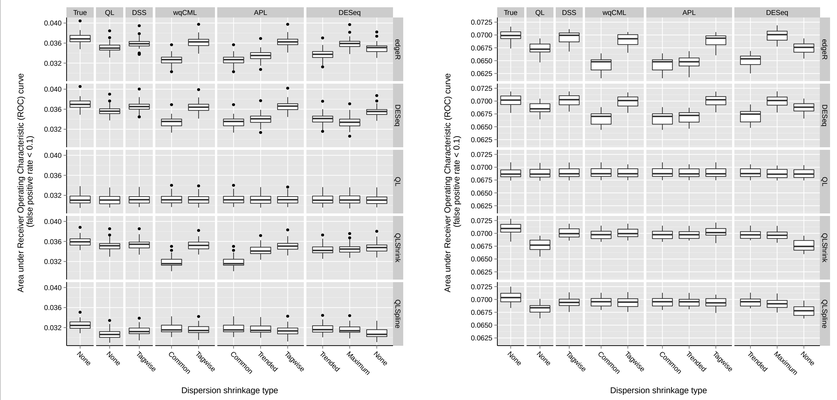
Mean squared error: simulation settings I through VI

Use the transformed dispersions for calculating mean squared error:

$$\text{MSE} = \sum_{g=1}^{G}\left(\frac{\widehat{\phi}_g}{1+\widehat{\phi}_g} - \frac{\phi_g}{1+\phi_g}\right)^2$$



Performance in five tests for differential expression: simulation settings II (left) and VI (right)



Conclusions

- Best dispersion estimation methods: DSS, Tagwise wqCML, Tagwise APL
- Best tests for differential expression:
  - edgeR exact test (performance varies with dispersion estimation method).
  - DESeq exact test (performance varies with dispersion estimation method).
  - QLShrink test (more robust robust under dispersion estimation method and departures from the negative binomial model).
- The best dispersion estimation methods use a moderate degree of dispersion shrinkage. However, the *kind* of shrinkage varies within this optimal group.
- Estimation error is not always indicative of test performance.
  - Maximum DESeq is one of the worst in terms of mean squared error, but often performs well in tests for differential expression.