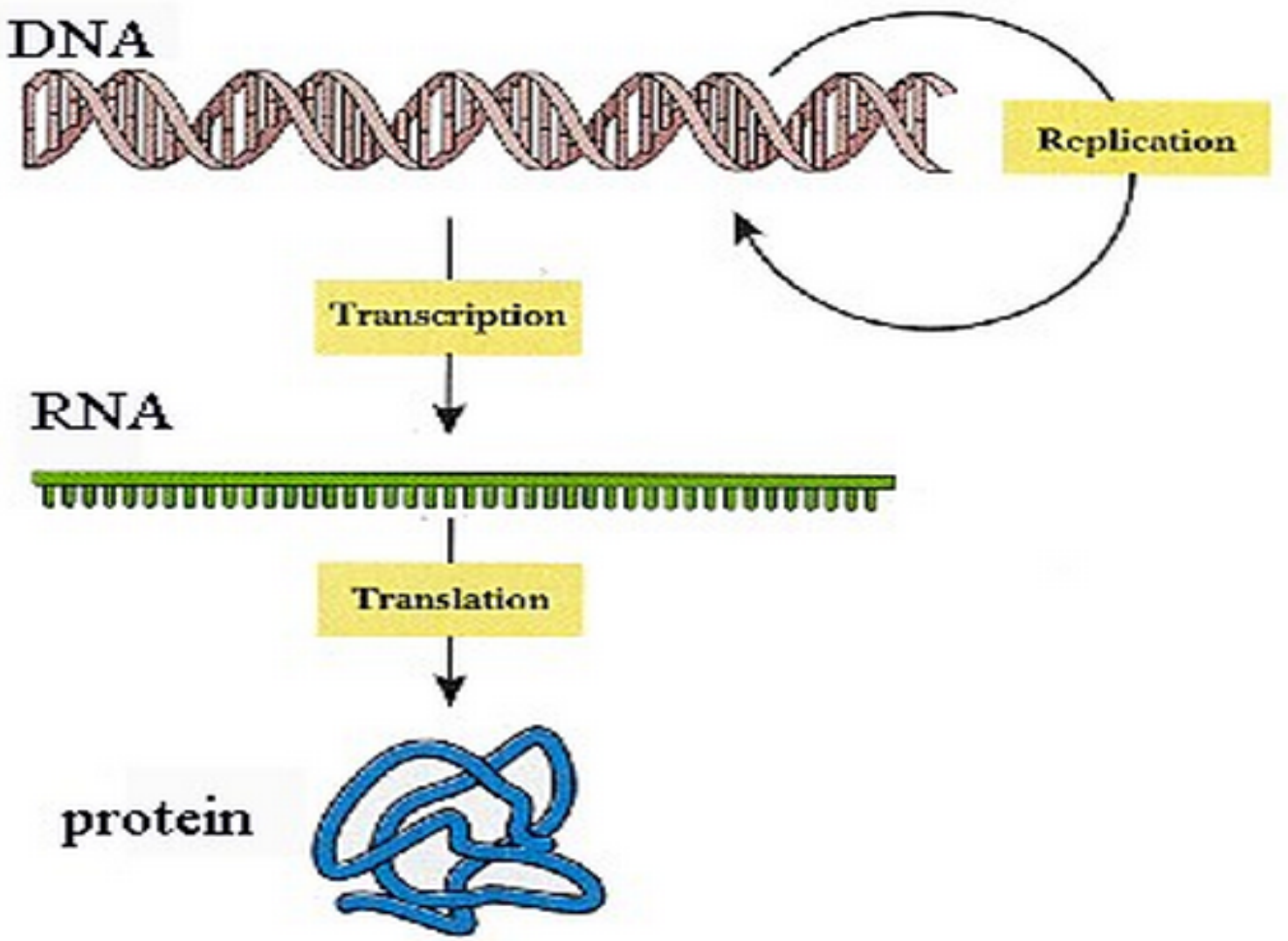
Dispersion Estimation and Its Effect on Test Performance in RNA-seq Data Analysis

Will Landau
Dr. Peng Liu

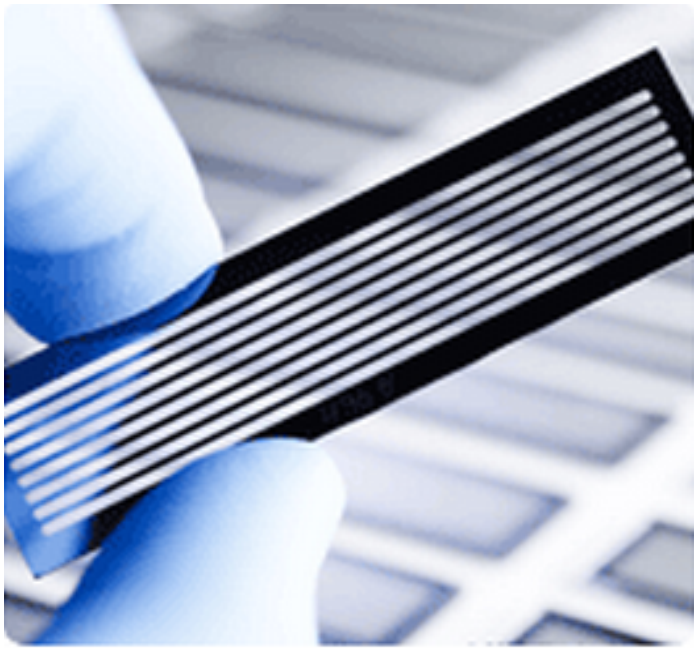
March 1, 2013

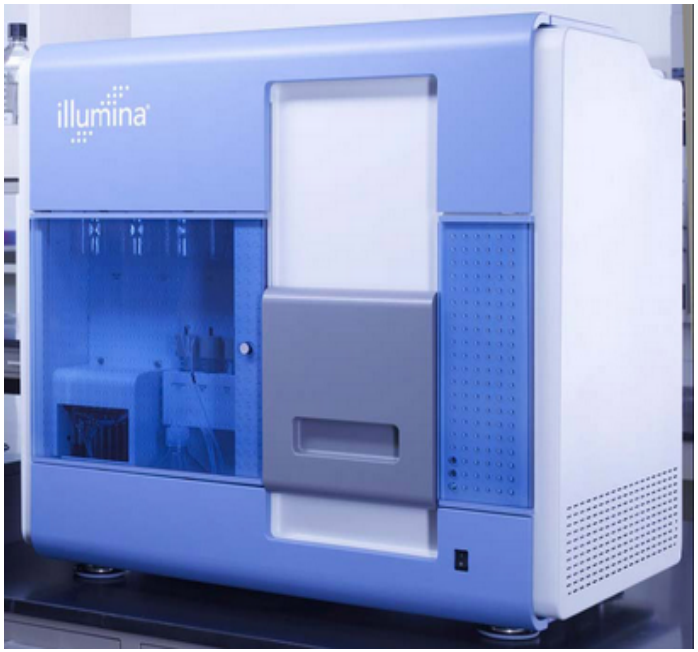
Outline



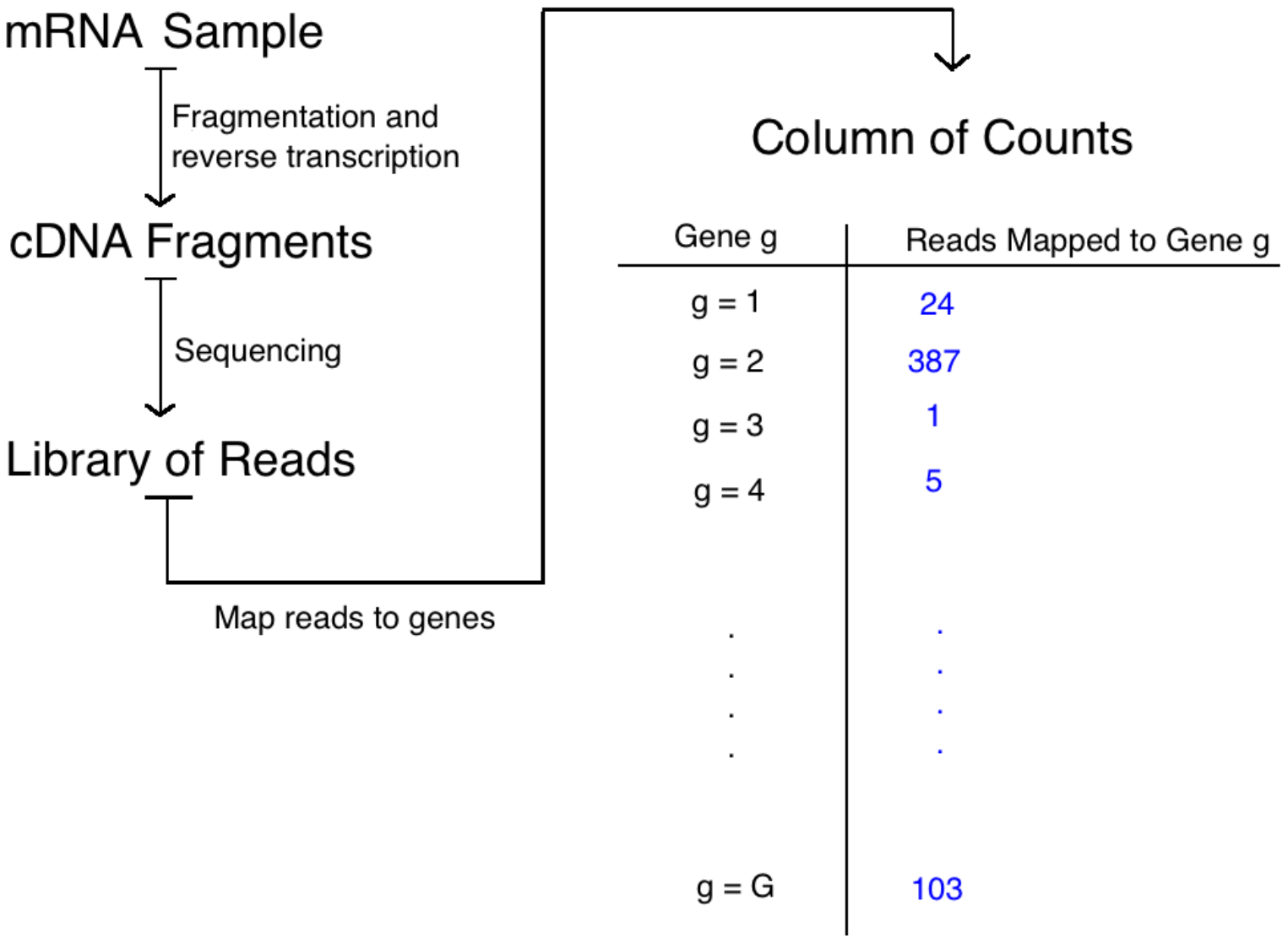
Next Generation Sequencing (NGS) Technologies

- A NGS platform measures the relative abundance of each RNA sequence in a sample.
- Example: Illumina's Genome Analyzer.





RNA-Seq Workflow



RNA-Seq Experiments

- Sequence multiple RNA samples from two or more treatment groups.
  - Biological replicates: original samples of genetic material (experimental units).
  - Technical replicates: repeated sequencing trials of the same sample of genetic material (observational units).
    - Libraries from different technical replicates may be pooled within each biological replicate.
- Central question: *which genes are differentially expressed across treatment conditions?*

An RNA-Seq Dataset

	Treatment Group k(i) = 1		Treatment Group k(i) = 2	
Gene g	Library i = 1	Library i = 2	Library i = 3	Library i = 4
g = 1	24	84	8	3
g = 2	387	110	27	32
g = 3	1	3	0	1
g = 4	5	4	4	6
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
g = G	103	94	100	98

The Negative Binomial Model

- Let $Y_{g,i}$ be the number of reads in library i mapped to gene g.
- If $Y_{g,i} \sim \text{NB}(\mu_{g,i}, \phi_g)$, then:

$$f(y \mid \mu_{g,i}, \phi_g) = \frac{\Gamma(y + {\phi_g}^{-1})}{\Gamma({\phi_g}^{-1})\Gamma(y+1)} \left(\frac{{\phi_g}^{-1}}{\mu_{g,i} + {\phi_g}^{-1}}\right)^{{\phi_g}^{-1}} \left(1 - \frac{{\phi_g}^{-1}}{\mu_{g,i} + {\phi_g}^{-1}}\right)^{y}$$

- As $\phi_g \to 0$, f converges to the Poisson pmf.
- $\text{Var}(Y_{g,i}) = \mu_{g,i} + \mu_{g,i}^2 \phi_g$
- $E(Y_{g,i}) = \mu_{g,i} = s_i \cdot \nu_{g,k(i)}$, where:
  - s_i is the normalization factor of library i.
  - $k(i)$ is the treatment group of library i.
  - $\nu_{g,k(i)}$ is the normalized true mean expression level of gene g in the libraries of treatment group $k(i)$.

Normalization factors

- The normalization factors, s_i, account for differences in library sizes caused by different sequencing depths and other technical factors.
- Si and Liu (2012) show that the following method, proposed by Anders and Huber (2010), performs well:

$$s_i = \text{Median}_g \frac{y_{g,i}}{\left(\prod_{i=j}^{n} y_{g,j}\right)^{1/n}}$$

  where n is the total number of libraries.
- Note: to avoid dividing by zero, in practice, all zero counts are set to a small constant for this calculation.

Objectives

- Review current methods for estimating dispersion parameters in negative binomial models for RNA-Seq data
- Use a simulation study to evaluate and compare the effectiveness of these methods in terms of:
  - Point estimation quality.
  - The performance of tests to detect differentially expressed genes.

Outline

The quasi-likelihood (QL) method (Robinson and Smyth, 2007)

- Implementation: package `AMAP.Seq` (Si and Liu, 2012)
- Iteratively estimate:
  - The negative binomial MLE, $\widehat{\mu}_{g,i}$, of $\mu_{g,i}$, given $\phi_g = \widehat{\phi}_g$.
  - $\widehat{\phi}_g$, the quasi-likelihood tagwise dispersion estimate given $\mu_{g,i} = \widehat{\mu}_{g,i}$, which is calculated by solving for $\widehat{\phi_g}$:

$$2\sum_{i=1}^{n}\left\{y_{g,i}\log\left[\frac{y_{g,i}}{\widehat{\mu}_{g,i}}\right] - (y_{g,i}+\widehat{\phi}_g{}^{-1})\log\left[\frac{y_{g,i}+\widehat{\phi}_g{}^{-1}}{\widehat{\mu}_{g,i}+\widehat{\phi}_g{}^{-1}}\right]\right\} = n-1$$

The dispersion shrinkage for sequencing (DSS) method (Wu, Wang, and Wu, 2012)

- Idea: shrink $\widehat{\phi}_g$ towards a common *prior* instead of a common value or trend.
- Decompose the negative binomial into a Poisson-Gamma hierarchical model:

$$\begin{aligned} Y_{g,i} \mid \theta_{g,i} &\sim \text{Poisson}(\theta_{g,i} s_i) \\ \theta_{g,i} \mid \phi_g &\sim \text{Gamma}(\nu_{g,k(i)}, \phi_g) \\ \phi_g &\sim \text{log-normal}(m_0, \tau^2) \end{aligned}$$

- The marginal distribution of the $Y_{g,i}$'s is $\text{NB}(\mu_{g,i}, \phi_g)$, where $\mu_{g,i} = s_i \nu_{g,k(i)}$ as before.
- Each $\widehat{\phi}_g$ is the mode of the posterior density of ϕ_g.
- Implementation: package `DSS`

The weighted quantile-adjusted conditional maximum likelihood (wqCML) method (Robinson & Smyth, 2007)

- Implementation:
  - Package edgeR.
  - Use the `estimateTagwiseDisp()` function.
  - Optionally, set α with the `prior.n` argument.
- Maximize the weighted log likelihood:

$$\text{WLL}(\phi_g) = l_g(\phi_g) + \alpha l_C(\phi_g)$$

  - l_C: the "common" log likelihood, the negative binomial log likelihood under the restriction that all genes share the same dispersion value.
  - l_g: the log likelihood used in the quantile-adjusted conditional maximum likelihood method (qCML).
    - CML constructs a negative binomial likelihood for each $Y_{g,i}$ conditioned on $\sum_{k(j)=k(i)} Y_{g,j}$.
    - qCML modifies the CML method to account for unequal library sizes.
  - α: tuning parameter, typically calculated via empirical Bayes.

The Cox-Reid adjusted profile likelihood (APL) method (McCarthy, Chen, and Smyth, 2012)

- Apply a negative binomial generalized linear model:

$$\log \mu_{g,i} = \mathbf{x}_i^T \boldsymbol{\beta}_g + \log m_i$$

  - $\mathbf{x}_i^T$: vector of covariate values specifying the experimental conditions on library i
  - $\boldsymbol{\beta}_g$: parameter vector for gene g, which does not include ϕ_g.
  - m_i: total number of reads in library i.

- Cox-Reid adjusted profile likelihood (APL) of gene g:

$$\text{APL}_g(\phi_g) = l(\phi_g \mid y_{g,i}, \widehat{\boldsymbol{\beta}}_g) - \frac{1}{2} \log \det I_g$$

  - l: the log-likelihood function of the loglinear model.
  - I_g is the Fisher information matrix of $\boldsymbol{\beta}_g$.
  - The estimate, $\widehat{\boldsymbol{\beta}}_g$, of $\boldsymbol{\beta}_g$ is computed independently from ϕ_g using Fisher's scoring algorithm.

Three ways to estimate ϕ_g

- Common: Take $\widehat{\phi}_g = \widehat{\phi}$, the dispersion that maximizes the shared likelihood function:

$$\text{APL}_S(\phi) = \frac{1}{G}\sum_{g=1}^{G} APL_g(\phi)$$

- Trended
  - Model ϕ_g as a smooth function of average gene-wise read count.
  - Default method:
    - Divide the genes into bins by average read count.
    - Estimate a common dispersion for each bin as above.
    - Fit a spline curve through the estimated dispersions.
- Tagwise
  - Maximize the weighted likelihood:

$$APL_g(\phi_g) + G_0 APL_{S_g}(\phi_g)$$

  - APL_{S_g} is a local shared log likelihood function for gene g.
  - G_0 is the weight on APL_{S_g}.
  - $G_0 = 20/df$ is suitable, where df is the number of residual degrees of freedom used to estimate ϕ_g.

Implementation of the APL method

Package edgeR:

- Common: `estimateGLMCommonDisp()`
- Trended: `estimateGLMTrendedDisp()`
- Tagwise: `estimateGLMTagwiseDisp()`

The differential expression for sequence count data (DESeq) method (Anders and Huber, 2010)

- Reparameterize the negative binomial model in terms of the variance, $\sigma^2_{g,i}$:

$$Y_{g,i} \sim \text{NB}(\mu_{g,i}, \sigma^2_{g,i})$$
$$\mu_{g,i} = s_i \cdot \nu_{g,k(i)}$$
$$\sigma^2_{g,i} = \underbrace{\mu_{g,i}}_{\text{“shot noise”}} + \underbrace{s_i^2 \cdot \eta_{g,k(i)}}_{\text{“raw variance”}}$$

- $\eta_{g,k(i)}$ is called the *raw variance parameter*.
- After estimating the $\nu_{g,k(i)}$'s and the $\eta_{g,k(i)}$'s, calculate the $\sigma^2_{g,i}$'s solve for estimates of the per-gene, per-library dispersions, $\phi_{g,i}$, using:

$$\sigma^2_{g,i} = \mu_{g,i} + \mu^2_{g,i}\phi_{g,i}$$

and then pool the $\widehat{\phi}_{g,i}$'s within each gene to obtain the $\widehat{\phi}_g$'s.

Implementation of the DESeq method: package `DESeq`, function `estimateDispersions()`

- The `sharingMode` argument
  - `"gene-est-only"`: The $\eta_{g,k(i)}$'s are estimated pointwise.
  - `"fit-only"`: The $\widehat{\eta}_{g,k(i)}$'s calculated as smooth functions of the $\widehat{\nu}_{g,k(i)}$'s.
  - `"maximum"`: Each $\widehat{\eta}_{g,k(i)}$ is the maximum of the pointwise estimate and the estimate from the smooth function.
- The `fitType` argument:
  - `"parametric"`: the smooth functions are computed with a parametric regression.
  - `"local"`: the smooth functions are computed with a local regression.

Available DE Testing Methods

- `edgeR` exact test: A modified version of Fisher's exact test in the package, `edgeR`.
- `DESeq` exact test: another modified version of Fisher's exact test in the package, `DESeq`.
- `QuasiSeq` QL method: apply a GLM to the data, parameterize $Var(y_{g,i})$ as $(\mu_{g,i} + \phi_g \mu_{g,i}^2)\Phi_g$, and test for DE with a quasi-likelihood ratio test.
  - ϕ_g is still the negative binomial dispersion.
  - Φ_g is called the *generalized linear model (GLM) dispersion*.
- `QuasiSeq` QLShrink method: same as the QL method, except that information is shared across genes to estimate the Φ_g's.
- `QuasiSeq` QLSpline method: same as the QLShrink method, but estimates the GLM dispersions using a spline to account for the mean-variance relationship in RNA-Seq data.

Outline

Generating a Pseudo-dataset

Pick a real dataset, which has n libraries and counts $y_{g,i}$. Simulate the pseudo-counts $\widetilde{y}_{h,j}$ for pseudo-gene $h = 1, \ldots, 10000)$ and pseudo-library $j = 1, \ldots, \widetilde{n}$:

1. Randomly pick a gene g from the real dataset.
2. Compute the geometric mean of the counts for gene g:

$$\overline{y}_{g.} = \left(\prod_{i=1}^{n} y_{g,i} \right)^{1/n}$$

   where all zero counts are set to a small constant for the above calculation.
3. Randomly select pseudo-gene h to be either differentially expressed (DE) or equivalently expressed (EE). (In all, 20% of pseudo-genes are DE.)
4. For EE genes, $\delta_h = 0$. For DE genes, the δ_hs are multivariate normal with mean 0 and a random block-diagonal variance-covariance matrix.
5. For treatment levels $k = 1$ and 2, set true mean expression levels:

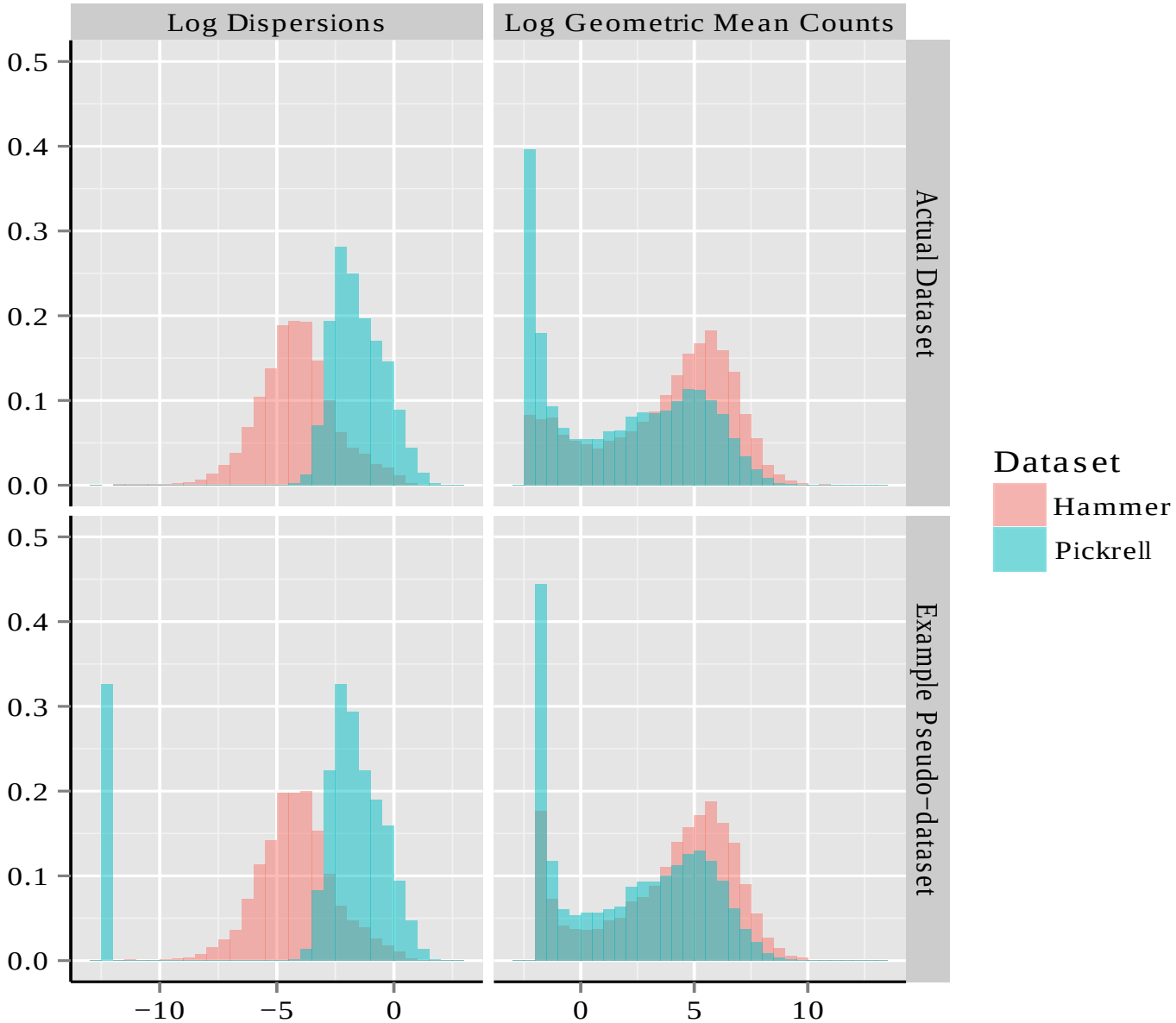
$$\nu_{h,k} = \overline{y}_{g.} \exp\left[(-1)^k \frac{\delta_h}{2} \right]$$

6. Simulate pseudocounts $\widetilde{y}_{h,j} \sim NB(\nu_{h,k(j)}, \widehat{\phi}_g)$, calculating $\widehat{\phi}_g$ from the real dataset using the QL Method.
7. If $\widetilde{y}_{h,1} = \cdots = \widetilde{y}_{h,\tilde{n}} = 0$, redraw gene g from the real dataset and return to step 1.

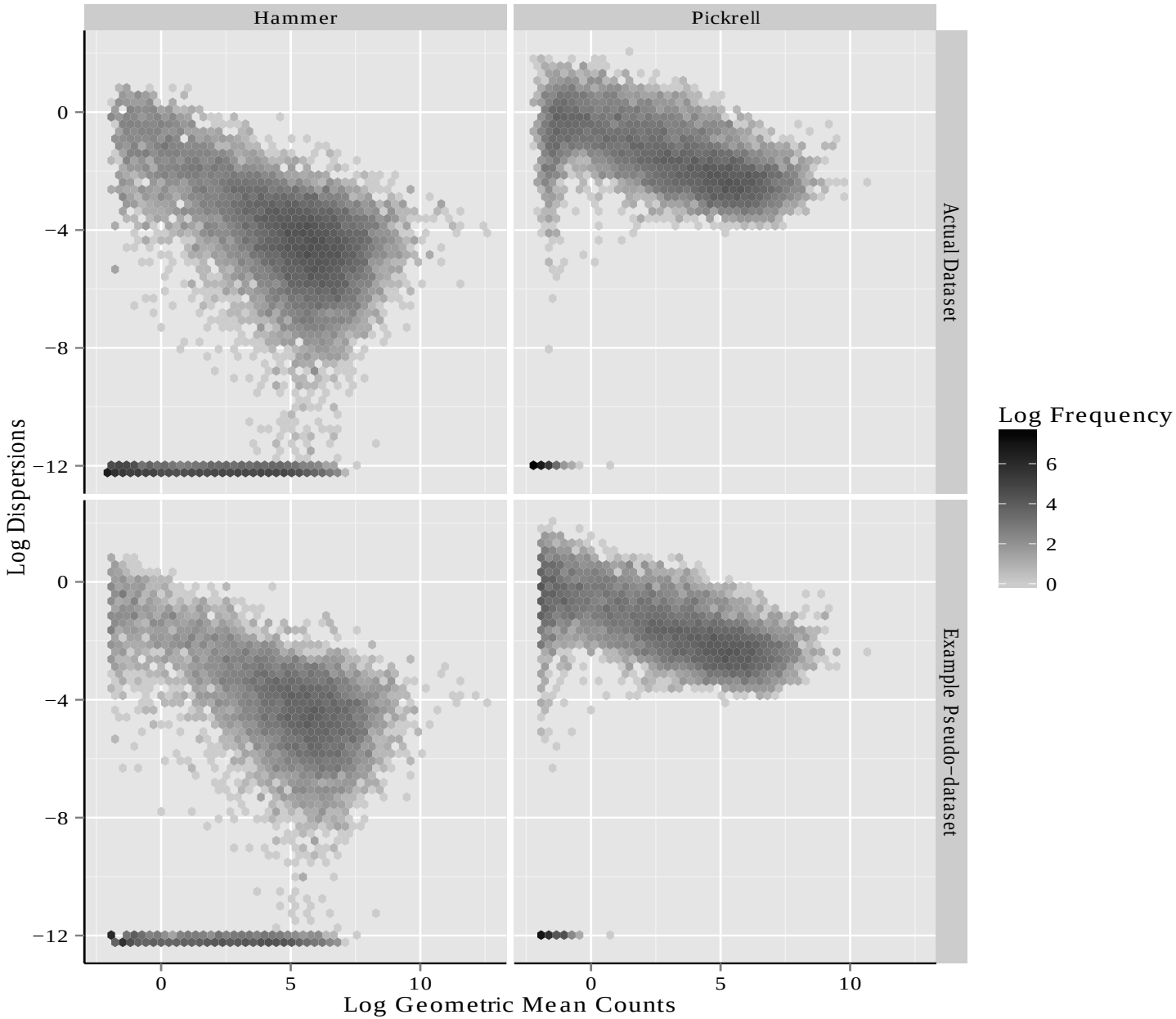
The Underlying Real Datasets

- "Hammer data" (Hammer, et al [4]). Data show gene expression in the L4 dorsal root ganglia in control rats and in those of rats with experimentally induced chronic neuropathic pain.
  - 18635 expressed genes.
- "Pickrell data" (Pickrell, et al. [10]). 69 lymphoblastoid cell lines derived from unrelated Nigerian individuals who were subjects in the International HapMap Project.
  - 12531 expressed genes.

The Underlying Real Datasets



The Underlying Real Datasets



Simulation Settings

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

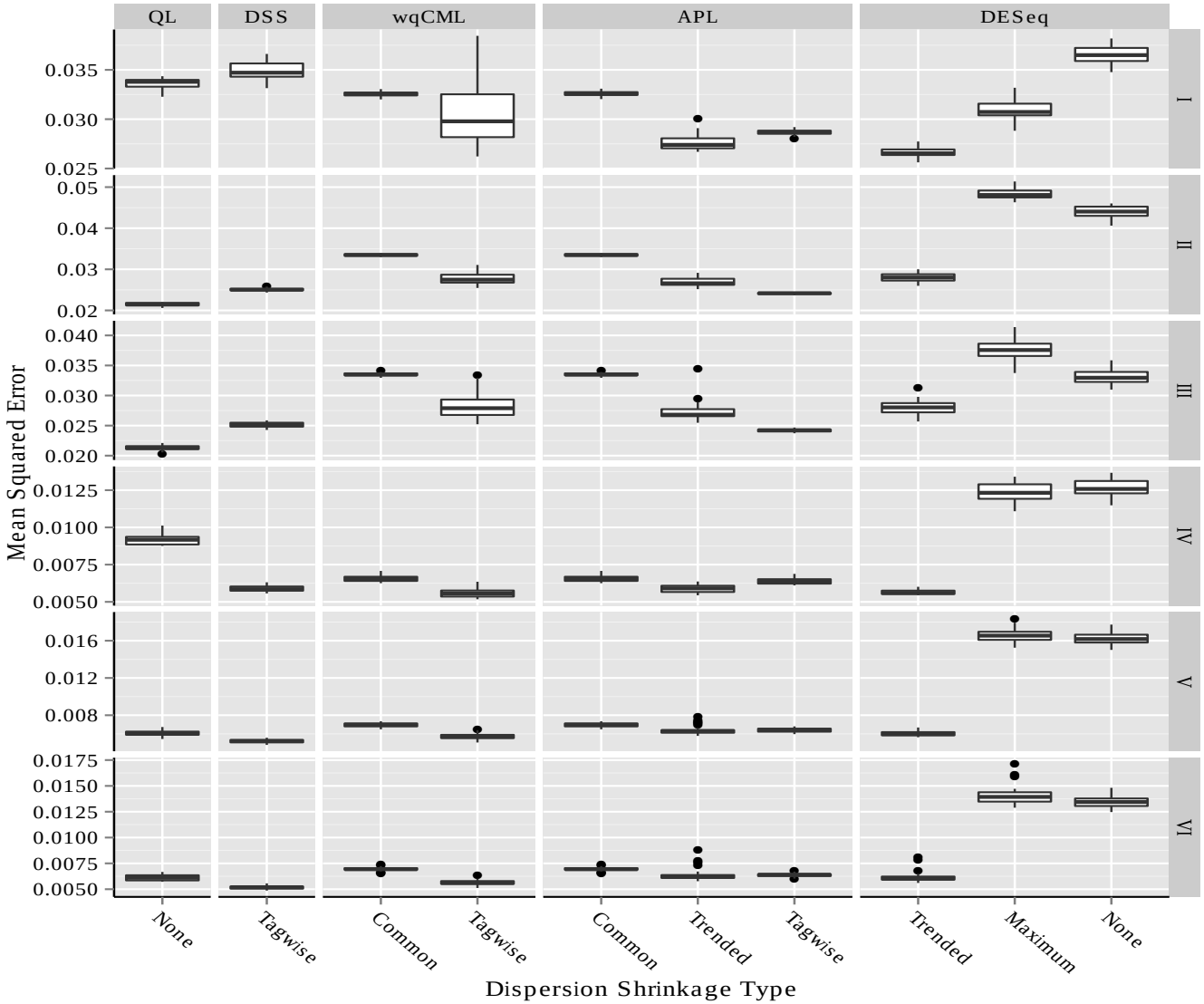
Outline

Mean Squared Error

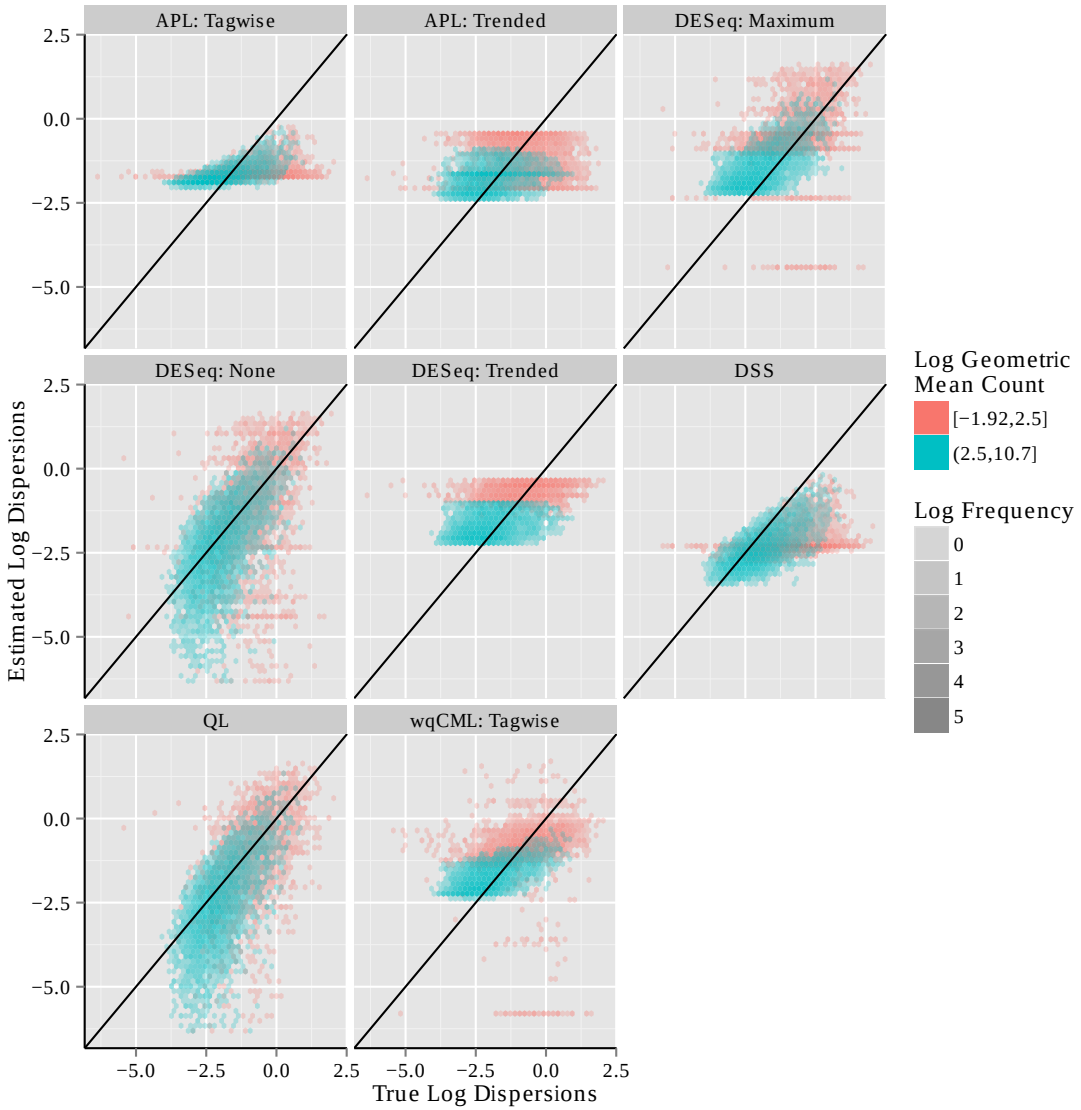
Mean squared error of adjusted dispersions:

$$\text{MSE} = \frac{1}{10000} \sum_{h=1}^{10000} \left[\frac{\widehat{\phi}_h}{1 + \widehat{\phi}_h} - \frac{\phi_h}{1 + \phi_h} \right]^2$$

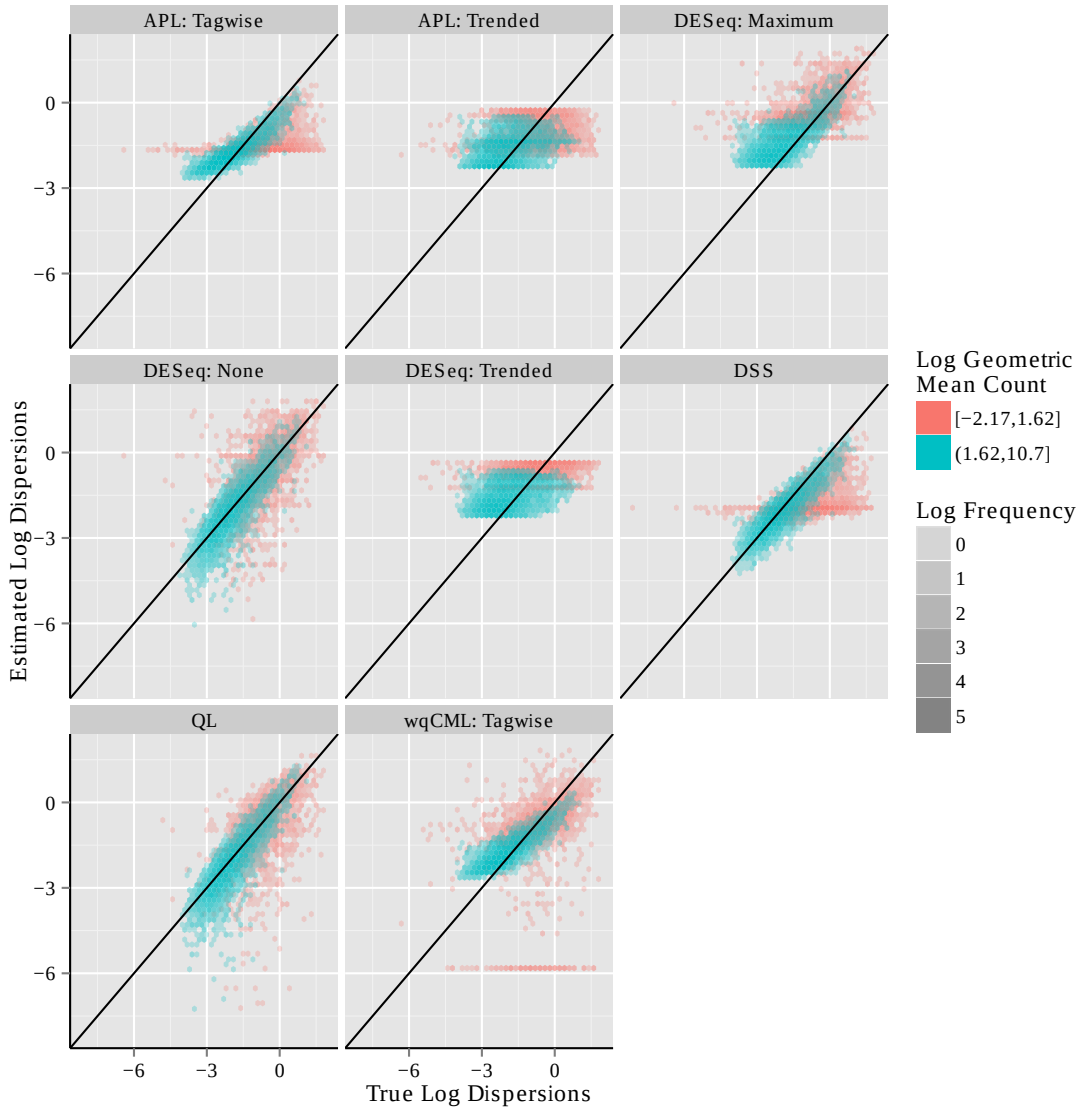
MSEs of the pseudo-datasets of simulation settings I, II, and III



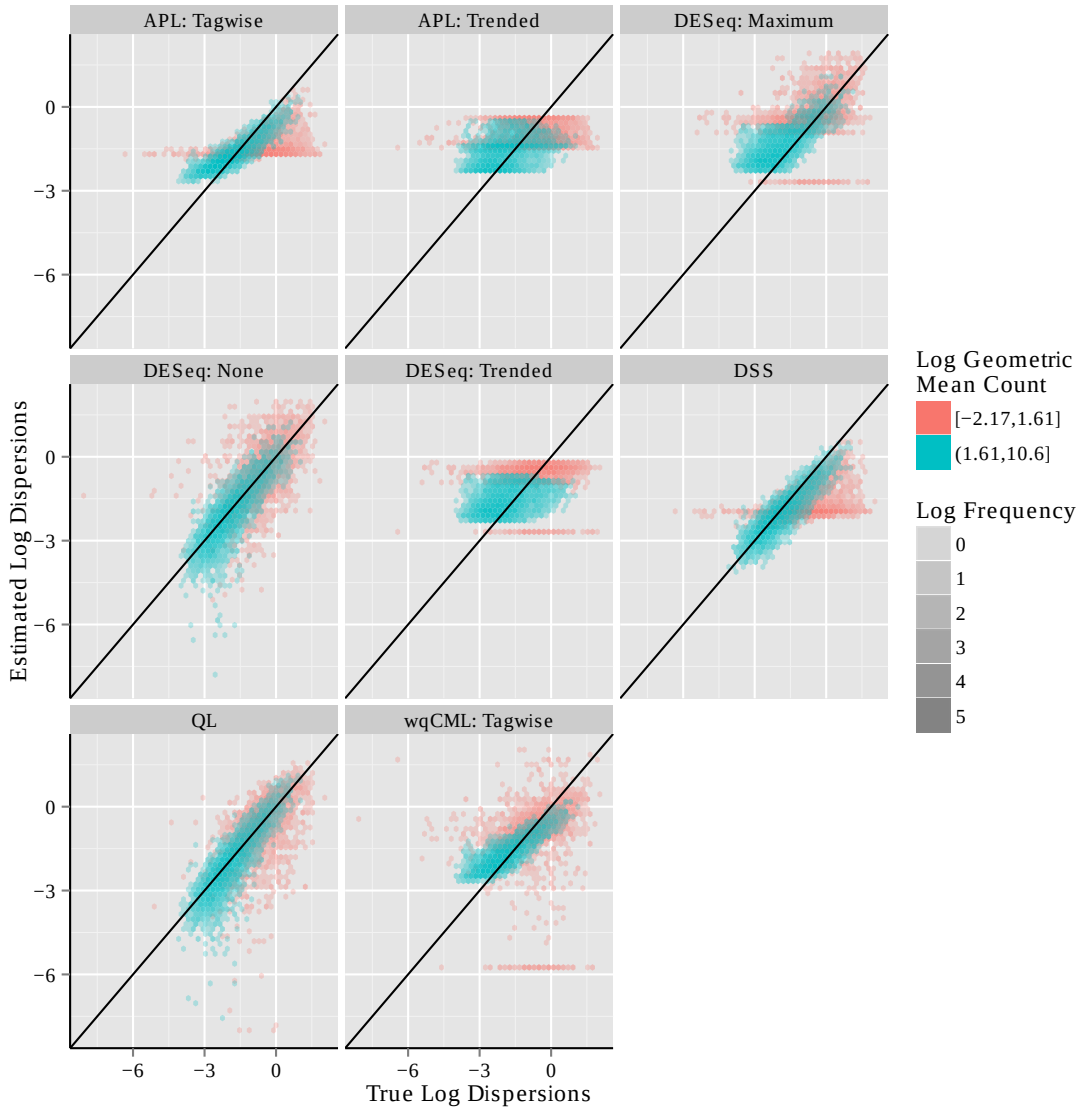
Estimated vs. True Dispersions: Simulation Setting I



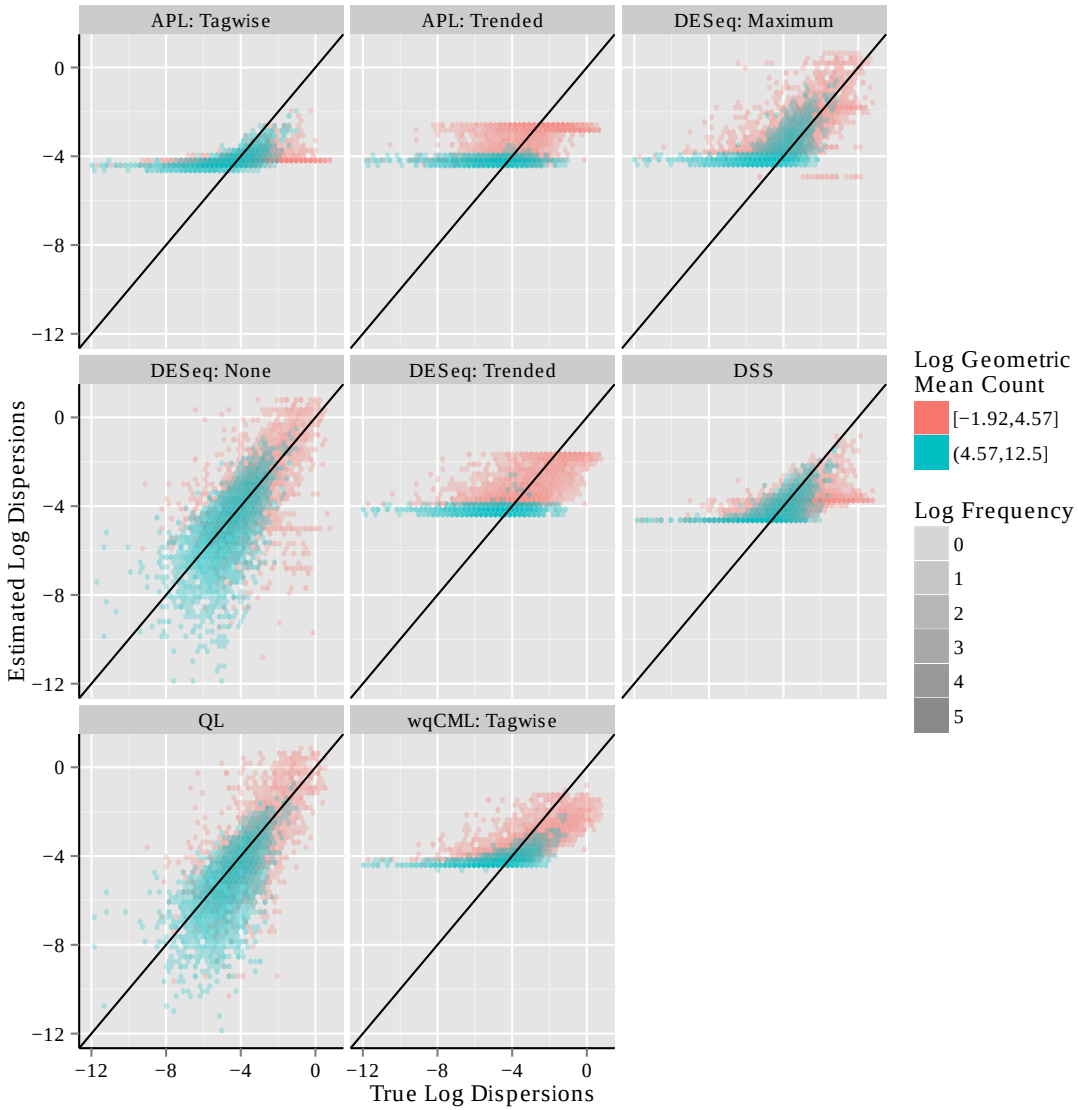
Estimated vs. True Dispersions: Simulation Setting II



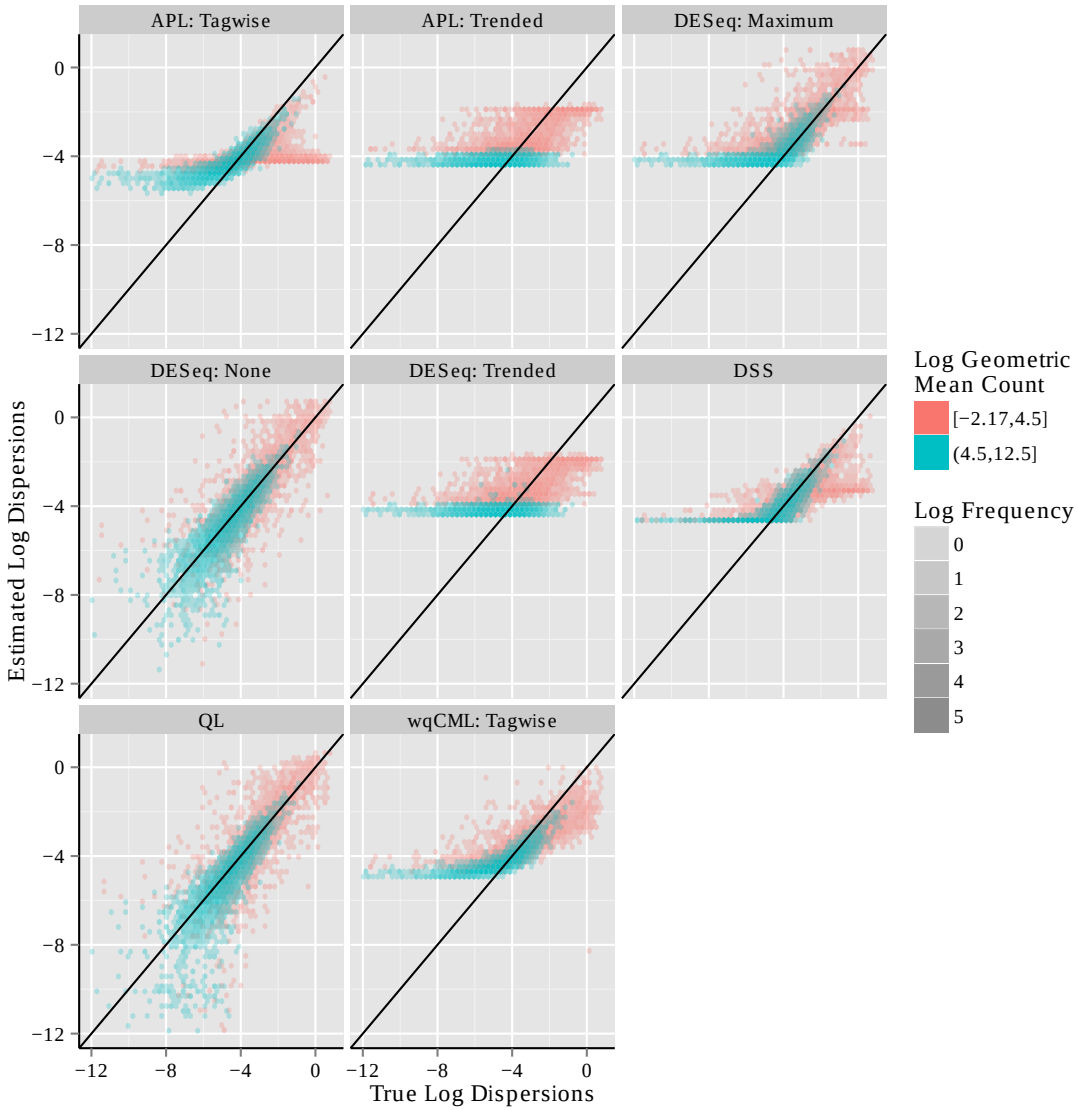
Estimated vs. True Dispersions: Simulation Setting III



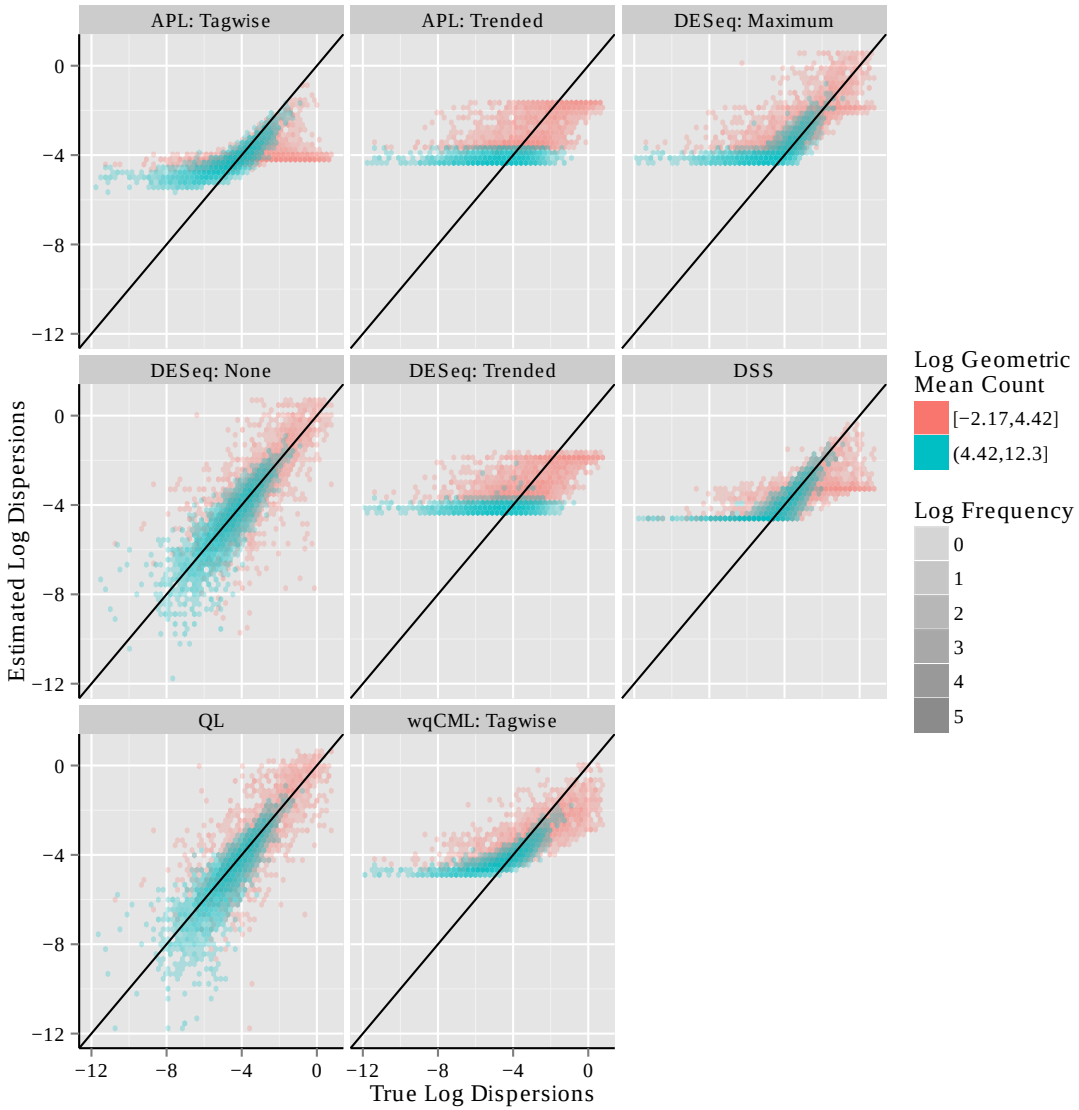
Estimated vs. True Dispersions: Simulation Setting IV



Estimated vs. True Dispersions: Simulation Setting V

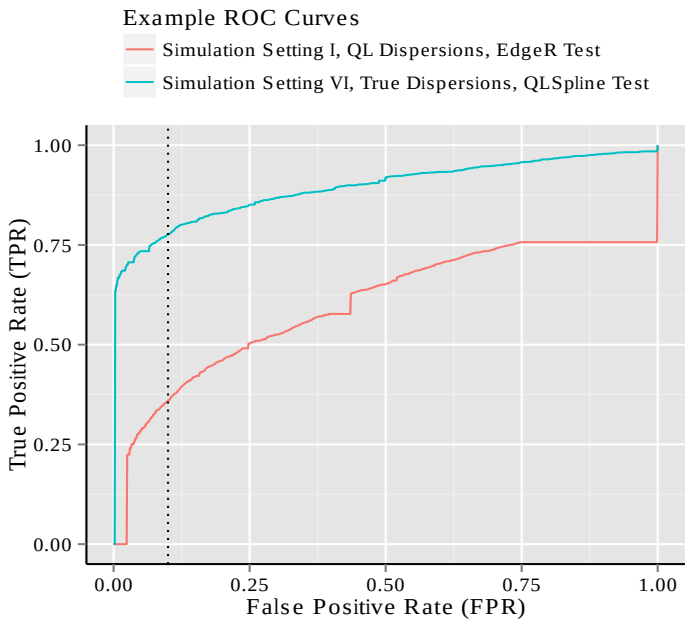


Estimated vs. True Dispersions: Simulation Setting VI

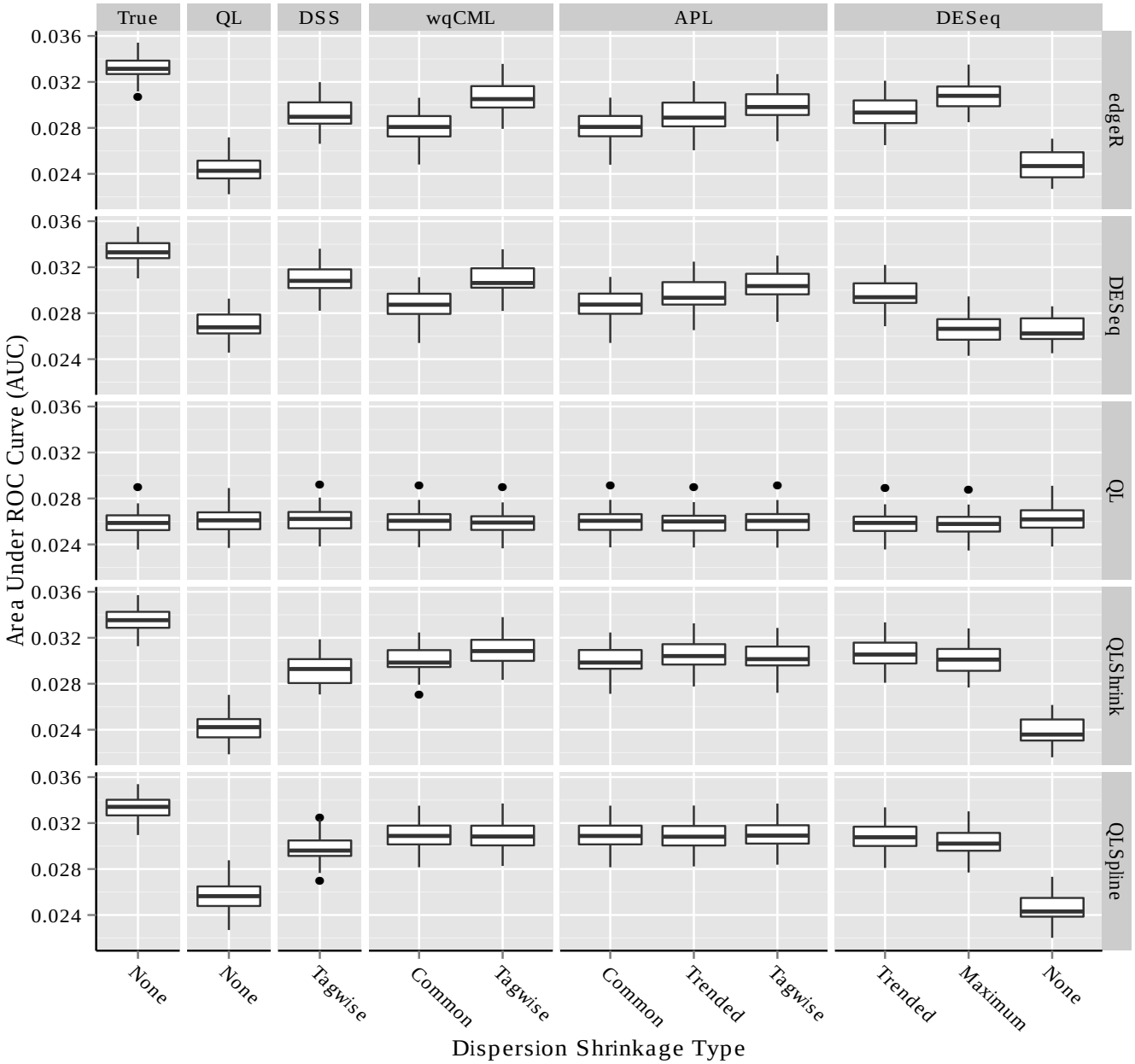


Effect of dispersion estimation method on DE test performance

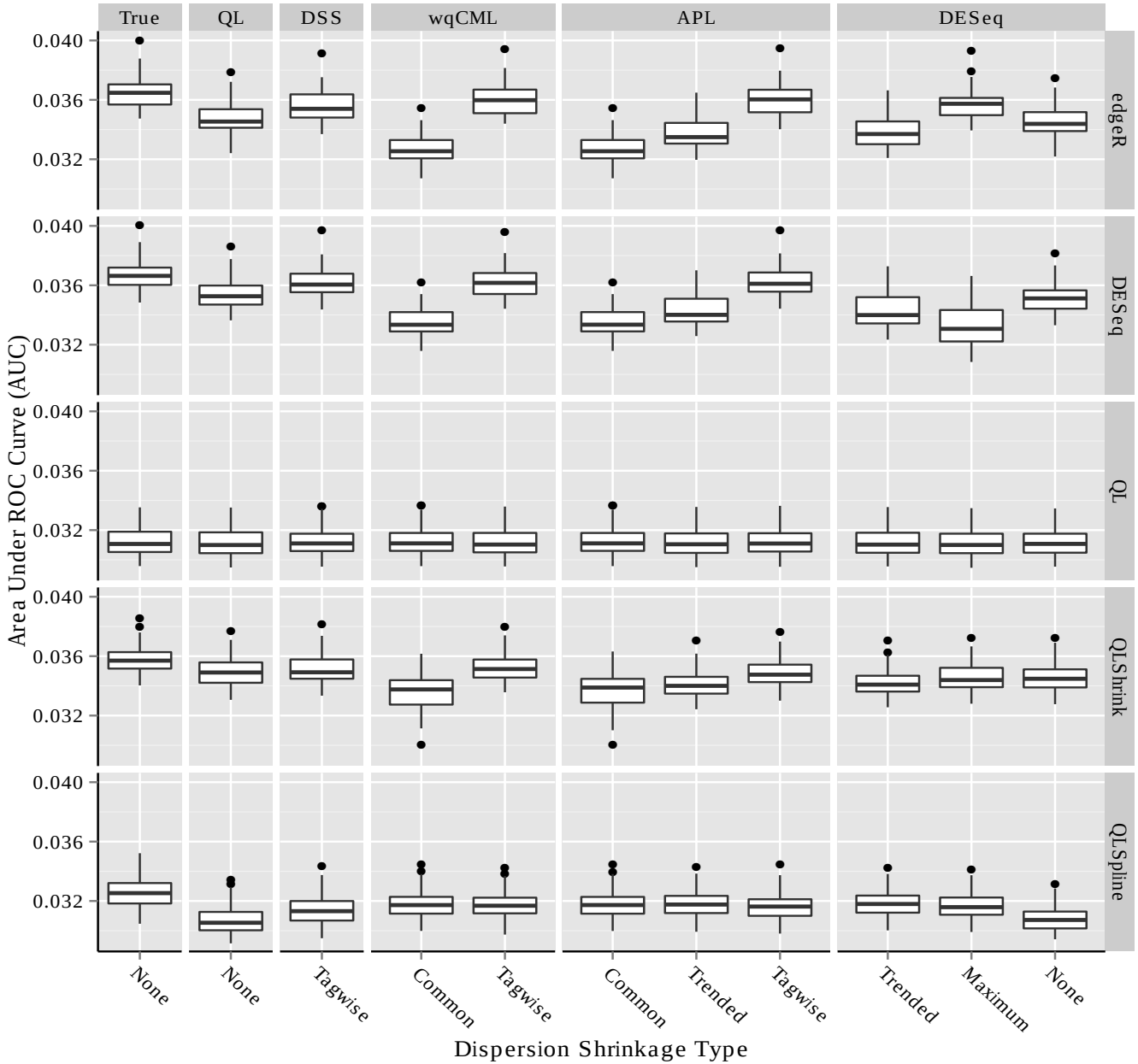
- Receiver operating characteristic (ROC) curve: graph of the true positive rate (TPR) of DE gene detection on the false positive rate (FPR) for several values of FPR from 0 to 1.
  - TPR: ratio of correctly identified DE genes to all the actually DE genes.
  - FPR ratio of genes incorrectly identified as DE to all the actually EE genes.
- The areas under the ROC curves (AUC) for FPR < 0.2 were plotted for each combination of simulation, dispersion, and test settings.



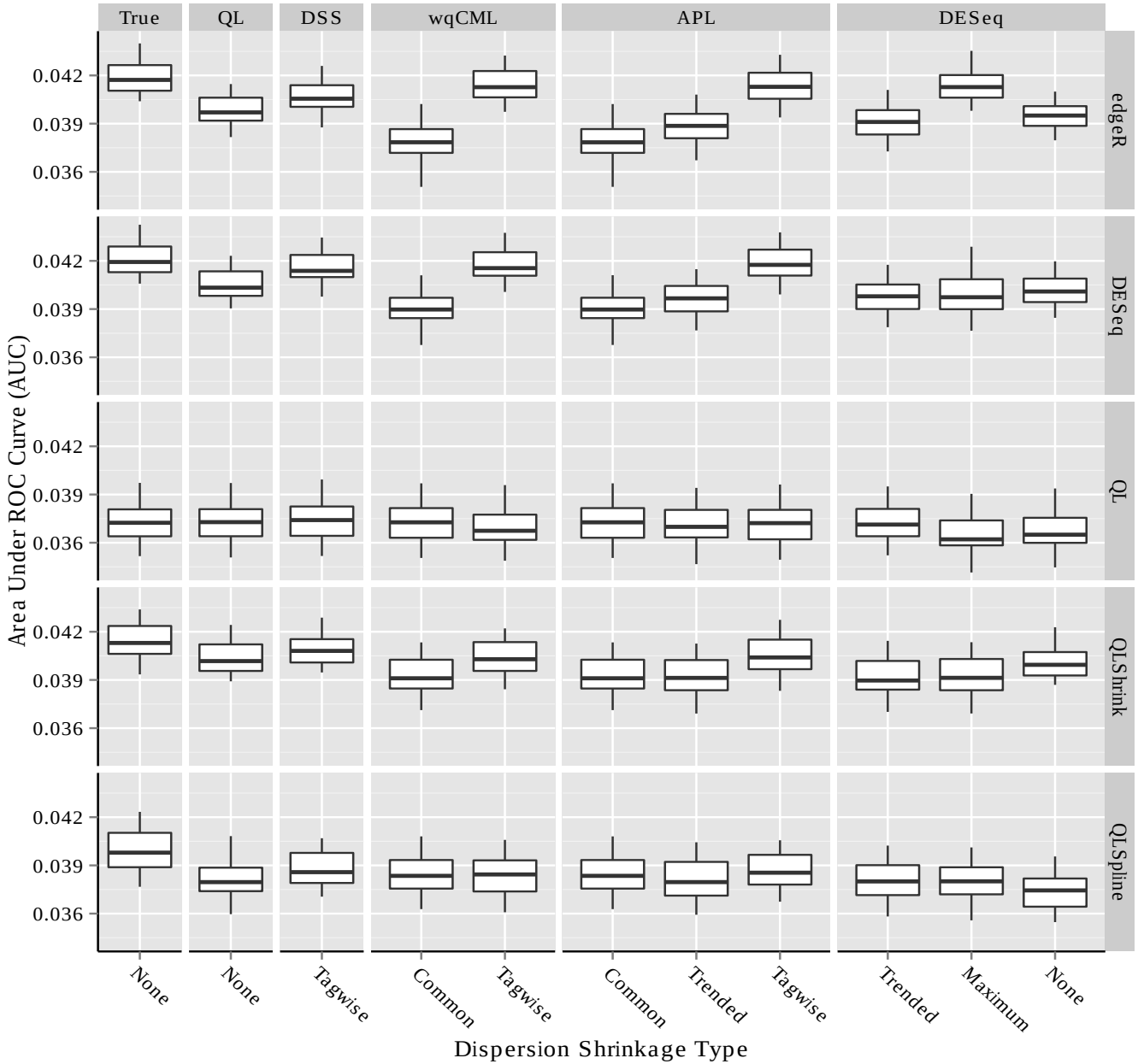
AUCs: Simulation Setting I



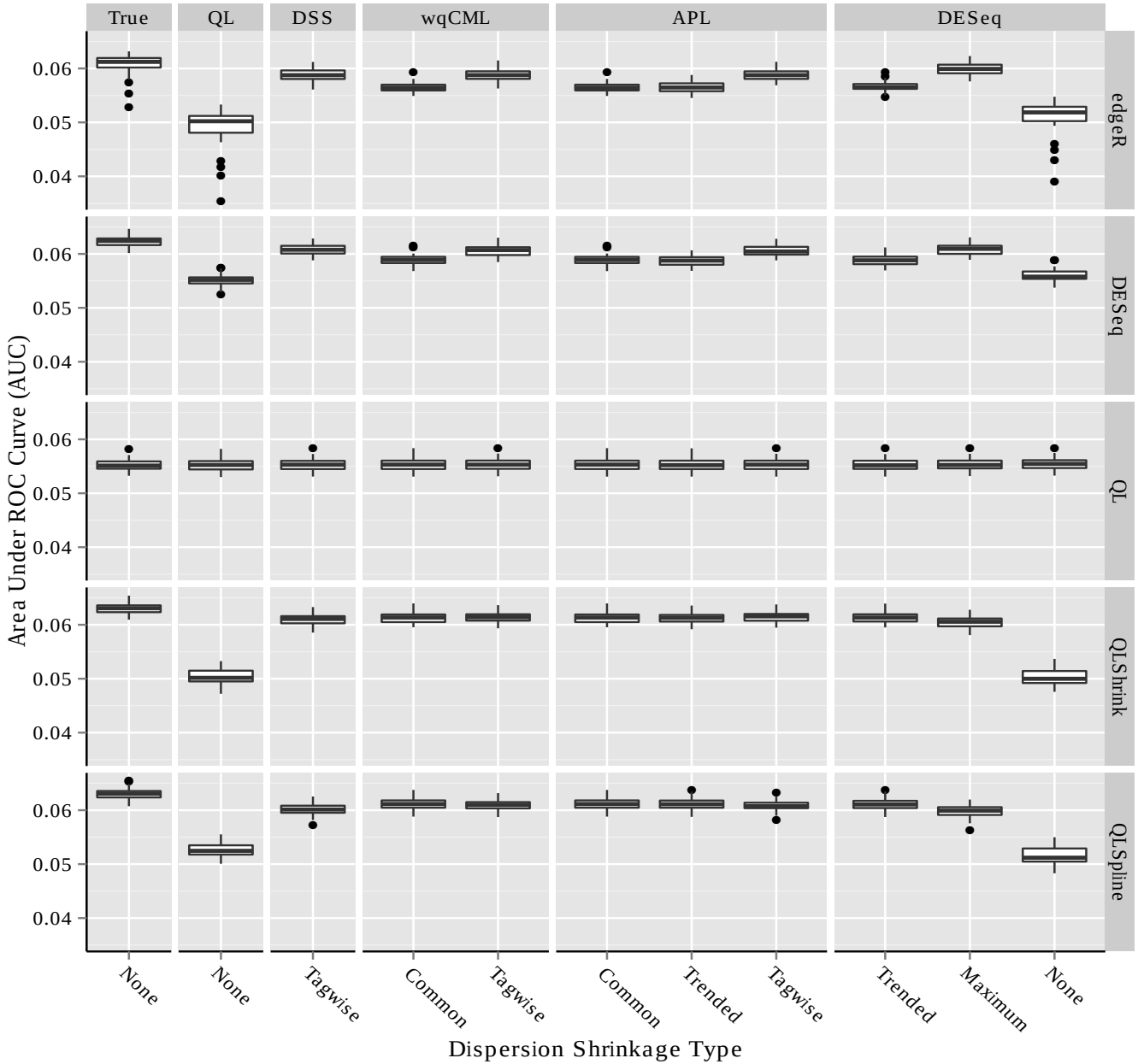
AUCs: Simulation Setting II



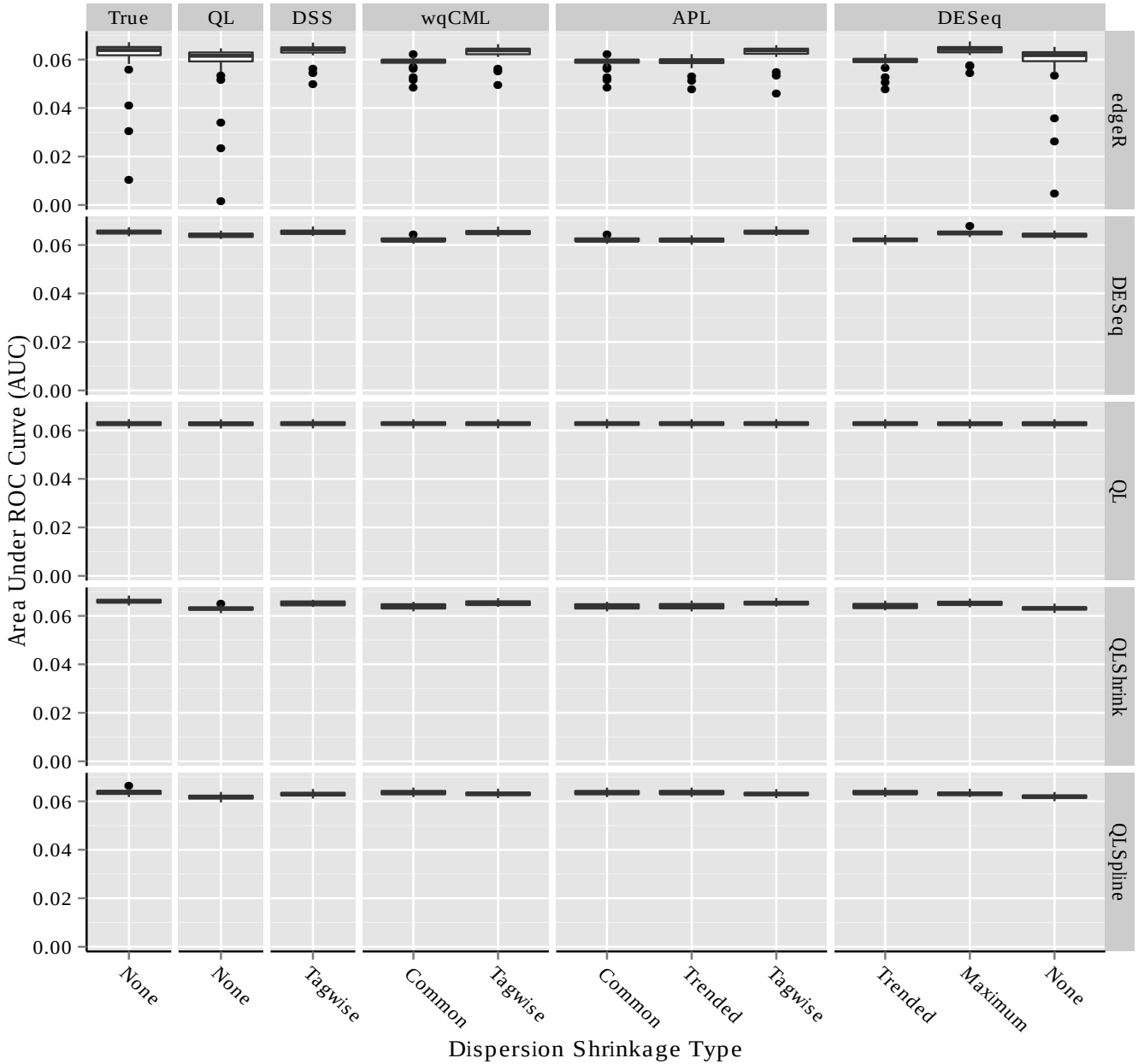
AUCs: Simulation Setting III



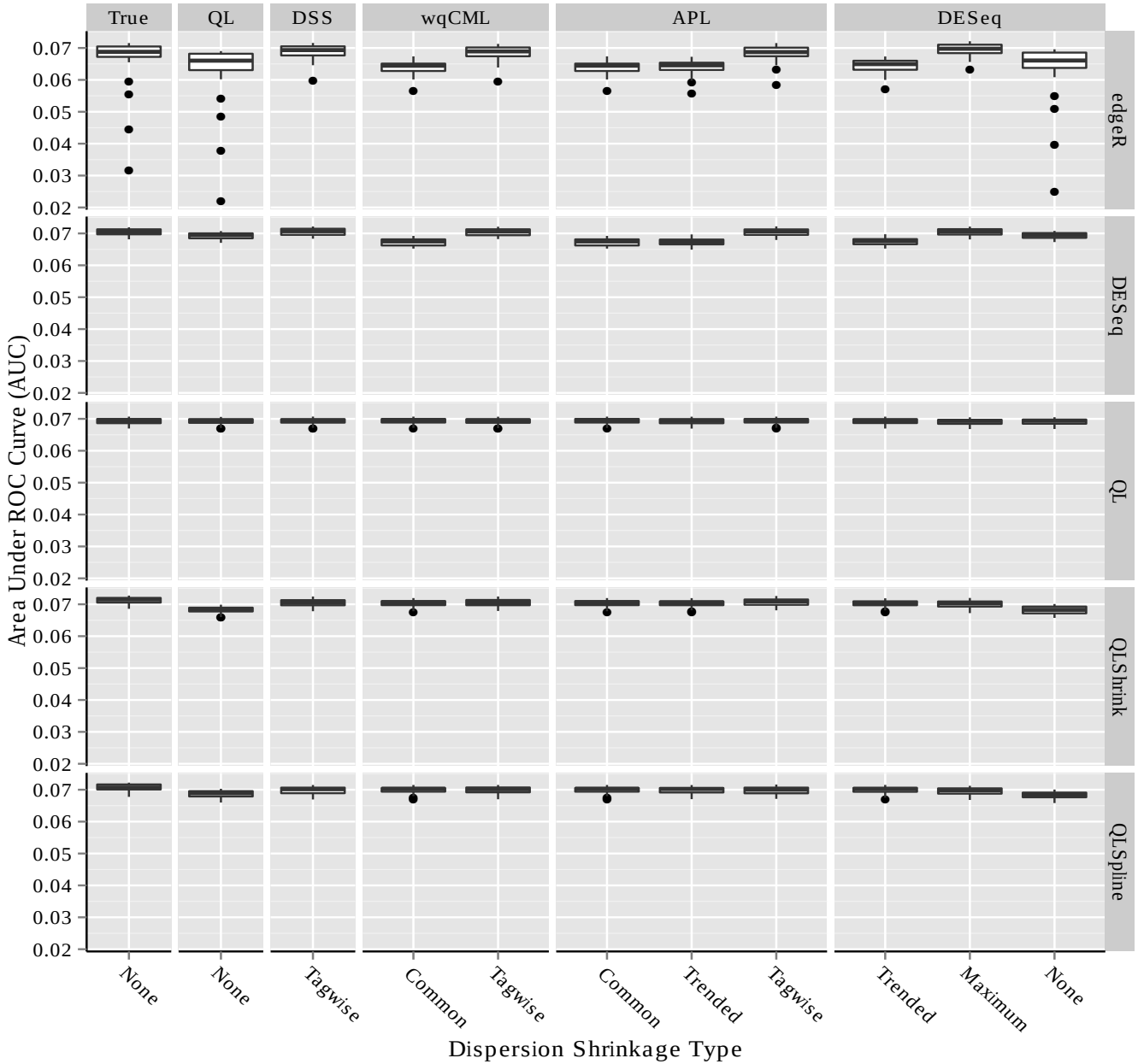
AUCs: Simulation Setting IV



AUCs: Simulation Setting V



AUCs: Simulation Setting VI



Outline

Conclusions

- Overall, the mean squared error-best methods are the ones that set the dispersions to a common trend (heavy shrinkage to a trend).
- The dispersions estimated independently for each gene (no shrinkage) have the strongest linear relationships with the true dispersions.
- The ones that maximize the performance of tests for differential expression are the ones that use a moderate degree of dispersion shrinkage, regardless of whether this shrinkage is toward a common value, trend, or prior distribution.

Exceptions

- Tagwise APL method is one of the mean squared error-best methods even though it is not one of the trended methods.
- The DSS and wqCML dispersions have strong linear relationships with the true dispersions for many of the Pickrell simulation settings despite the fact that these methods use some form of dispersion shrinkage.

Acknowledgements

We would like to thank Dr. Dan Nettleton, Dr. Dianne Cook, Dr. Long Qu, Emily King, Yet Tien Nguyen, and Fangfang Liu of Iowa State University for their useful feedback. We would also like to thank Dr. Gordon Smyth of the Walter and Eliza Hall Institute in Australia for answering our questions.

Sources I

1. S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10), October 2010.
2. A. C. Cameron and P. K. Travedi. *Regression Analysis of Count Data*. Cambridge University Press, 1998.
3. Robert C Gentleman, Vincent J. Carey, Douglas M. Bates, and others. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004. http://genomebiology.com/2004/5/10/R80.
4. P. Hammer, M. S. Banck, R. Amberg, et al. mrna-seq with agnostic splice site discovery for nervous system transcriptomics tested in chronic pain. *Genome Research*, 20(6):84760, June 2010.
5. S. Lund. Package QuasiSeq. http://cran.r-project.org/web/packages/QuasiSeq/QuasiSeq.pdf, June 2012.
6. S. P. Lund, D. Nettleton, D. J. McCarthy, and G. K. Smyth. Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Statistical Applications in Genetics and Molecular Biology*, 11(5), October 2012.
7. D. J. McCarthy, Y. Chen, and G. K. Smyth. Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. *Nucleic Acids Research*, 40:428897, 2012.

Sources II

8. P. McCullagh. Quasi-likelihood functions. *Annals of Statistics*, 11:5967, 1983.
9. A. Oshlack, M. D. Robinson, and M. D. Young. From RNA-seq reads to differential expression results. *Genome Biology*, 11(220), 2010.
10. J. K. Pickrell, J. C. Marioni, A. A. Pai, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):76872, April 2010.
11. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2012. http://www.R-project.org.ISBN3-900051-07-0.
12. M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3), 2010.
13. M. D. Robinson and G. K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):28812887, 2007.
14. M. D Robinson and G. K. Smyth. Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, 9(2): 321332, 2008.

Sources III

15. M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139140, October 2009.
16. M. D. Robinson, D. J. McCarthy, Y. Chen, and G. K. Smyth. Package edgeR. http://www.bioconductor.org/packages/2.10/bioc/manuals/edgeR/man/edgeR.pdf, September 2012.
17. Y. Si. Package AMAP.Seq. http://cran.r-project.org/web/packages/AMAP.Seq/AMAP.Seq.pdf, June 2012.
18. Y. Si and P. Liu. An optimal test with maximum average power while controlling FDR with application to RNA-seq data. Accepted, 2012.
19. L. Wang, P. Li, and T. P. Brutnell. Exploring plant transcriptomes using ultra high-throughput sequencing. *Briefings in Functional Genomics*, 9 (2):118128, 2010. doi: 10.1093/bfgp/elp057. http://bfg.oxfordjournals.org/content/9/2/118.abstract.
20. H. Wu, C. Wang, and Z. Wu. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics*, 1 (1):124, 2012.

Outline

Appendix: Dispersion Estimation Methods in Detail

- QL
- DSS
- wqCML
- APL
- DESeq

The quasi-likelihood (QL) method (Robinson and Smyth, 2007)

- Implementation: package `AMAP.Seq` (Si and Liu, 2012)
- Algorithm:
  1. Set $\widehat{\phi}_g$, the estimate of ϕ_g, to some initial value.
  2. Calculate the MLE, $\widehat{\mu}_{g,i}$, of each "true" unnormalized count mean, $\mu_{g,i}$, by maximizing the negative binomial log likelihood given $\phi_g = \widehat{\phi}_g$ and count $y_{g,i}$.
  3. Update $\widehat{\phi}_g$ to be the quasi-likelihood tagwise dispersion estimate given $\mu_{g,i} = \widehat{\mu}_{g,i}$ by solving for $\widehat{\phi_g}$:

$$2\sum_{i=1}^{n}\left\{y_{g,i}\log\left[\frac{y_{g,i}}{\widehat{\mu}_{g,i}}\right] - (y_{g,i}+\widehat{\phi}_g{}^{-1})\log\left[\frac{y_{g,i}+\widehat{\phi}_g{}^{-1}}{\widehat{\mu}_{g,i}+\widehat{\phi}_g{}^{-1}}\right]\right\} = n-1$$

  4. Iterate steps 2-4 a pre-determined number of times, each time using the most current value of $\widehat{\phi}_g$.

The dispersion shrinkage for sequencing (DSS) method (Wu, Wang, and Wu, 2012)

- Idea: shrink $\widehat{\phi}_g$ towards a common *prior* instead of a common value.
- Model:

$$\begin{aligned} Y_{g,i} \mid \theta_{g,i} &\sim \text{Poisson}(\theta_{g,i} s_i) \\ \theta_{g,i} \mid \phi_g &\sim \text{Gamma}(\nu_{g,k(i)}, \phi_g) \\ \phi_g &\sim \text{log-normal}(m_0, \tau^2) \end{aligned}$$

- The marginal distribution of the $Y_{g,i}$'s is $\text{NB}(\mu_{g,i}, \phi_g)$, where $\mu_{g,i} = s_i \nu_{g,k(i)}$ as before.
- Implementation: package `DSS`

Estimating ϕ_g

- Estimate hyperparameters m_0 and τ^2 using the method of moments estimates of ϕ_g.
- $\widehat{\nu}_{g,k(i)} = \frac{\sum_{j:k(j)=k(i)} Y_{g,j}/s_j}{n_{k(i)}}$, where $n_{k(i)}$ is the number of libraries in the same treatment group as replicate i.
- Set $\widehat{\mu}_{g,i} = s_i \widehat{\nu}_{g,k(i)}$ as before.
- Take $\widehat{\phi}_g$ to be the mode of the posterior density, $f(\phi_g \mid Y_{g,i}, \mu_{g,i}, i = 1, \ldots, n)$, given by:

$$
\begin{aligned}
&\log[f(\phi_g \mid Y_{g,i}, \mu_{g,i}, i = 1, \ldots, n)] \\
&\qquad \propto \sum_i \psi({\phi_g}^{-1} + Y_{g,i}) \\
&\qquad - n\psi({\phi_g}^{-1}) - {\phi_g}^{-1} \sum_i \log(1 + \mu_{g,i}\phi_g) \\
&\qquad + \sum_i Y_{g,i}[\log(\mu_{g,i}\phi_g) - \log(1 + \mu_{g,i}\phi_g)] \\
&\qquad - \frac{[\log(\phi_g) - m_0]^2}{2\tau^2} - \log(\phi_g) - \log(\tau)
\end{aligned}
$$

The weighted quantile-adjusted conditional maximum likelihood (wqCML) method (Robinson & Smyth, 2007)

- Implementation:
  - Package `edgeR`.
  - Use the `estimateTagwiseDisp()` function.
  - Set α with the `prior.n` argument.
- Maximize the weighted likelihood:

$$\mathrm{WL}(\phi_g) = l_g(\phi_g) + \alpha l_C(\phi_g)$$

  - l_C: the "common" log likelihood, the negative binomial log likelihood under the restriction that all genes share the same dispersion value.
  - l_g: the log likelihood given by quantile-adjusted conditional maximum likelihood (qCML).
  - α: tuning parameter, typically calculated via empirical Bayes.

Conditional maximum likelihood (CML): the basis for qCML

- Assume:
  - Each library has m total reads.
  - $Y_{g,1}, \ldots, Y_{g,n}$ are mutually independent.
- Then:
  - $Y_{g,i} \sim \text{NB}(m\nu_{g,k(i)}, \phi_g)$.
  - $Z_g = \sum_{i=1}^{n} Y_{g,i} \sim \text{NB}(nm\nu_{g,k(i)}, \phi_g)$
- CML selects the $\widehat{\phi}_g$ that maximizes the log likelihood of $Y_g = (Y_{g,1}, \ldots, Y_{g,n})$ conditioned on Z_g in terms of ϕ_g:

$$l_{Y_g|Z_g=z}(\phi_g) = \left[\sum_{i=1}^{n} \log\Gamma(y_{g,i} + {\phi_g}^{-1})\right] + \log\Gamma(n{\phi_g}^{-1}) - \log\Gamma(z + n{\phi_g}^{-1}) - \log\Gamma({\phi_g}^{-1})$$

Quantile-adjusted CML (qCML)

Now, assume $y_{g,i}$ is drawn from $Y_{g,i} \sim \text{NB}(m_i \nu_{g,k(i)}, \phi_g)$, where the library sizes, m_i, may be different. Calculate $\widehat{\phi}_g$:

1. Select the unadjusted CML dispersion as the starting value for $\widehat{\phi}_g$, temporarily assuming each $Y_i \sim NB(m^* \nu_{g,k(i)}, \phi)$, where $m^* = \left(\prod_{i=1}^{n} m_i\right)^{\frac{1}{n}}$.
2. Calculate $\widehat{\nu}_{g,k(i)}$, an estimate of $\nu_{g,k(i)}$, given $\phi_g = \widehat{\phi}_g$.
3. For $i = 1, \ldots, n$, calculate the probabilities:

$$p_{g,i} = P(Y_{g,i} < y_{g,i}) + \frac{1}{2} P(Y_{g,i} = y_{g,i})$$

4. Using a linear interpolation of the quantile function of the $\text{NB}(m^* \widehat{\nu}_{g,k(i)}, \widehat{\phi}_g)$ distribution, calculate the $\text{NB}(m^* \widehat{\nu}_{g,k(i)}, \widehat{\phi}_g)$ quantiles that correspond to the $p_{g,i}$'s. These interpolated quantiles are the pseudodata.
5. Set $\widehat{\phi}_g$ to be the CML estimate of ϕ_g using the pseudodata.
6. Repeat steps 2-5, each time using the most current dispersion estimate, $\widehat{\phi}_g$, until convergence.

The Cox-Reid adjusted profile likelihood (APL) method (McCarthy, Chen, and Smyth, 2012)

- Apply the negative binomial GLM:

$$\log \mu_{g,i} = \mathbf{x}_i^T \boldsymbol{\beta}_g + \log m_i$$

  - $\mathbf{x}_i^T$: vector of covariate values specifying the experimental conditions on library i
  - $\boldsymbol{\beta}_g$: parameter vector for gene g, which includes ϕ_g
  - m_i: total number of reads in library i.

- Cox-Reid adjusted profile likelihood (APL) of gene g:

$$\text{APL}_g(\phi_g) = l(\phi_g \mid y_{g,i}, \widehat{\boldsymbol{\beta}}_g) - \frac{1}{2} \log \det I_g$$

  - l: the log-likelihood function of the loglinear model.
  - I_g is the Fisher information matrix of $\boldsymbol{\beta}_g$.
  - The estimate, $\widehat{\boldsymbol{\beta}}_g$, of $\boldsymbol{\beta}_g$ is computed independently from ϕ_g using Fisher's scoring algorithm.

Three ways to estimate ϕ_g

- Common: Take $\widehat{\phi}_g = \widehat{\phi}$, the dispersion that maximizes the shared likelihood function:

$$\text{APL}_S(\phi) = \frac{1}{G}\sum_{g=1}^{G} APL_g(\phi)$$

- Trended
  - Model ϕ_g as a smooth function of average gene-wise read count.
  - Default method:
    - Divide the genes into bins by average read count.
    - Estimate a common dispersion for each bin as above.
    - Fit a spline curve through the estimated dispersions.
- Tagwise
  - Maximize the weighted likelihood:

$$APL_g(\phi_g) + G_0 APL_{S_g}(\phi_g)$$

  - APL_{S_g} is a local shared log likelihood function for gene g.
  - G_0 is the weight on APL_{S_g}.
  - $G_0 = 20/df$ is suitable, where df is the number of residual degrees of freedom used to estimate ϕ_g.

Implementation of the APL method

Package edgeR:

- Common: `estimateGLMCommonDisp()`
- Trended: `estimateGLMTrendedDisp()`
- Tagwise: `estimateGLMTagwiseDisp()`

The differential expression for sequence count data (DESeq) method (Anders and Huber, 2010)

- Reparameterize the negative binomial model in terms of the variance, $\sigma^2_{g,i}$:

$$Y_{g,i} \sim \text{NB}(\mu_{g,i}, \sigma^2_{g,i})$$
$$\mu_{g,i} = s_i \cdot \nu_{g,k(i)}$$
$$\sigma^2_{g,i} = \underbrace{\mu_{g,i}}_{\text{“shot noise”}} + \underbrace{s_i^2 \cdot \eta_{g,k(i)}}_{\text{“raw variance”}}$$

- $\eta_{g,k(i)}$ is called the *raw variance parameter*.
- After estimating the variance, solve for estimates of the per-gene, per-library dispersions, $\phi_{g,i}$, using:

$$\sigma^2_{g,i} = \mu_{g,i} + \mu^2_{g,i}\phi_{g,i}$$

and then pool the $\widehat{\phi}_{g,i}$'s within each gene to obtain the $\widehat{\phi}_g$'s.

To estimate $\sigma^2_{g,i}$, it suffices to estimate $\nu_{g,k(i)}$, and $\eta_{g,k(i)}$

- $\nu_{g,k(i)}$:

$$\widehat{\nu}_{g,k(i)} = \frac{1}{n_{k(i)}} \sum_{j:k(j)=k(i)} \frac{y_{g,i}}{s_i}$$

where $n_{k(i)}$ is the number of replicates with the same treatment group as replicate i.

- For $\eta_{g,k(i)}$, define:

$$w_{g,k(i)} = \frac{1}{n_{k(i)} - 1} \sum_{j:k(j)=k(i)} \left(\frac{y_{g,i}}{s_i} - \widehat{\nu}_{g,k(i)} \right)^2$$

$$z_{g,k(i)} = \frac{\widehat{\nu}_{g,k(i)}}{n_{k(i)}} \sum_{j:k(j)=k(i)} \frac{1}{s_i}$$

$w_{g,k(i)} - z_{g,k(i)}$ is an unbiased estimator of $\eta_{g,k(i)}$.

Implementation of the DESeq method: package `DESeq`, function `estimateDispersions()`

- The `sharingMode` argument (smooth functions $h_{k(i)}()$ determined by `fitType`)
  - `"gene-est-only"`: $\eta_{g,k(i)}$ estimated by $w_{g,k(i)} - z_{g,k(i)}$
  - `"fit-only"`: $\widehat{\eta}_{g,k(i)} = h_{k(i)}(\widehat{\nu}_{g,k(i)}) - z_{g,k(i)}$
  - `"maximum"`: $\widehat{\eta}_{g,k(i)} = \max(w_{g,k(i)}, h_{k(i)}(\widehat{\nu}_{g,k(i)})) - z_{g,k(i)}$
- The `fitType` argument:
  - `"parametric"`: the $h_{k(i)}()$'s are computed with a parametric regression of $w_{g,k(i)}$ on $\widehat{\nu}_{g,k(i)}$.
  - `"local"`: the $h_{k(i)}()$'s are computed with a local regression of $w_{g,k(i)}$ on $\widehat{\nu}_{g,k(i)}$.

Implementation of the DESeq method: package `DESeq`, function `estimateDispersions()`

- The `method` argument:
  - `"pooled"`: pool the $\widehat{\phi}_{g,i}$'s within each gene to obtain the $\widehat{\phi}_g$'s.
  - `"pooled-CR"`: use the APL method to calculate the pooled $\widehat{\phi}_g$'s.
  - `"per-condition"`: estimate a dispersion for each gene and each treatment level.
  - `"blind"`: pool the $\widehat{\phi}_{g,i}$'s to calculate the $\widehat{\phi}_g$'s as if all libraries were in a single treatment group.