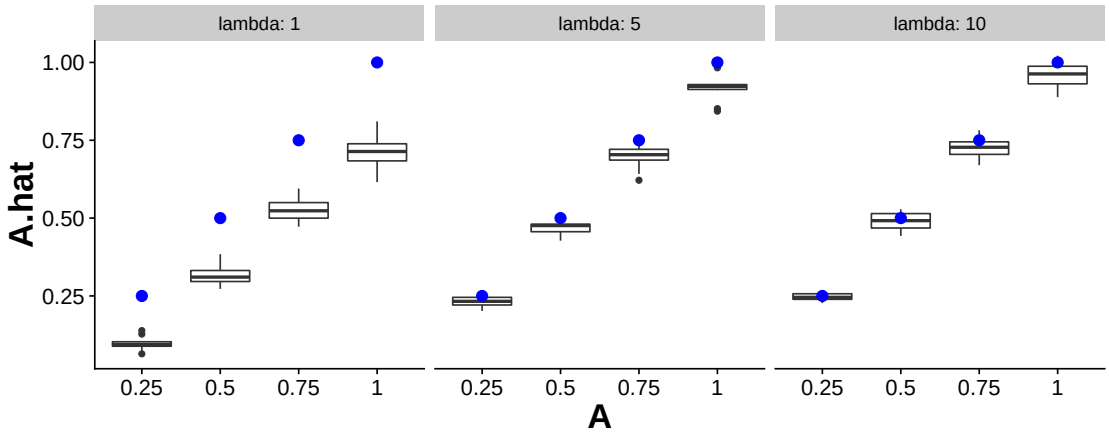
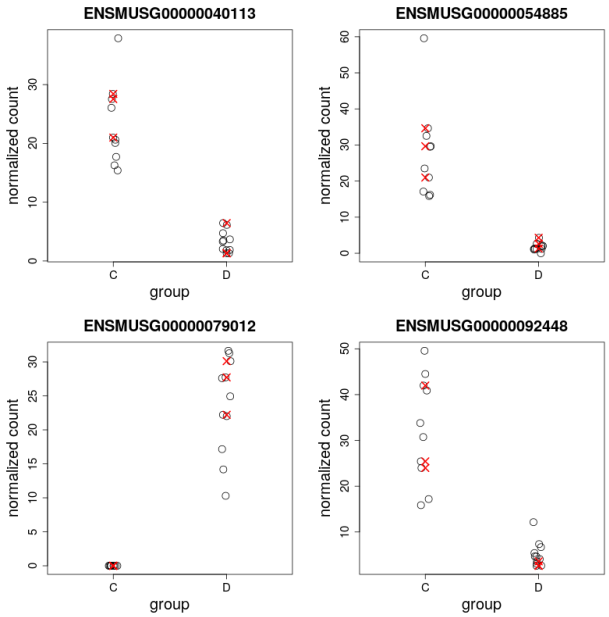
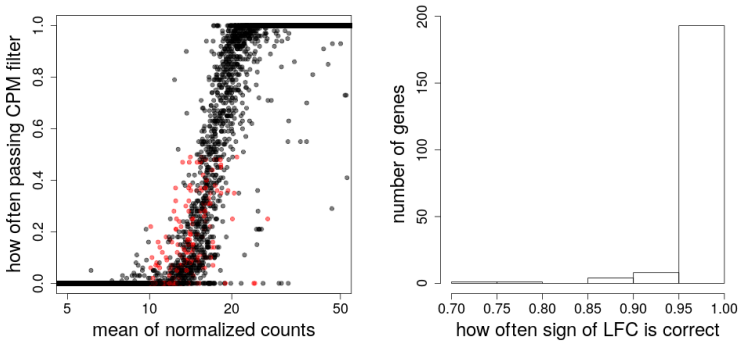
Supplementary Figures



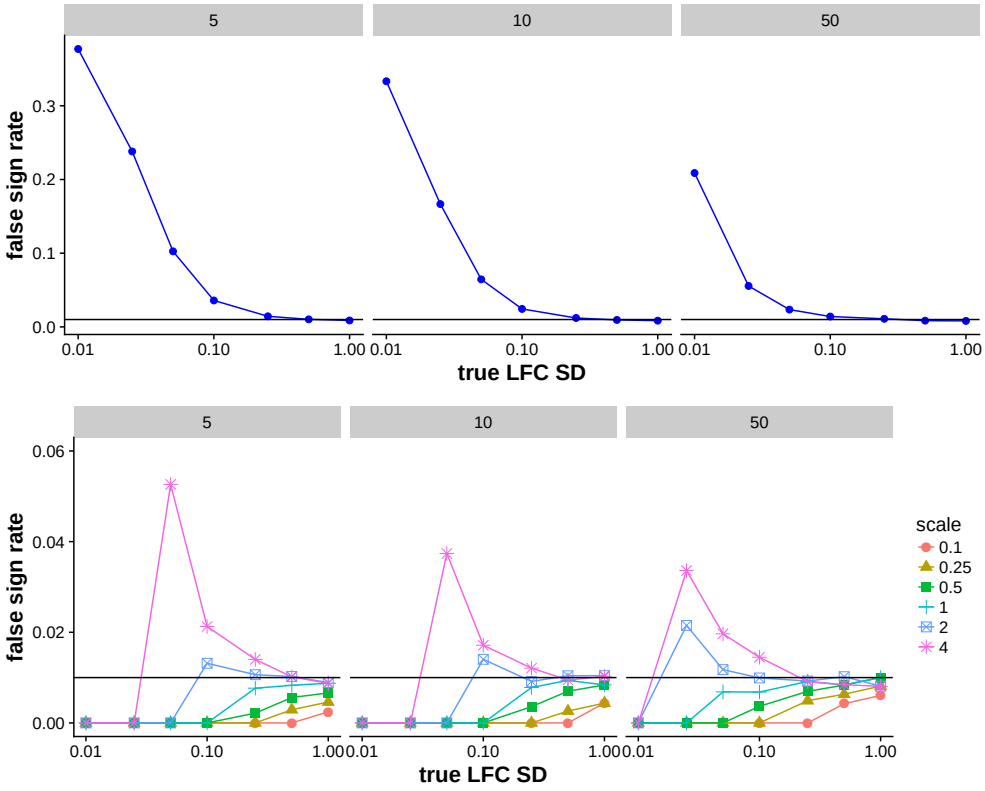
Supplementary Figure 1: Effects of estimating A when counts are very low. On the x-axis is the true value of A across simulations, and on the y-axis is the estimate $\hat{A}$ using the method implemented in *apeglm*. The boxes denote the result of 10 iterations, and the blue dot denotes the true value A. In the Methods, the estimated standard errors are treated as known in order to estimate the scale of a Normal distribution from which the true LFCs may have originated. We simulated n=5 vs 5 Poisson counts, with a baseline rate λ either set to 1, 5, or 10. We then simulated an LFC between the two groups, and 1,000 such "genes" were simulated. The LFC was drawn from a Normal distribution with mean 0 and variance A, with $A \in \{0.25, 0.5, 0.75, 1\}$. Our estimate $\hat{A}$ using the formula proposed by Efron and Morris [1] assuming known variances had negative bias for $\lambda = 1$ (although it roughly tracks the trend), but performed well for $\lambda = 5$ or $\lambda = 10$, where it only slightly underestimated A.



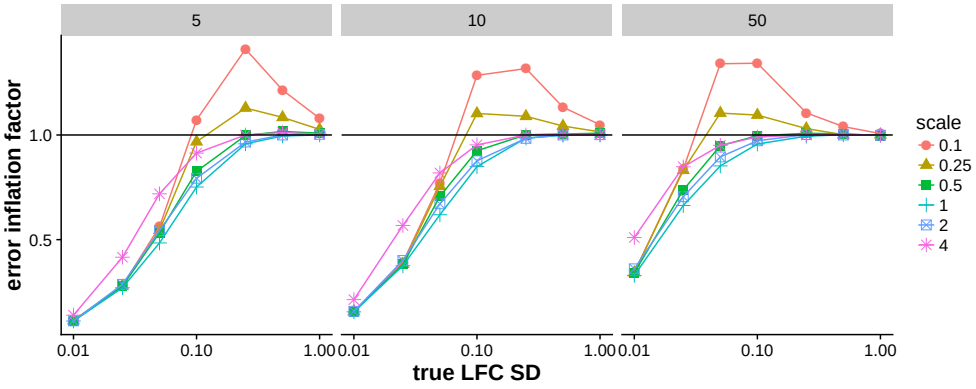
Supplementary Figure 2: The scaled counts for four example genes in the Bottomly et al. [2] dataset. These four genes were filtered more than 50% of the time by a CPM filtering rule applied to 3 vs 3 samples, but were reported as DE by *DESeq2* in the full dataset, with E(FDR) $< 5\%$. The random subsets were balanced with respect to three batches and the full analysis controlled for batch in the design. The red X's are examples of the scaled counts for one random subset, where this gene would be removed by the CPM rule, which requires ≥ 3 samples with CPM greater than the CPM value for a raw count of 10 for the least sequenced sample.



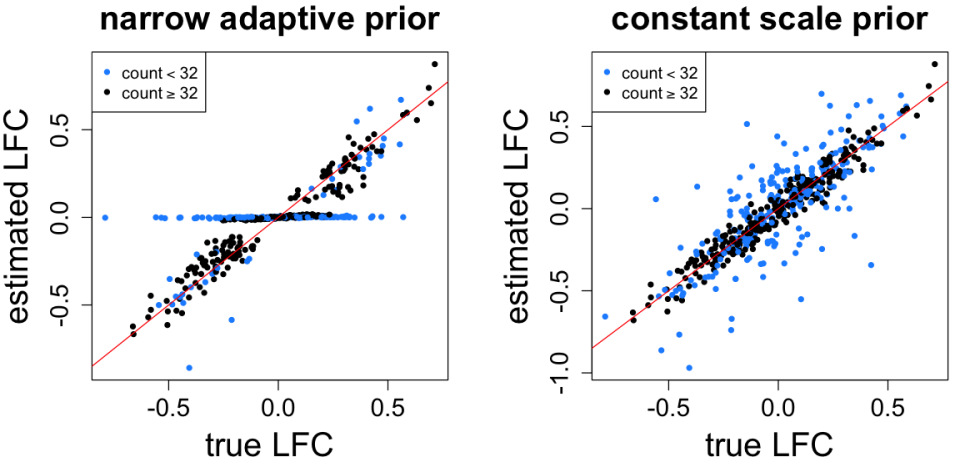
Supplementary Figure 3: Details on 207 filtered genes in the Bottomly et al. [2] dataset. Shown in red are 207 genes (left) that were filtered more than 50% of the time by a CPM filtering rule in a 3 vs 3 random subset, which also have a mean of scaled counts in the full dataset greater than 10 and with E(FDR) $< 5\%$ on the full dataset. The histogram (right) indicates that these genes often had the correct sign of log fold change in random subsets (99% on average), indicating there was meaningful signal for these filtered genes.



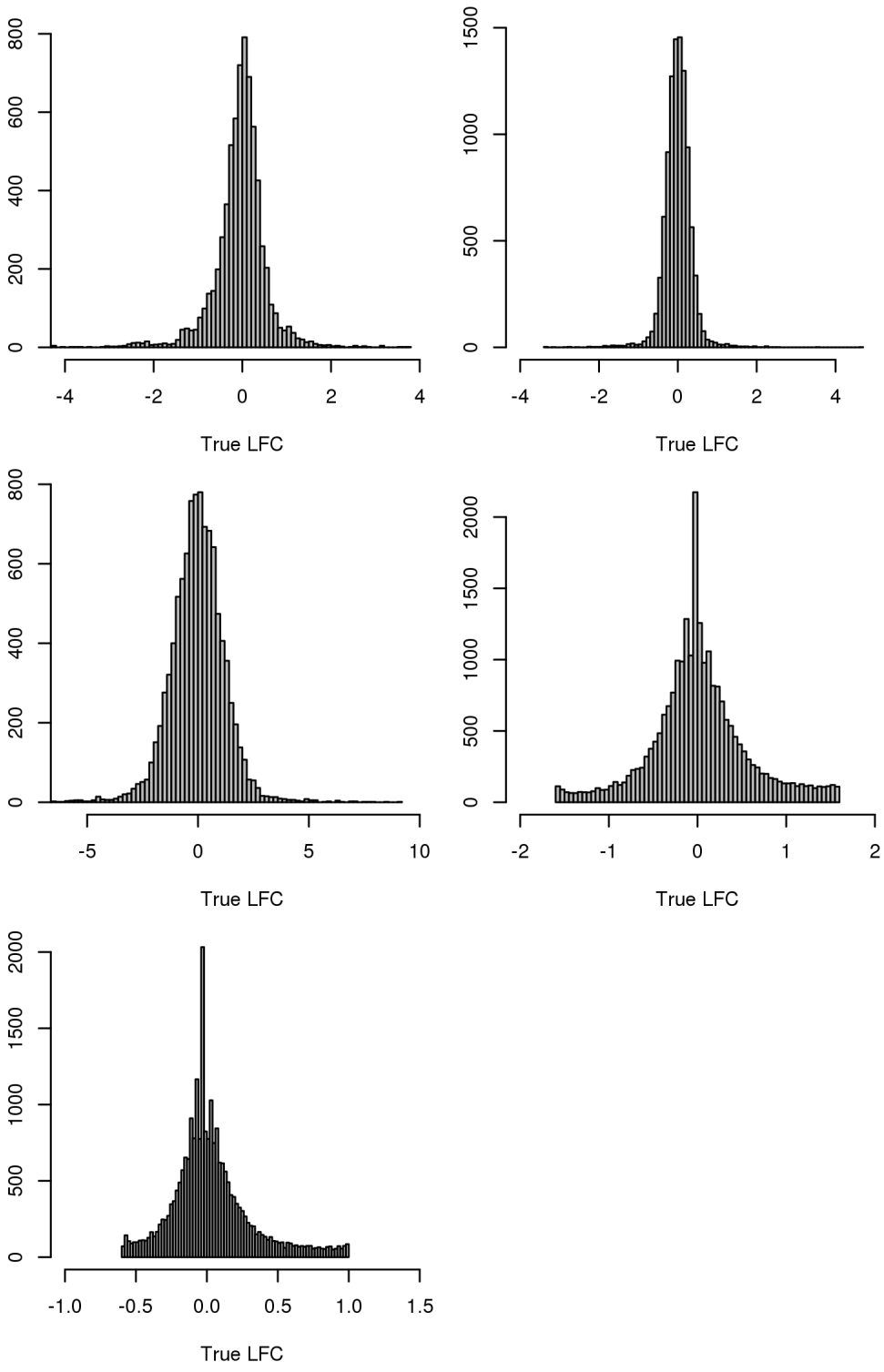
Supplementary Figure 4: Achieved false sign rate on a simulation dataset. The top row represents using a fixed scaled (S=1) for the prior in *apeglm* while the bottom row represents scaling the prior to match the distribution of true LFCs, with various *multipliers* tested in the range $[0.1, 4]$. The three columns indicate different per-group sample sizes in $\{5, 10, 50\}$. The points indicate the median over 100 iterations of the simulation. Here, the standard deviation (SD) of the distribution of true LFCs was provided as a *oracle* estimate to *apeglm* for setting the scale of the prior (*multiplier* * *true LFC SD*), whereas standard *apeglm* usage and for all other evaluations presented here the scale of the prior is estimated from the LFC MLE and their standard errors (Methods). The top row indicates that a fixed prior (S=1) did not control the FSR when the true LFC distribution was very narrow. The bottom row indicates that when the prior was 2 or 4 times the scale of the true distribution (blue square, pink asterisk), there was also some loss of FSR control, although to a lesser degree than using a fixed prior.



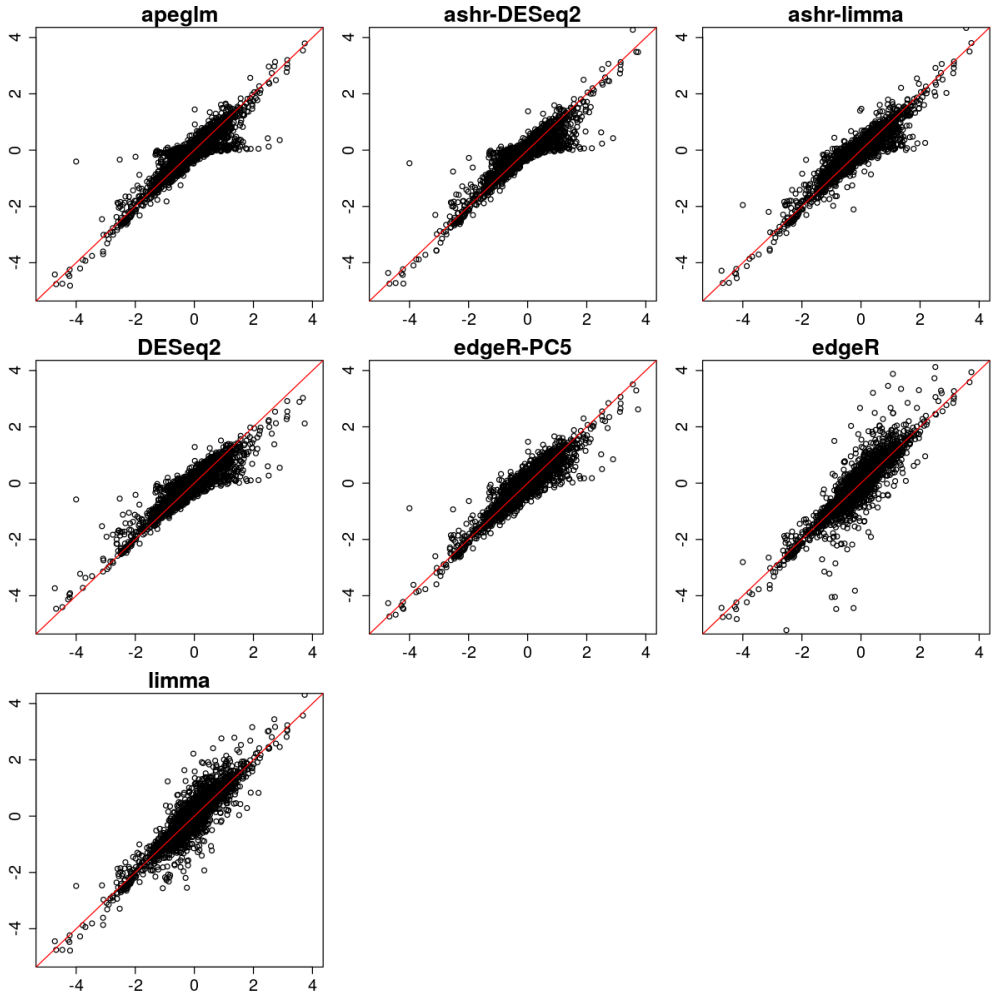
Supplementary Figure 5: Error inflation factor for various scales of prior on simulated data. We define the error inflation factor as the ratio of mean absolute error (MAE) of the estimated LFC with an adaptive prior over the MAE of estimated LFC when using a fixed prior (S=1). For a middle range of true LFC SD, having a too narrow prior (red circles, yellow triangles) lead to an increase in MAE relative to a fixed prior. While in Supplementary Figure 4, scaling the prior to be *wider* than the true LFC SD caused a problem, here scaling the prior to be *narrower* than the true LFC lead to increased errors. Setting the scale of the prior equal to the scale of the true LFC SD struck a good balance (cyan plus sign). The following Supplementary Figure 6 shows what these errors looked like for individual genes.



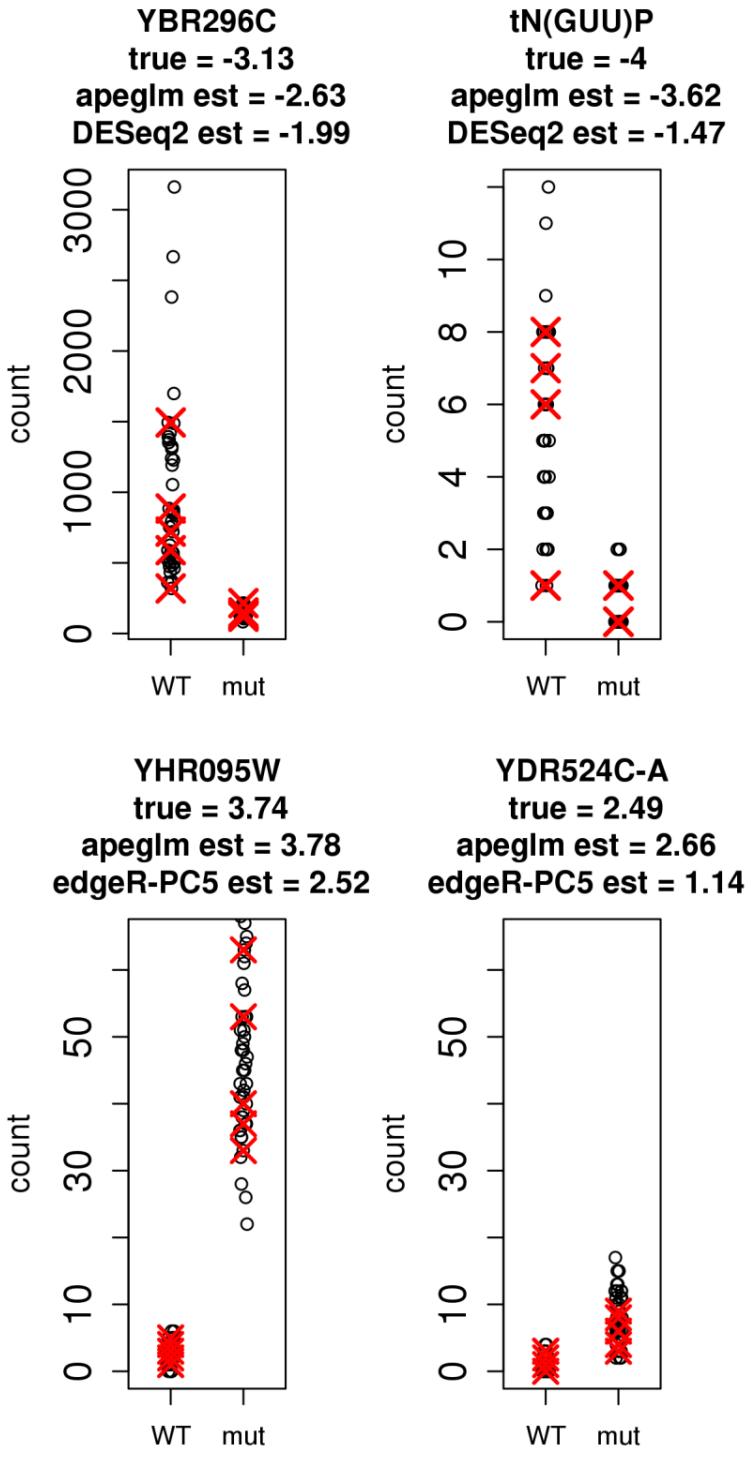
Supplementary Figure 6: Examples of estimated LFC over true LFC for a too narrow prior and a fixed prior (S=1). Shown are the estimates for a single iteration of a 5 vs 5 sample comparison where the *true LFC SD* was 0.25 and the scale of the prior was 0.1 * *true LFC SD* = 0.025, i.e. using a *multiplier* of 0.1.



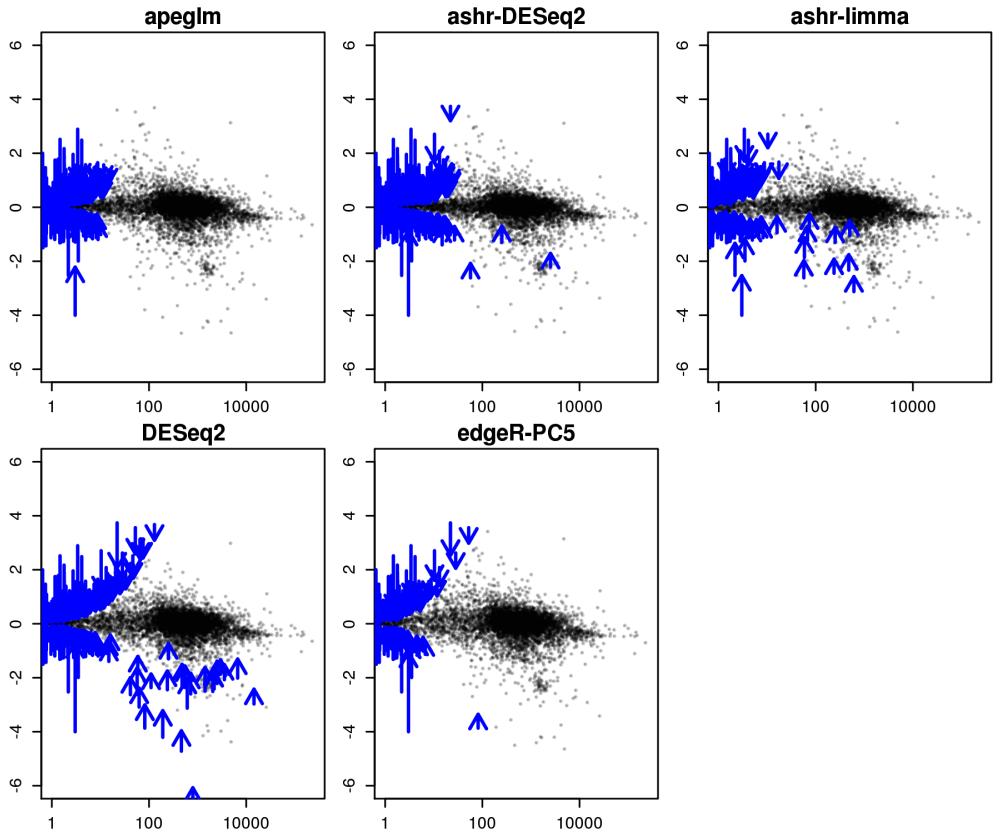
Supplementary Figure 7: From top to bottom, left to right are the distribution of true LFCs in highly replicated yeast data from Schurch et al. [3], simulation data from Pickrell et al. [4], simulation data from Bottomly et al. [2], and RNA-seq mixology dataset from Holik et al. [5] with the comparisons 075vs025 and 050vs025. Note the restriced range of LFC for the mixology dataset, for the mixtures of 075vs025 and 050vs025.



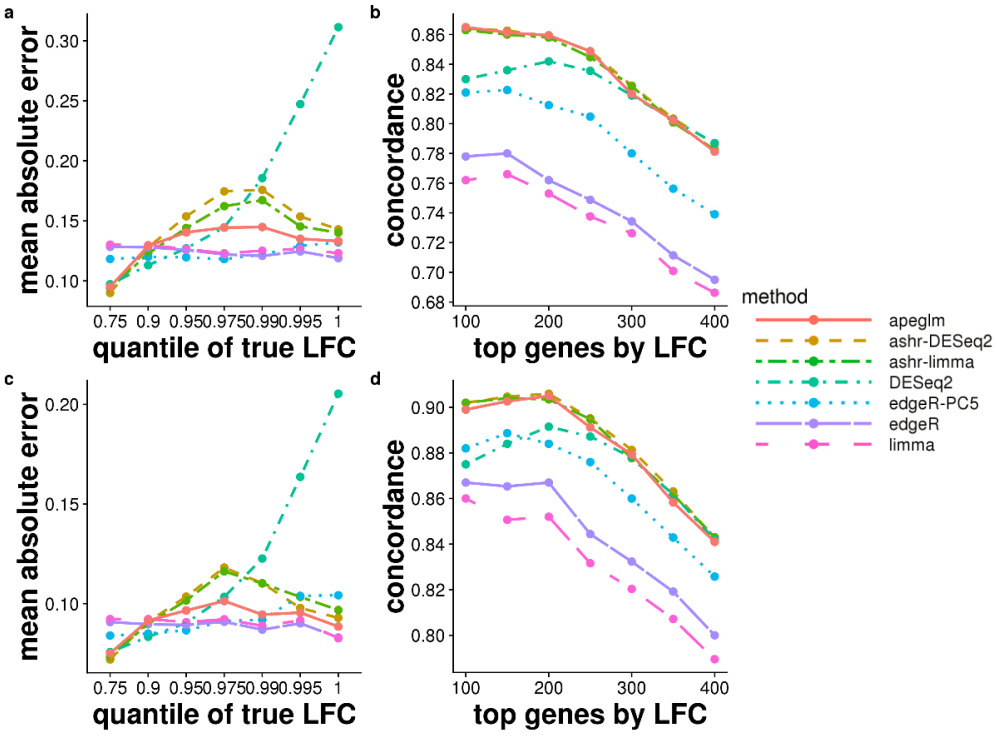
Supplementary Figure 8: Estimated LFCs from a single iteration of a 3 vs 3 samples for all genes, for the highly replicated yeast dataset from Schurch et al. [3]. The estimated LFC by seven different methods are plotted on y-axis against the reference LFC on x-axis. The red line denotes equality of estimated LFC and reference LFC. The bias for large effects for *DESeq2* can be observed, as well as the high variance for small effects for *edgeR* and *limma*.



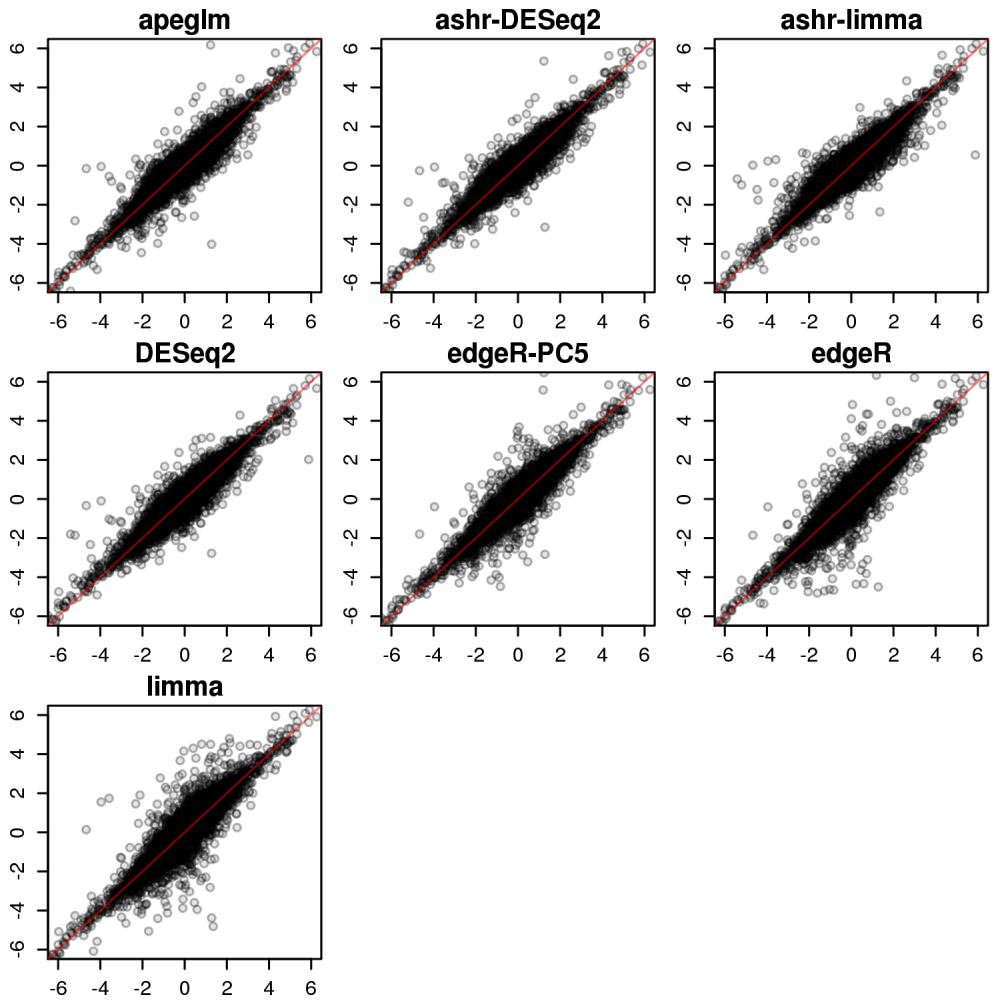
Supplementary Figure 9: The raw counts for four example genes from *DESeq2* (top) and *edgeR-PC5* (bottom) for highly replicated yeast dataset from Schurch et al. [3]. The title displays estimated LFC from different methods and the true, reference LFC. Red crosses display the 5 vs 5 samples that were drawn and black circles indicate all the samples (used to define the reference LFC).



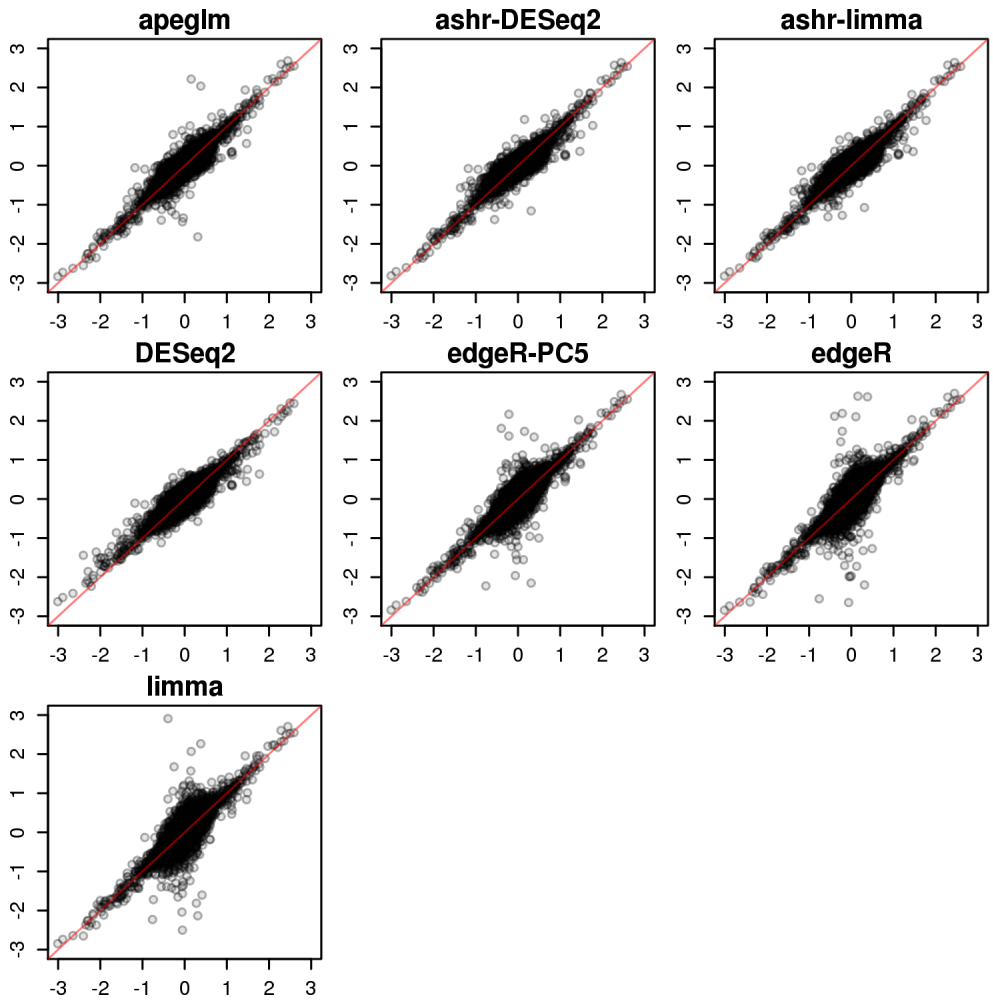
Supplementary Figure 10: The MA plot of shrinkage estimates of LFC, averaging the estimates over 100 iterations for all genes from 3 vs 3 samples, in the Schurch et al. [3] dataset. The x-axis shows the mean of scaled counts from the full dataset. This MA plot visualizes the bias of shrinkage estimators across the range of mean signal. The points show the average estimated LFC when it was less than 0.5 units from the reference LFC. The arrows show when the difference between average estimated LFC and reference LFC was larger than 0.5. The tail of the arrow is the reference LFC and the head of the arrow is the average estimated LFC from the method.



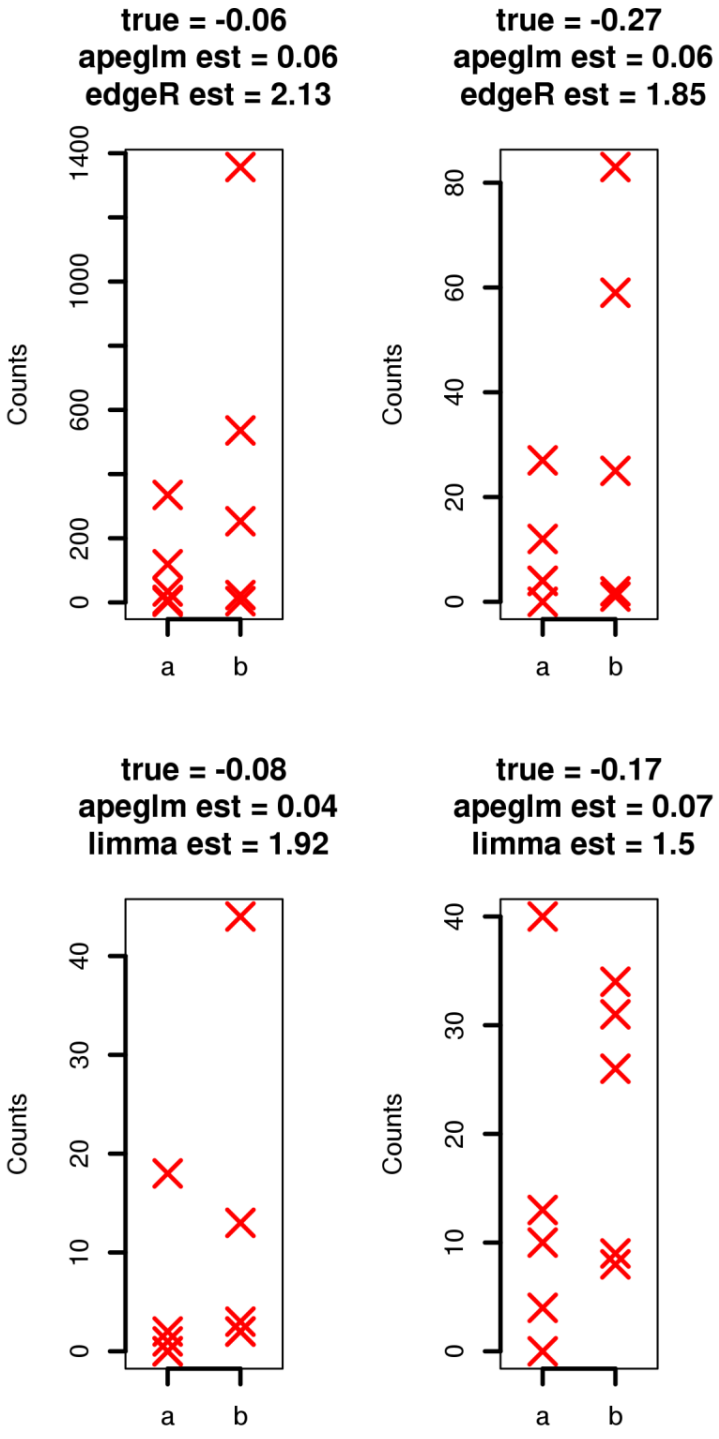
Supplementary Figure 11: MAE plot and CAT plot of the simulation dataset (top row, 5 vs 5, and bottom row, 10 vs 10) modeled on estimated parameters from the Bottomly et al. [2] dataset. Each point represents the average over 10 repeated simulations.



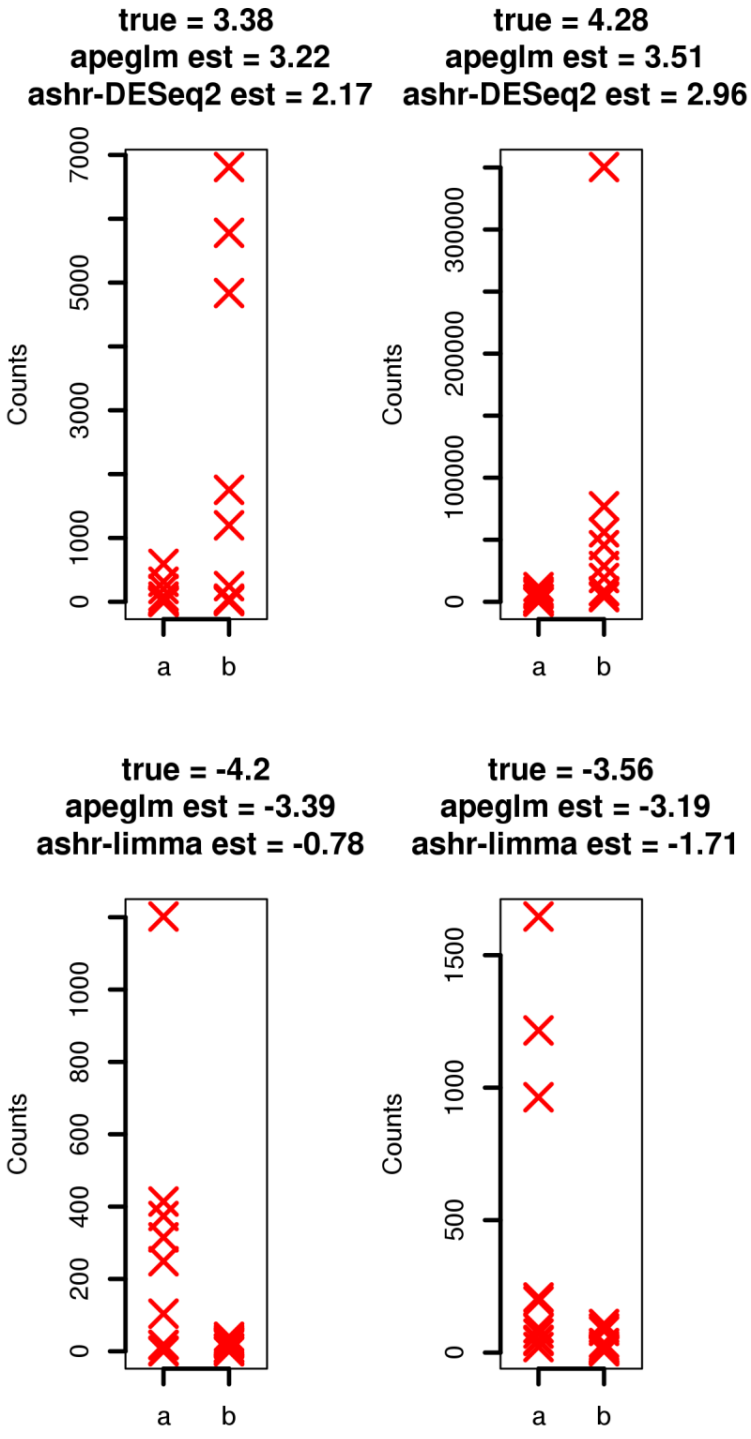
Supplementary Figure 12: Scatterplot of estimated over true LFC for one iteration (5 vs 5) of the Pickrell et al. [4] dataset.



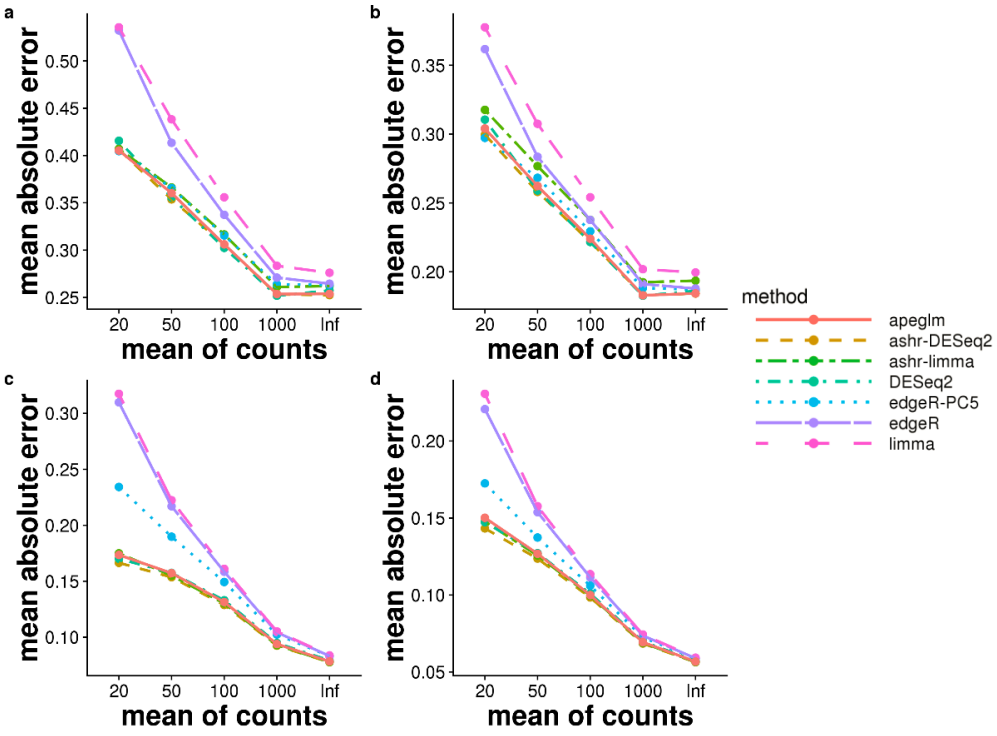
Supplementary Figure 13: Scatterplot of estimated over true LFC for one iteration (5 vs 5) of the Bottomly et al. [2] dataset.



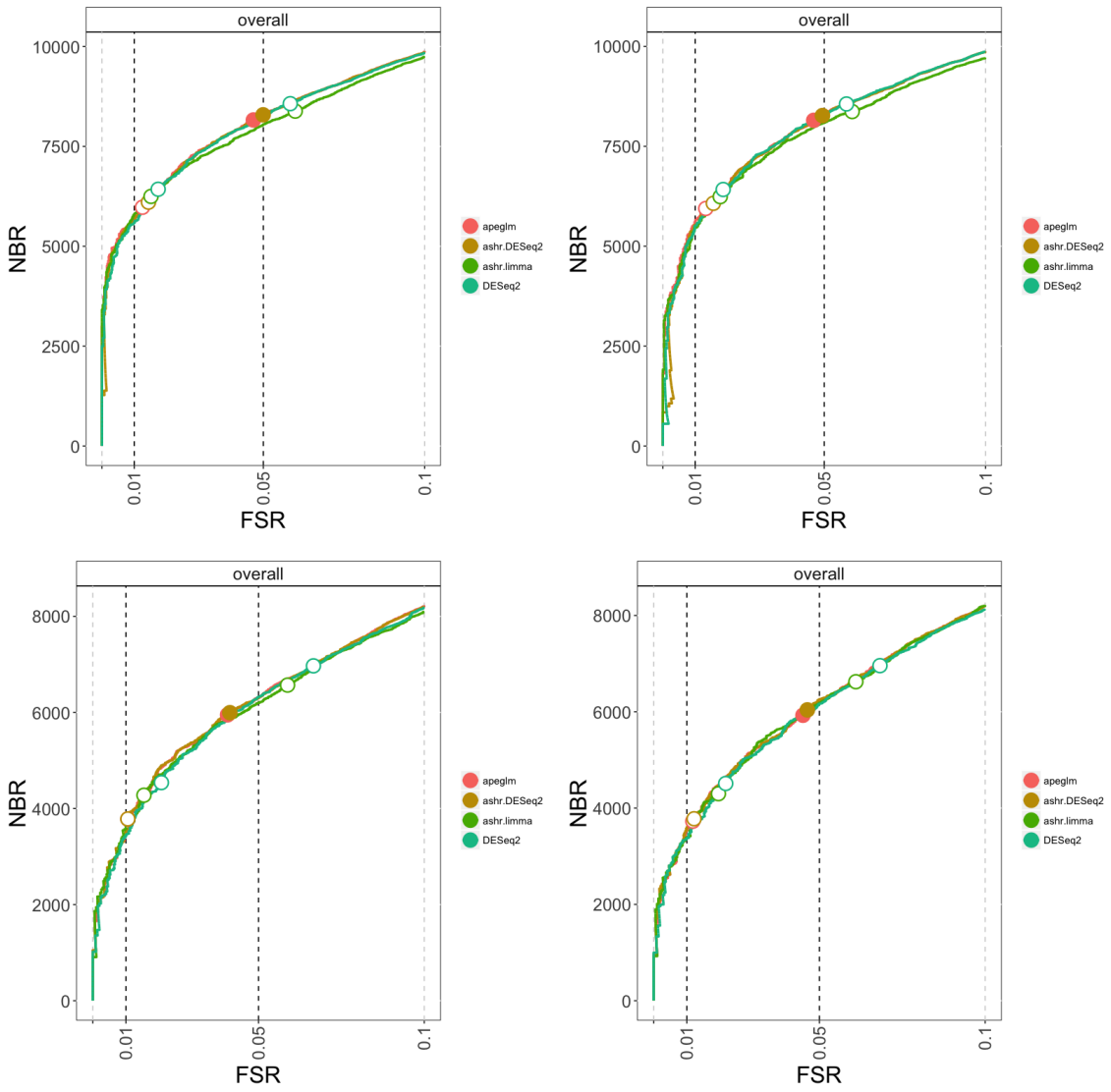
Supplementary Figure 14: The raw counts for four example genes from *edgeR* (top) and *limma* (bottom) for the 5 vs 5 simulation based on the Pickrell et al. [4] dataset. The title displays estimated LFC from different methods and the true, simulated LFC. Red crosses display the 5 vs 5 samples.



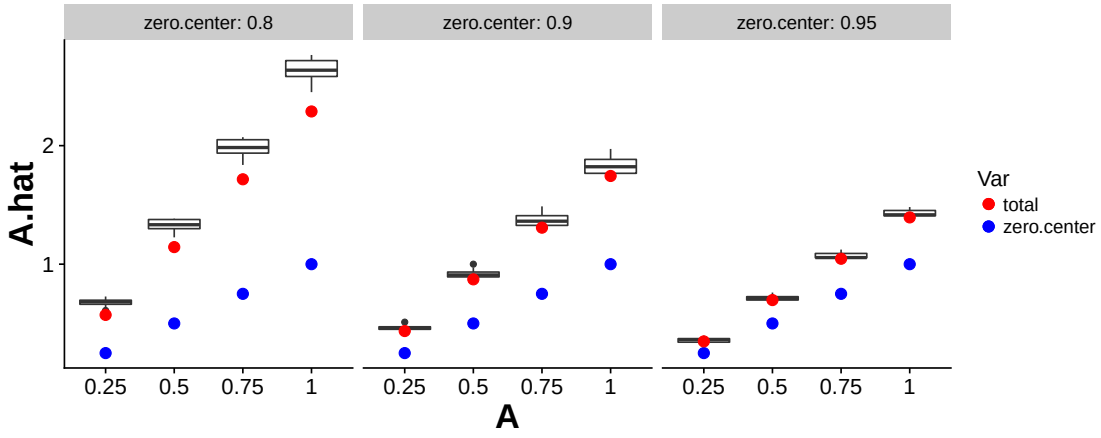
Supplementary Figure 15: The raw counts for four example genes from *ashr-DESeq2* (top) and *ashr-limma* (bottom) for the 10 vs 10 simulation based on the Bottomly et al. [2] dataset. The title displays estimated LFC from different methods and the true, simulated LFC. Red crosses display the 10 vs 10 samples.



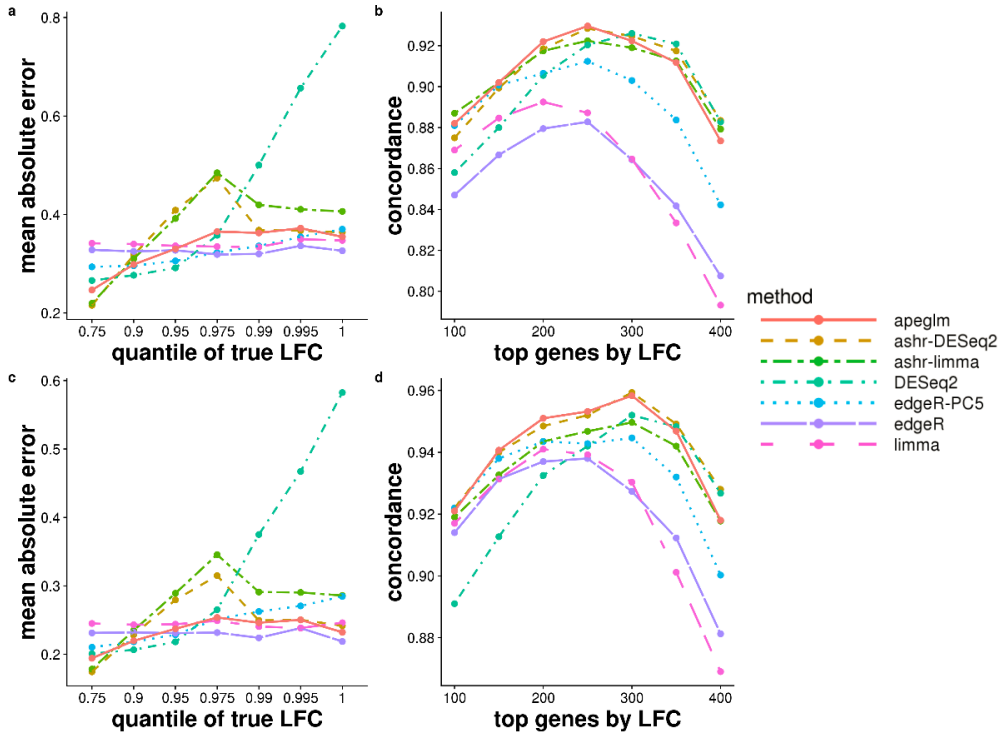
Supplementary Figure 16: MAE plot binned by the mean scaled counts of simulation dataset (top row, 5 vs 5, and bottom row, 10 vs 10) modeled on estimated parameters from the Pickrell et al. [4] (a and c) Bottomly et al. [2] dataset (b and d). Each point represents the average over 10 repeated simulations.



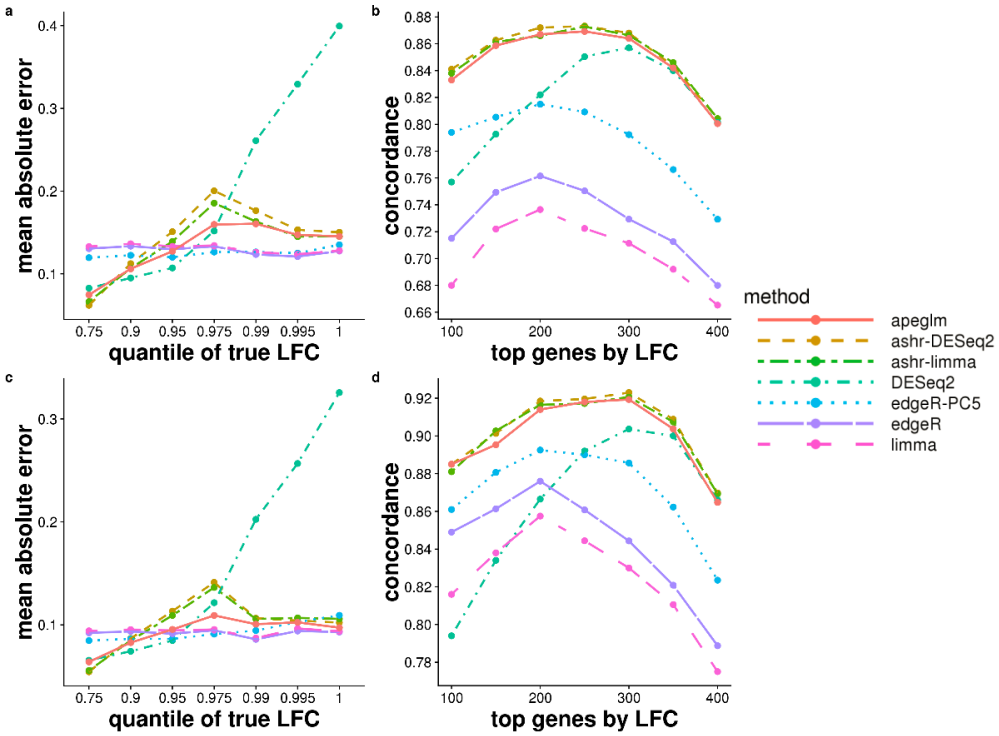
Supplementary Figure 17: Number of genes (NBR) found at various achieved false sign rate (FSR). The top row shows two iterations of the simulated dataset modeled on the Pickrell et al. [4] data, the bottom row shows two iterations of the simulated dataset modeled on the Bottomly et al. [2] data, both with sample size of 5 vs 5. The plots are generated with the *iCOBRA* [6] Bioconductor package, with the two sets of circles indicating *s*-value cutoffs of 1% and 5%, and filled circles indicating that a method has achieved the nominal FSR bound from the *s*-value cutoff. The `COBRAData` objects for these four simulated datasets can be accessed at `https://github.com/mikelove/apeglmPaper`, where there are instructions on how to launch an interactive Shiny app for exploring the *s*-values and estimated LFCs across methods.



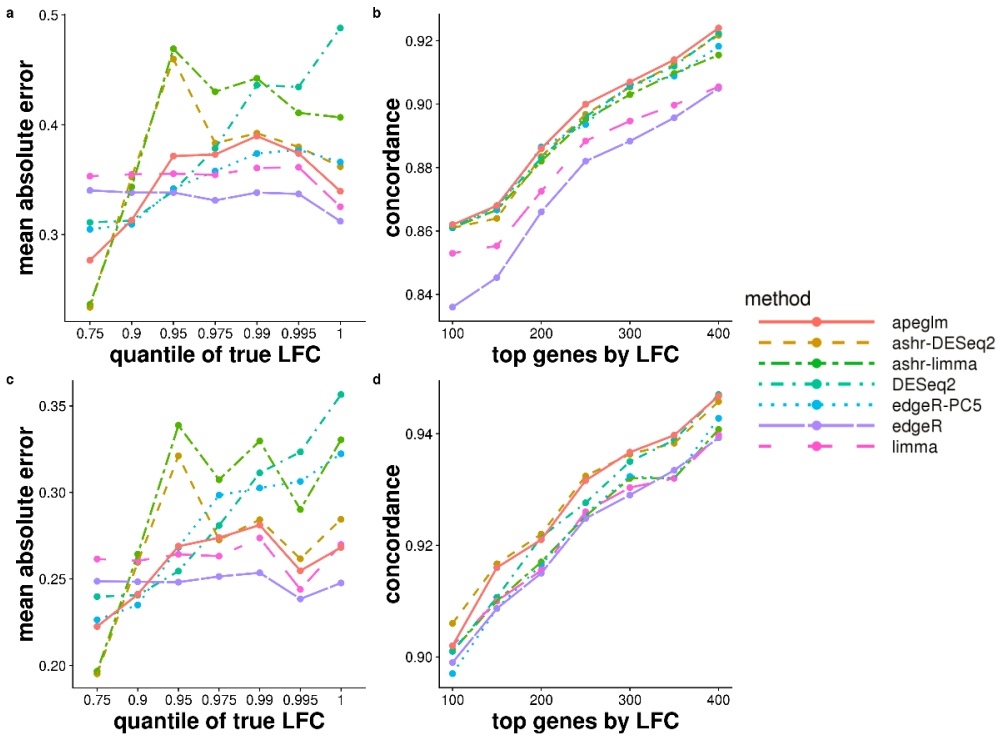
Supplementary Figure 18: Effects of estimating A when the true distribution is bimodal. Instead of simulating LFC from a zero-centered Normal distribution, LFC were simulated from a bimodal distribution with a zero-centered Normal component with variance A, and a Normal component centered at $3\sqrt{A}$ with variance $A/4$. The three panels represent cases where the fraction of LFCs from the zero-centered component was one of $\{0.8, 0.9, 0.95\}$. On the x-axis is the true value of A across simulations, and on the y-axis is the estimate $\hat{A}$ using the method implemented in *apeglm*. The boxes denote the result of 10 iterations, and the blue dot denotes the true value $\hat{A} = A$ for the zero-centered component. The red dot denotes the total variance of the true LFCs (not observed by *apeglm*). In general, the estimated scale of the prior tracked with the total variance of the true LFCs, so including both the zero-centered and non-zero-centered component.



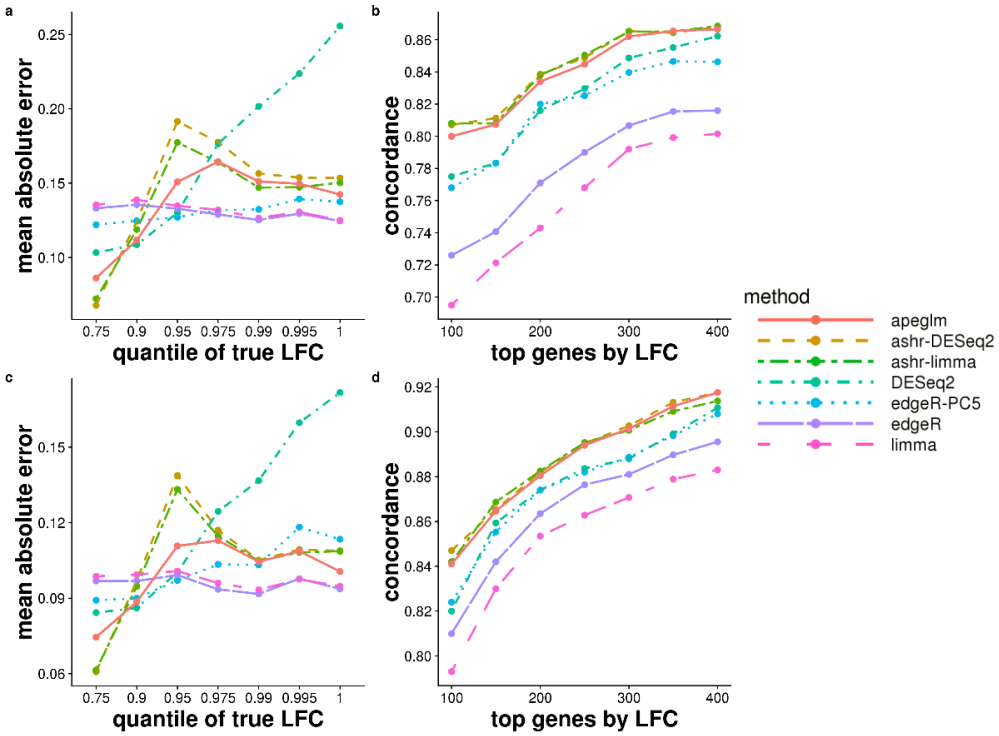
Supplementary Figure 19: MAE plots over true LFCs (left) and CAT plots (right) for simulation datasets for Pickrell et al. [4] with sample sizes 5 vs 5 (top) and 10 vs 10 (bottom), with 5% of genes have strong positive LFCs.



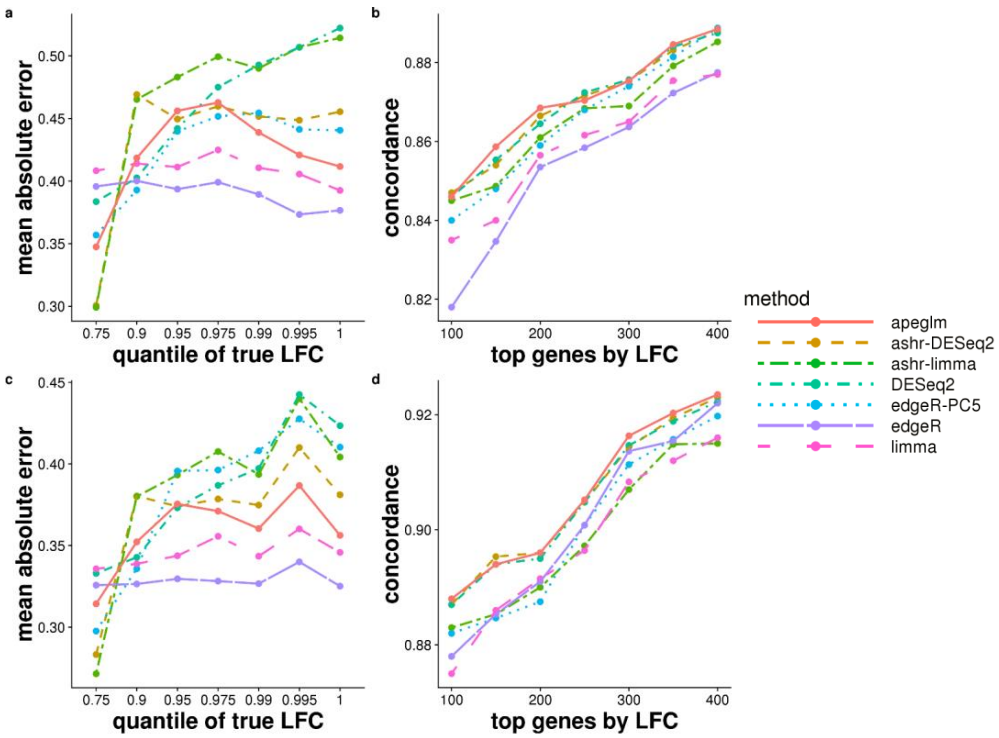
Supplementary Figure 20: MAE plots over true LFCs (left) and CAT plots (right) for simulation datasets for Bottomly et al. [2] with sample sizes 5 vs 5 (top) and 10 vs 10 (bottom), with 5% of genes have strong positive LFCs.



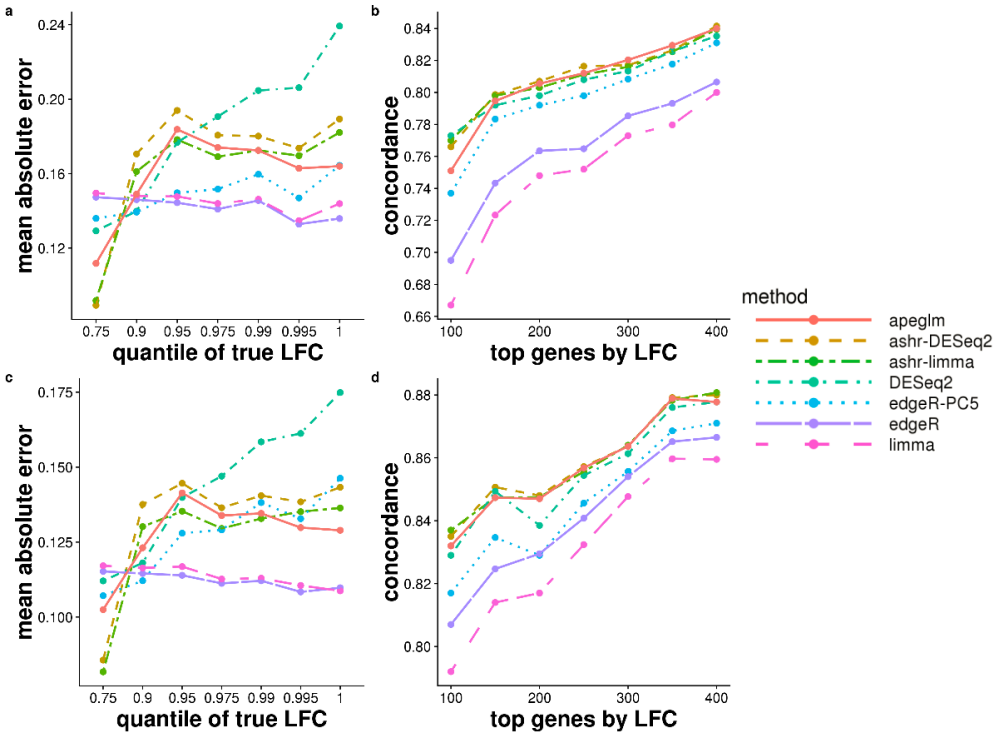
Supplementary Figure 21: MAE plots over true LFCs (left) and CAT plots (right) for simulation datasets for Pickrell et al. [4] with sample sizes 5 vs 5 (top) and 10 vs 10 (bottom), with 10% of genes have strong positive LFCs.



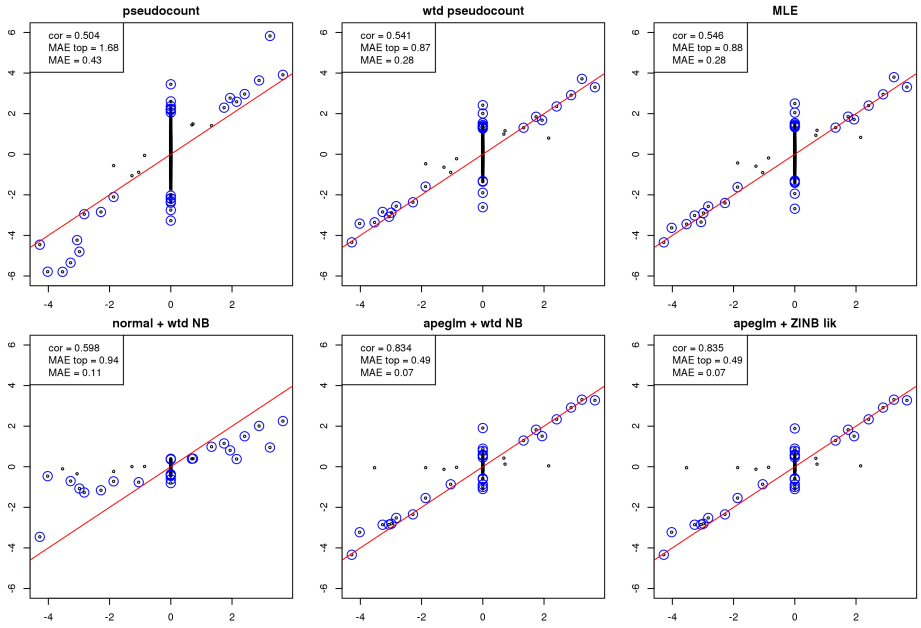
Supplementary Figure 22: MAE plots over true LFCs (left) and CAT plots (right) for simulation datasets for Bottomly et al. [2] with sample sizes 5 vs 5 (top) and 10 vs 10 (bottom), with 10% of genes have strong positive LFCs.



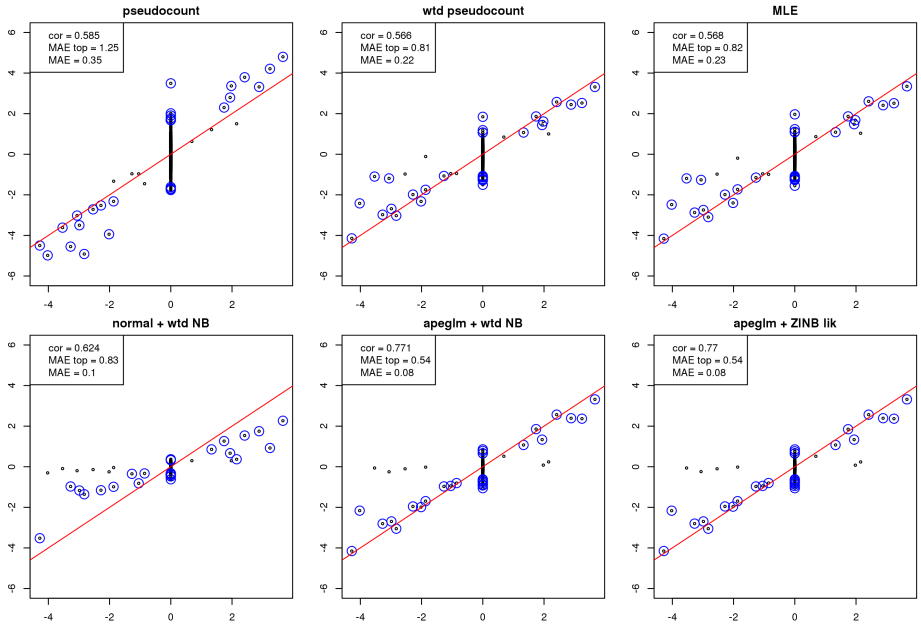
Supplementary Figure 23: MAE plots over true LFCs (left) and CAT plots (right) for simulation datasets for Pickrell et al. [4] with sample sizes 5 vs 5 (top) and 10 vs 10 (bottom), with 20% of genes have strong positive LFCs.



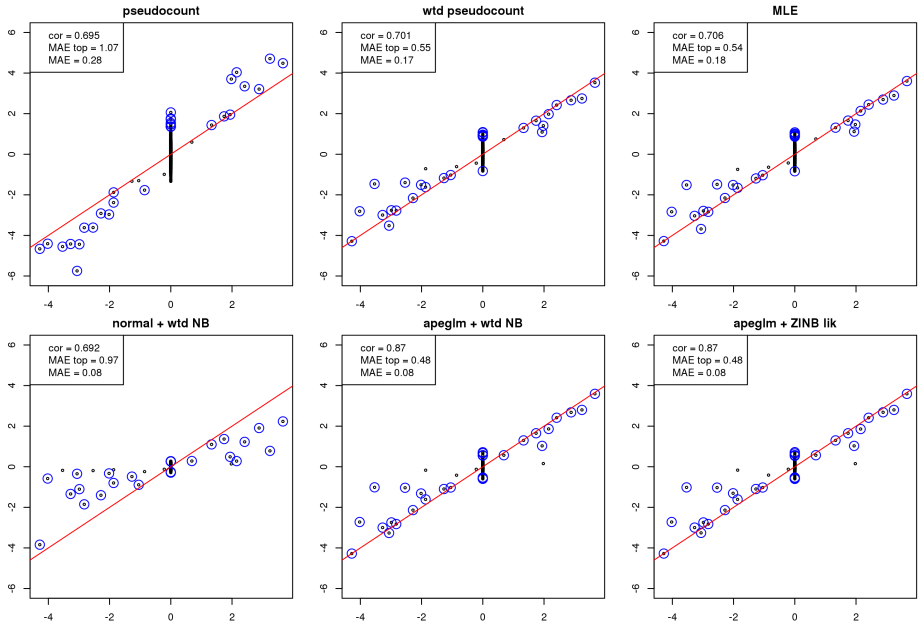
Supplementary Figure 24: MAE plots over true LFCs (left) and CAT plots (right) for simulation datasets for Bottomly et al. [2] with sample sizes 5 vs 5 (top) and 10 vs 10 (bottom), with 20% of genes have strong positive LFCs.



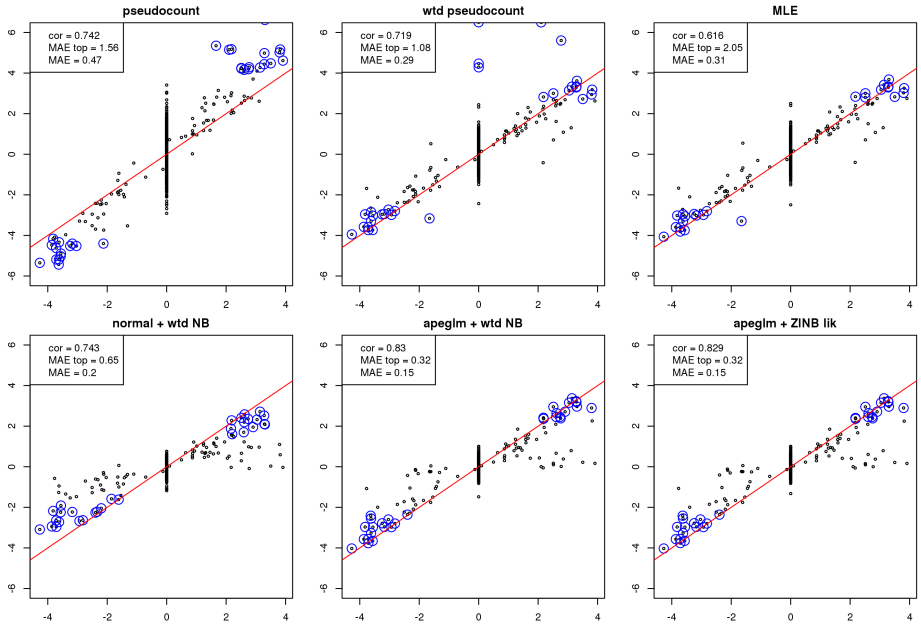
Supplementary Figure 25: Scatter plots of estimated LFCs over true LFCs for the *splatter* ZINB dataset simulating a single-cell experiment. 1% of the genes are differentially expressed genes and sample size is 20 vs 20. Dots in blue circles are the top 30 genes ranking by the estimated LFCs.



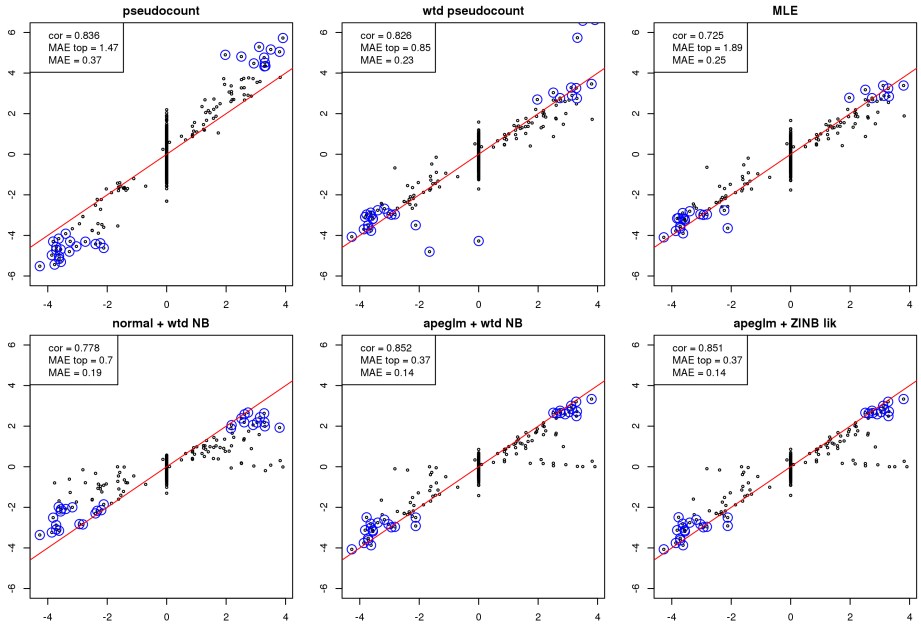
Supplementary Figure 26: Scatter plots of estimated LFCs over true LFCs for the *splatter* ZINB dataset simulating a single-cell experiment. 1% of the genes are differentially expressed genes and sample size is 30 vs 30. Dots in blue circles are the top 30 genes ranking by the estimated LFCs.



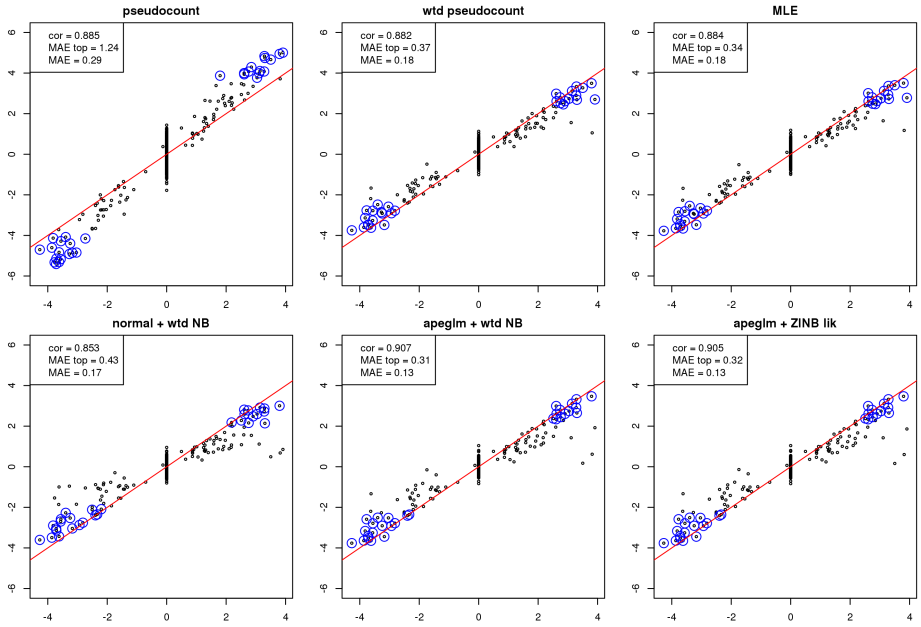
Supplementary Figure 27: Scatter plots of estimated LFCs over true LFCs for the *splatter* ZINB dataset simulating a single-cell experiment. 1% of the genes are differentially expressed genes and sample size is 50 vs 50. Dots in blue circles are the top 30 genes ranking by the estimated LFCs.



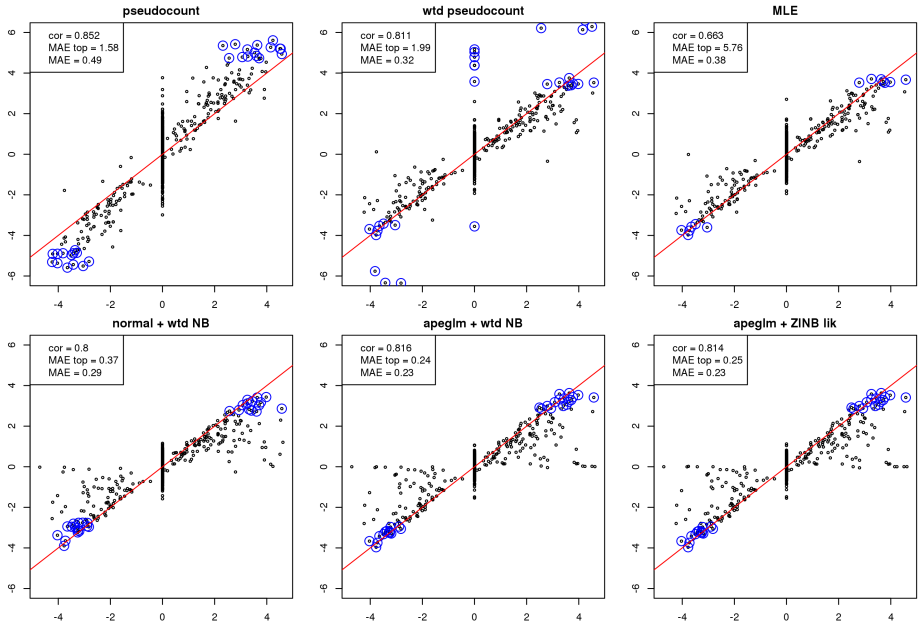
Supplementary Figure 28: Scatter plots of estimated LFCs over true LFCs for the *splatter* ZINB dataset simulating a single-cell experiment. 5% of the genes are differentially expressed genes and sample size is 20 vs 20. Dots in blue circles are the top 30 genes ranking by the estimated LFCs.



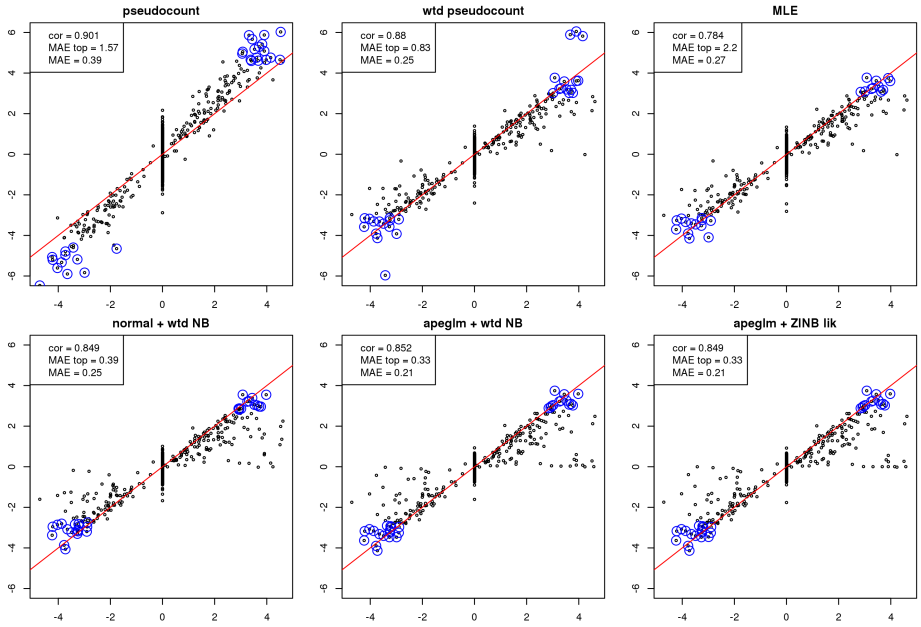
Supplementary Figure 29: Scatter plots of estimated LFCs over true LFCs for the *splatter* ZINB dataset simulating a single-cell experiment. 5% of the genes are differentially expressed genes and sample size is 30 vs 30. Dots in blue circles are the top 30 genes ranking by the estimated LFCs.



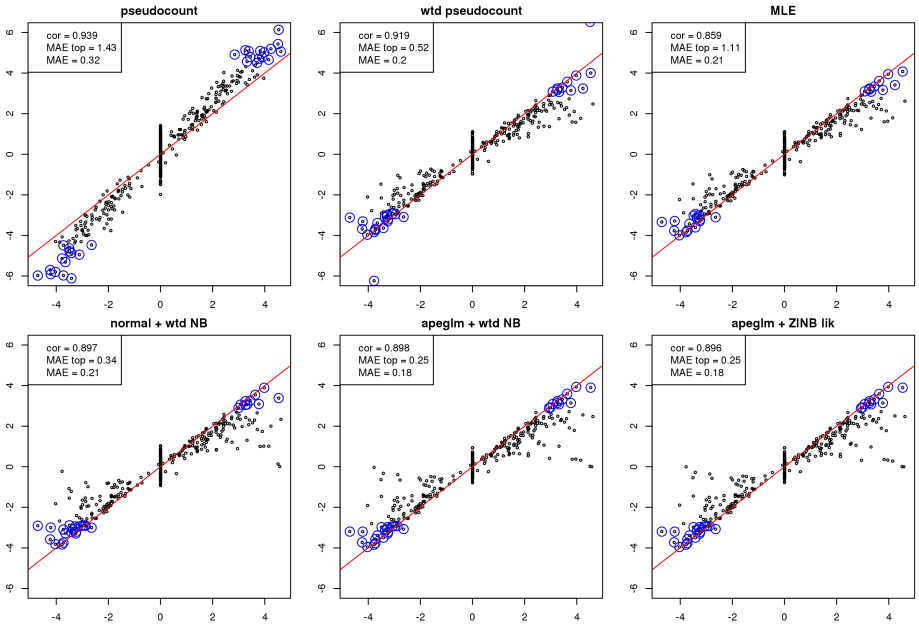
Supplementary Figure 30: Scatter plots of estimated LFCs over true LFCs for the *splatter* ZINB dataset simulating a single-cell experiment. 5% of the genes are differentially expressed genes and sample size is 50 vs 50. Dots in blue circles are the top 30 genes ranking by the estimated LFCs.



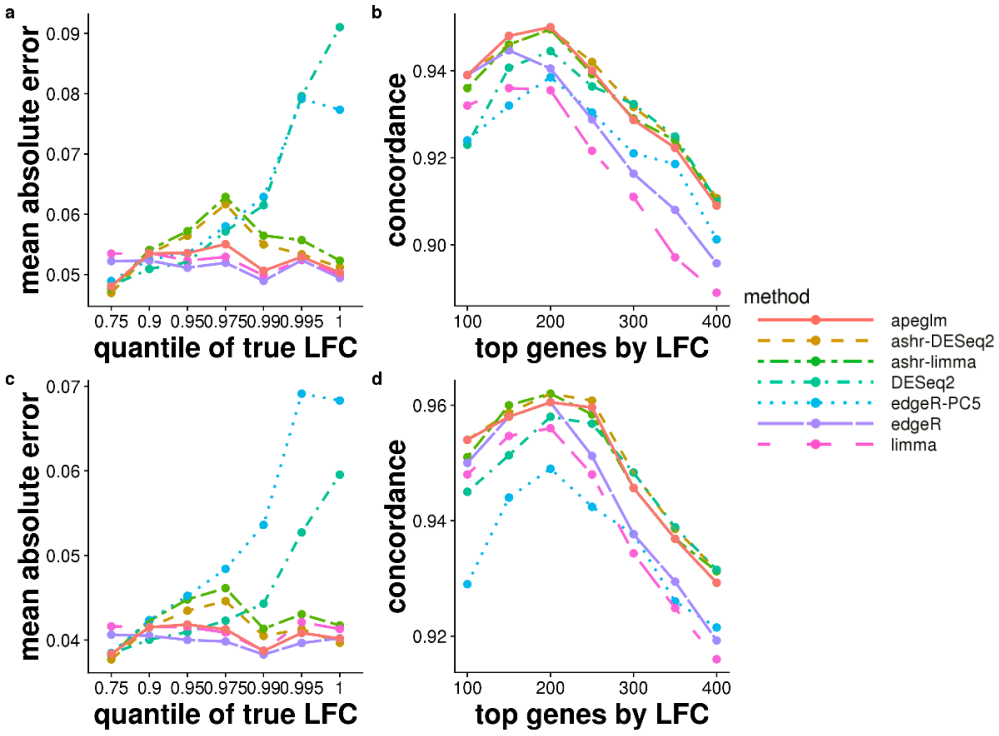
Supplementary Figure 31: Scatter plots of estimated LFCs over true LFCs for the *splatter* ZINB dataset simulating a single-cell experiment. 10% of the genes are differentially expressed genes and sample size is 20 vs 20. Dots in blue circles are the top 30 genes ranking by the estimated LFCs.



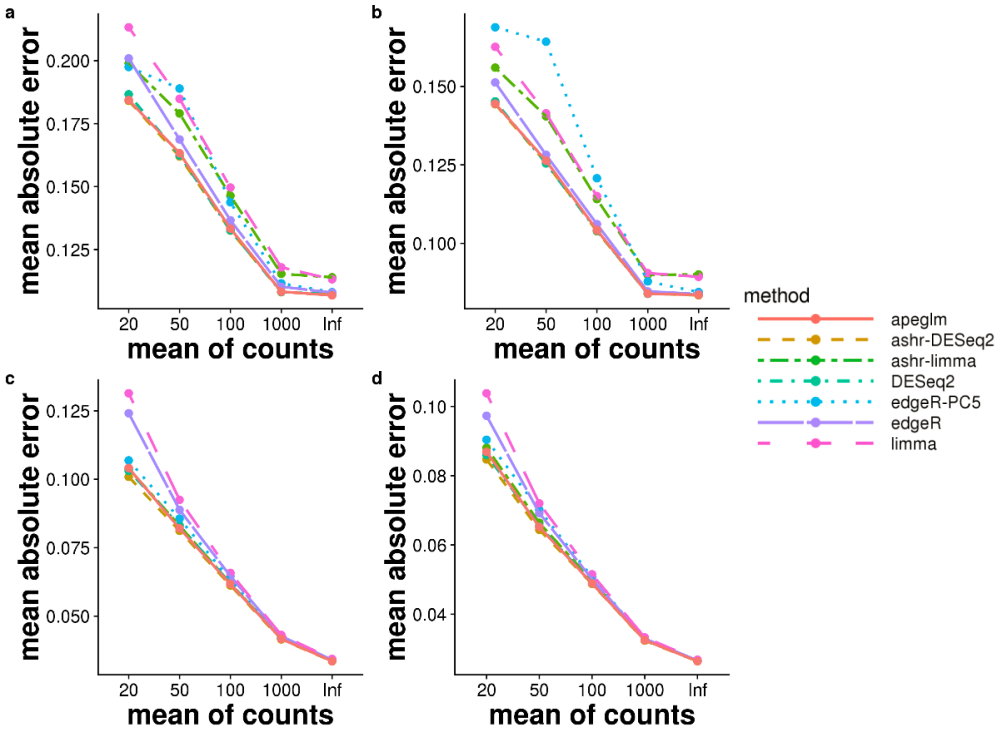
Supplementary Figure 32: Scatter plots of estimated LFCs over true LFCs for the *splatter* ZINB dataset simulating a single-cell experiment. 10% of the genes are differentially expressed genes and sample size is 30 vs 30. Dots in blue circles are the top 30 genes ranking by the estimated LFCs.



Supplementary Figure 33: Scatter plots of estimated LFCs over true LFCs for the *splatter* ZINB dataset simulating a single-cell experiment. 10% of the genes are differentially expressed genes and sample size is 50 vs 50. Dots in blue circles are the top 30 genes ranking by the estimated LFCs.



Supplementary Figure 34: Simulation dataset (top row, 30 vs 30, and bottom row, 50 vs 50) modeled on estimated parameters from the Bottomly et al. [2] dataset. Each point represents the average over 10 repeated simulations.



Supplementary Figure 35: MAE plots over mean normalized counts for simulation datasets for Pickrell et al. [4] (top) and Bottomly et al. [2] (bottom) with sample sizes 30 vs 30 (left) and 50 vs 50 (right).

References

[1] Bradley Efron and Carl Morris. Data Analysis Using Stein's Estimator and Its Generalization. *Journal of the American Statistical Association*, 70(350):311–319, 1975.

[2] Daniel Bottomly, Nicole A. R. Walter, Jessica Ezzell Hunter, Priscila Darakjian, Sunita Kawane, Kari J. Buck, Robert P. Searles, Michael Mooney, Shannon K. McWeeney, and Robert Hitzemann. Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. *PloS one*, 6(3):e17820, March 2011. ISSN 1932-6203.

[3] Nicholas J. Schurch, Piet Schofield, Marek Gierliski, Christian Cole, Alexander Sherstnev, Vijender Singh, Nicola Wrobel, Karim Gharbi, Gordon G. Simpson, Tom Owen-Hughes, Mark Blaxter, and Geoffrey J. Barton. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, 22(6):839–851, 2016.

[4] Joseph K. Pickrell, John C. Marioni, Athma A. Pai, Jacob F. Degner, Barbara E. Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K. Pritchard. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(768–772), 2010.

[5] Aliaksei Z. Holik, Charity W. Law, Ruijie Liu, Zeya Wang, Wenyi Wang, Jaeil Ahn, Marie-Liesse Asselin-Labat, Gordon K. Smyth, and Matthew E. Ritchie. RNA-seq mixology: designing realistic control experiments to compare protocols and analysis methods. *Nucleic Acids Research*, 45(5):e30–e30, 2016.

[6] Charlotte Soneson and Mark D. Robinson. iCOBRA: open, reproducible, standardized and live method benchmarking. *Nature Methods*, 13(283), 2016.