Supplementary Figures

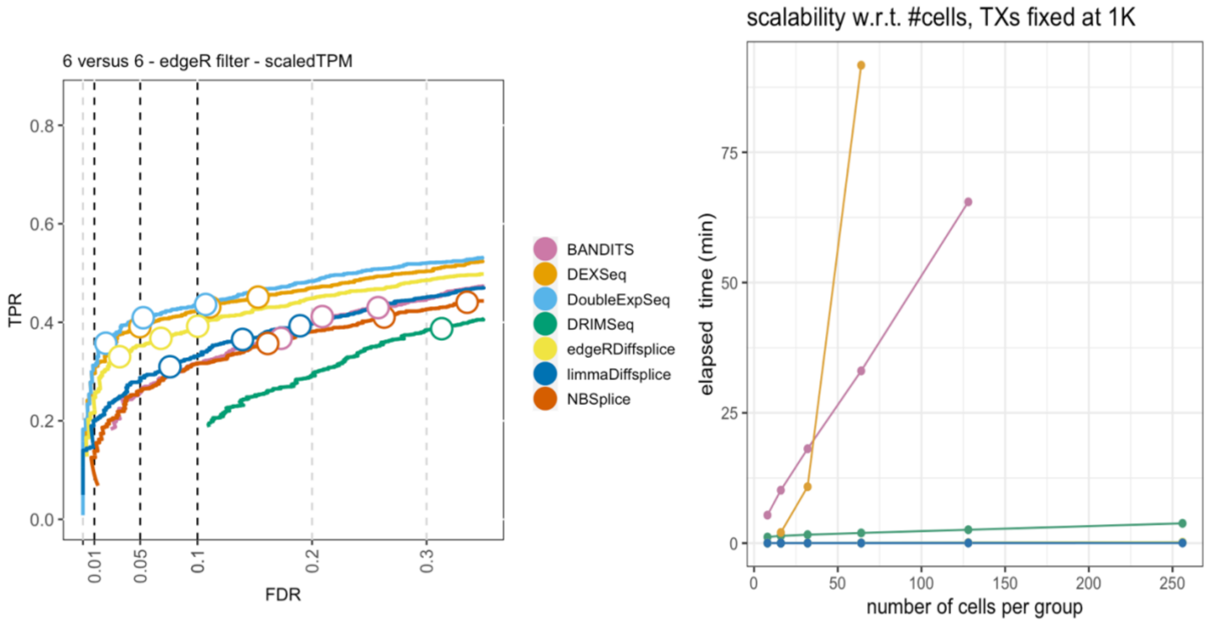


Figure S1: Performance and scalability evaluation on a subset of the Love et al. dataset. To allow for a performance and scalability evaluation of BANDITS, which does not scale to datasets with a large number of transcripts, we here perform a DTU analysis for the *6 versus 6* samples dataset of Love et al. with only 1000 transcripts. **Left panel: performance evaluation.** The results are in line with those of Figure 1A. The performance of BANDITS is indicated in pink. **Right panel: Scalability evaluation.** BANDITS scales linearly with respect to the number of cells (or samples) in the dataset. The slope of the linear trend, however, is considerably larger than those of the other DTU methods that scale linearly. Note that the profiles of limma diffsplice, edgeR diffsplice and DoubleExpSeq overlap in this figure.

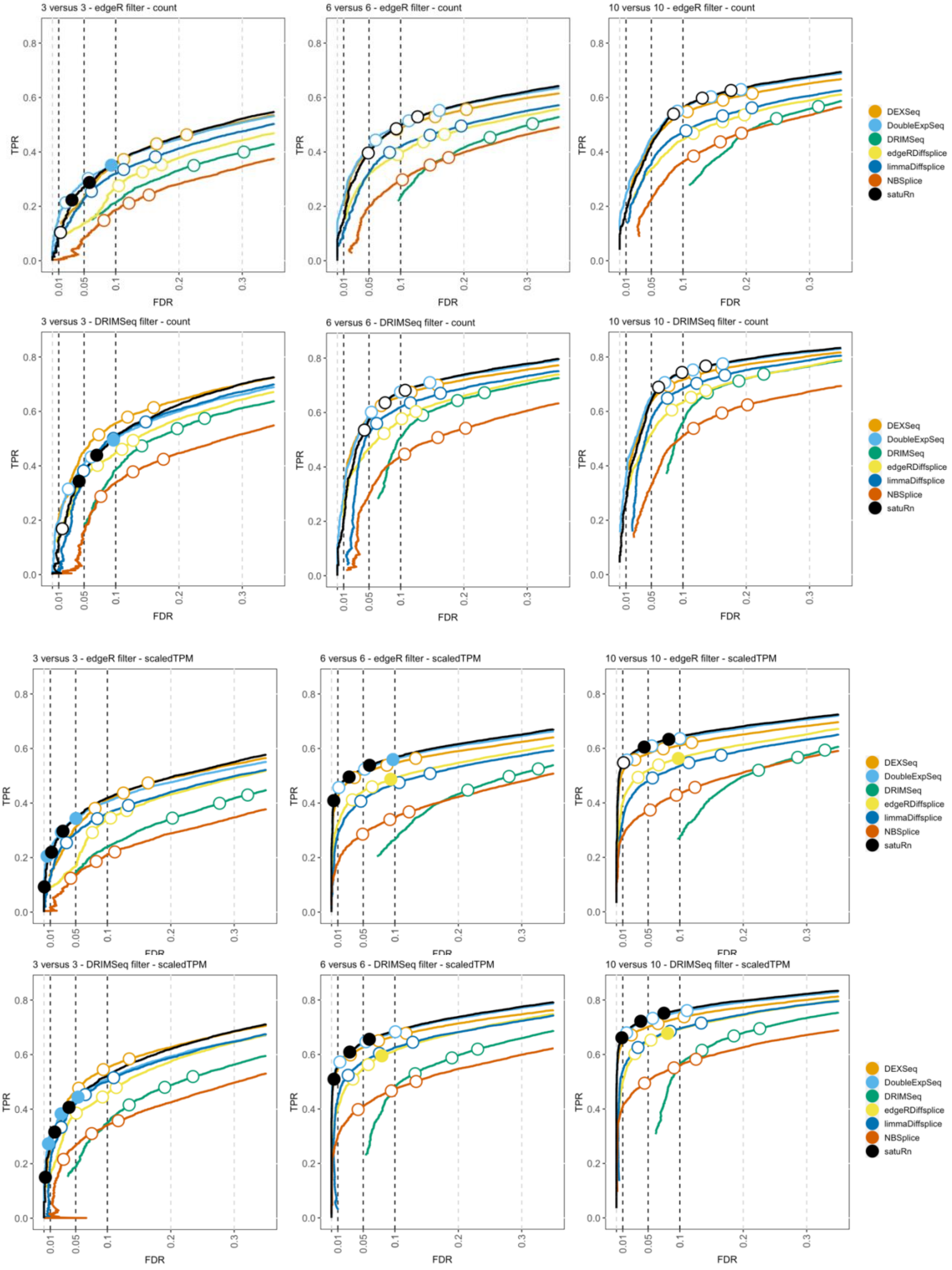
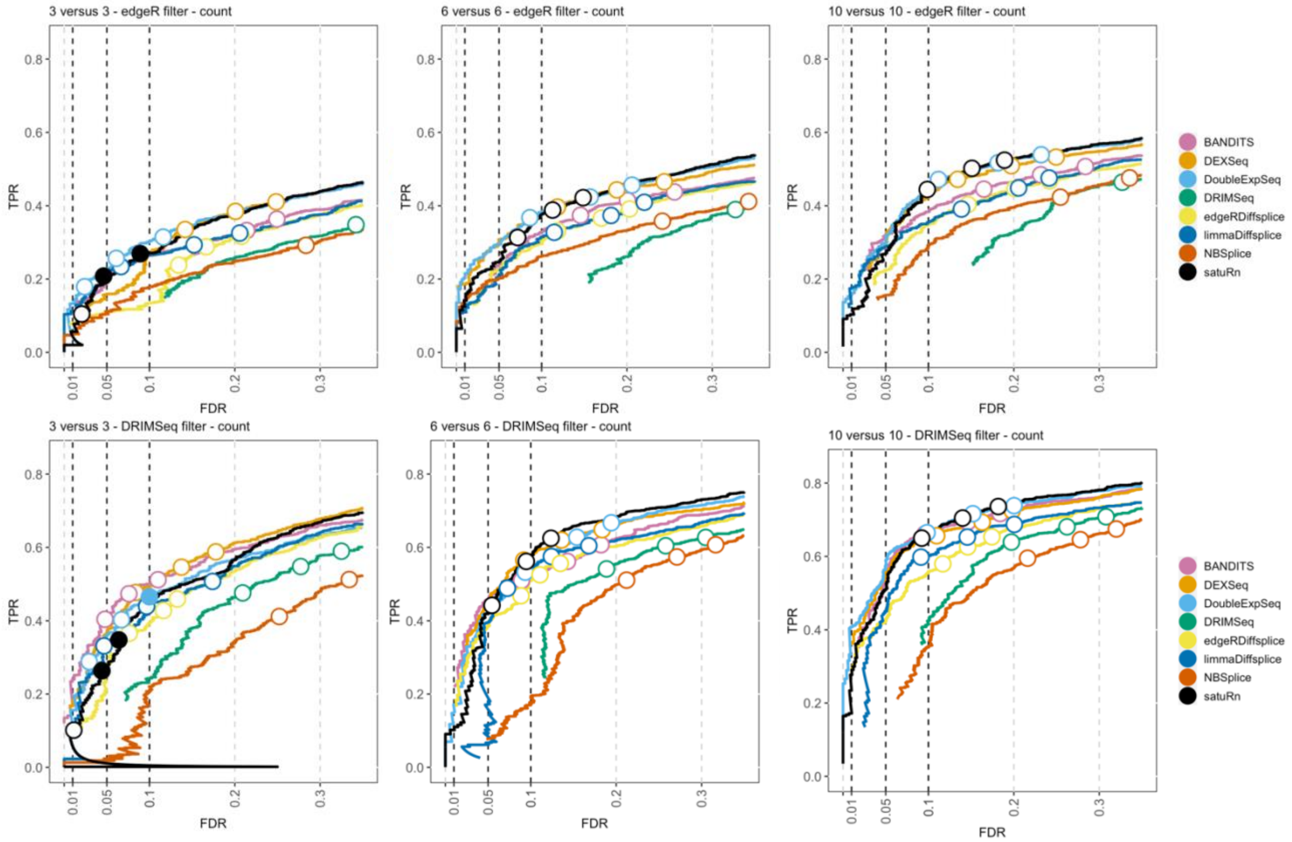
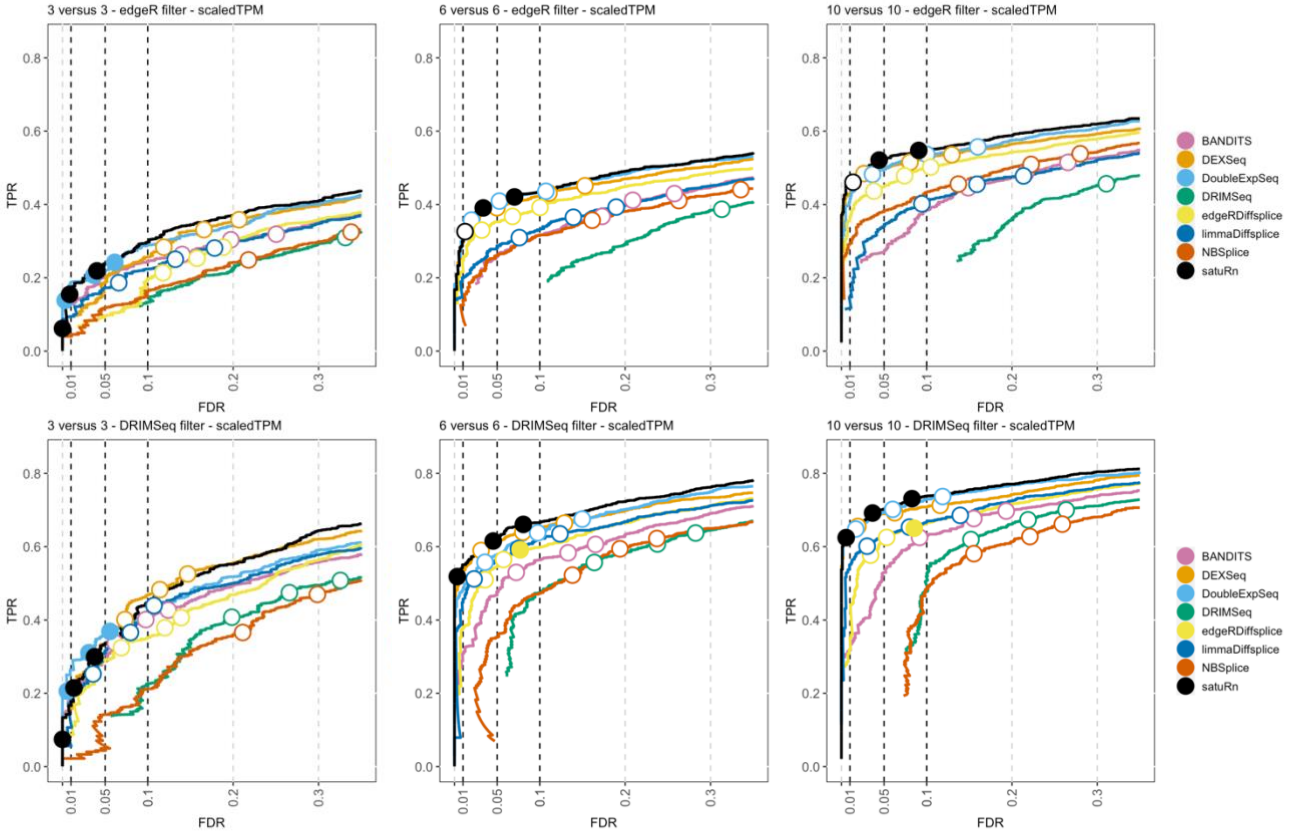


Figure S2: Performance evaluation of satuRn on different subsamples of the simulated bulk RNA-Seq dataset by Love et al. FDR-TPR curves visualize the performance of each method by displaying the sensitivity of the method (TPR) with respect to the false discovery rate (FDR). The three circles on each curve represent working

points when the FDR level is set at nominal levels of 1%, 5% and 10%, respectively. The circles are filled if the empirical FDR is equal or below the imposed FDR threshold. We subsampled two-group comparisons according to three different samples sizes; a *3 versus 3*, *6 versus 6* and *10 versus 10* comparison, as denoted in the panel titles. The benchmark was performed both on the raw counts **(rows 1 and 2)** or on scaled transcripts-per-million (TPM) **(rows 3 and 4)** as imported with the Bioconductor R package tximport[1]. We additionally adopted two different filtering strategies: an edgeR-based filtering **(rows 1 and 3)** and a DRIMSeq-based filtering **(rows 2 and 4)**. Overall, the performance of satuRn is on par with those of the best tools in the literature, DEXSeq and DoubleExpSeq. In addition, satuRn achieves a better control of the FDR on all datasets. For extremely small sample size, i.e. the *3 versus 3* comparison, the performance is slightly below that of DEXSeq, and inference does become slightly too conservative. Note that, as expected, the performances increase with increasing sample size, and a higher performance is achieved with the more stringent DRIMSeq filtering criterion (see Methods), which goes at the cost of retaining fewer transcripts for DTU analysis. Finally, we note that the performances and FDR control are consistently higher for the scaled TPM data as compared to the raw counts. Note that this was only observed for this particular dataset.



86



87 **Figure S3: Performance evaluation on different subsamples of the simulated bulk RNA-Seq dataset by Love et**
88 **al. with a reduced number of transcripts to allow for a comparison with BANDITS.** FDR-TPR curves visualize the
89 performance of each method by displaying the sensitivity of the method (TPR) with respect to the false discovery

rate (FDR). The three circles on each curve represent working points when the FDR level is set at nominal levels of 1%, 5% and 10%, respectively. The circles are filled if the empirical FDR is equal or below the imposed FDR threshold. We subsampled two-group comparisons according to three different samples sizes; a *3 versus 3*, *6 versus 6* and *10 versus 10* comparison, as denoted on top of the panels. The benchmark was performed both on the raw counts **(rows 1 and 2)** or on scaled transcripts-per-million (TPM) **(rows 3 and 4)** as imported with the Bioconductor R package tximport[1]. We additionally adopted two different filtering strategies: an edgeR-based filtering **(rows 1 and 3)** and a DRIMSeq-based filtering **(rows 2 and 4)**. Note that, in contrast to Figure S2, we additionally randomly subsampled 1000 genes (~3000-5000 transcripts) after filtering, in order to reduce the number of transcripts in the data and thereby allowing for a DTU analysis with BANDITS. In concordance with Figure S2, the performance of satuRn is on par with the best tools of the literature with a better control of the FDR in general. While the performance of BANDITS is good for the settings for which it was originally developed, (i.e., small datasets with a stringent filtering criterium), its performance is reduced in larger, more leniently filtered datasets and inference is also overly liberal in these settings. In addition, while all other methods perform much better on the scaledTPM data (rows 3 and 4) than on the raw count data (rows 1 and 2), BANDITS has a similar performance on both input data types. This can be explained by the fact that BANDITS inherently corrects for differences in transcript length, even when raw counts are used as an input.

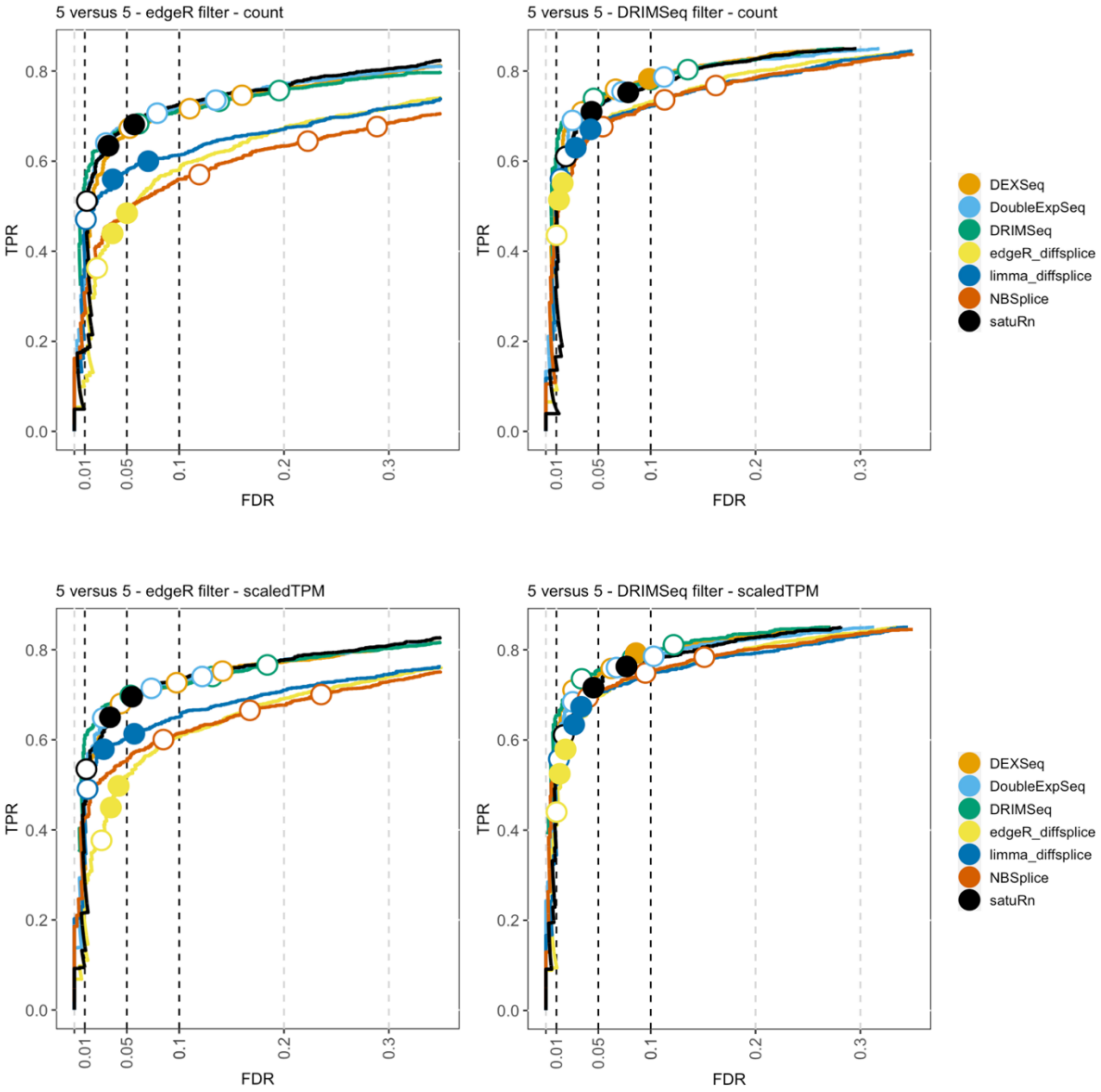


Figure S4: Performance evaluation of satuRn on the "Dmelanogaster" simulated bulk RNA-Seq dataset by Van den Berge et al. FDR-TPR curves visualize the performance of each method by displaying the sensitivity of the method (TPR) with respect to the false discovery rate (FDR). The three circles on each curve represent working points when the FDR level is set at nominal levels of 1%, 5% and 10%, respectively. The circles are filled if the empirical FDR is equal or below the imposed FDR threshold. The benchmark was performed both on the raw counts **(row 1)** and on scaled TPM **(row 2)** as imported with the Bioconductor R package tximport[1]. We additionally adopted two different filtering strategies; an edgeR-based filtering **(column 1)** and a DRIMSeq-based filtering **(column 2)**. Overall, the performance of satuRn is on par with those of the best tools in the literature, DEXSeq and DoubleExpSeq. In contrast to the performance evaluation on the dataset by Love et al. (Figures 1A and S2), there is a limited difference in performances based on the data input type (i.e., counts versus scaled TPM), and DRIMSeq also performs well on these datasets.

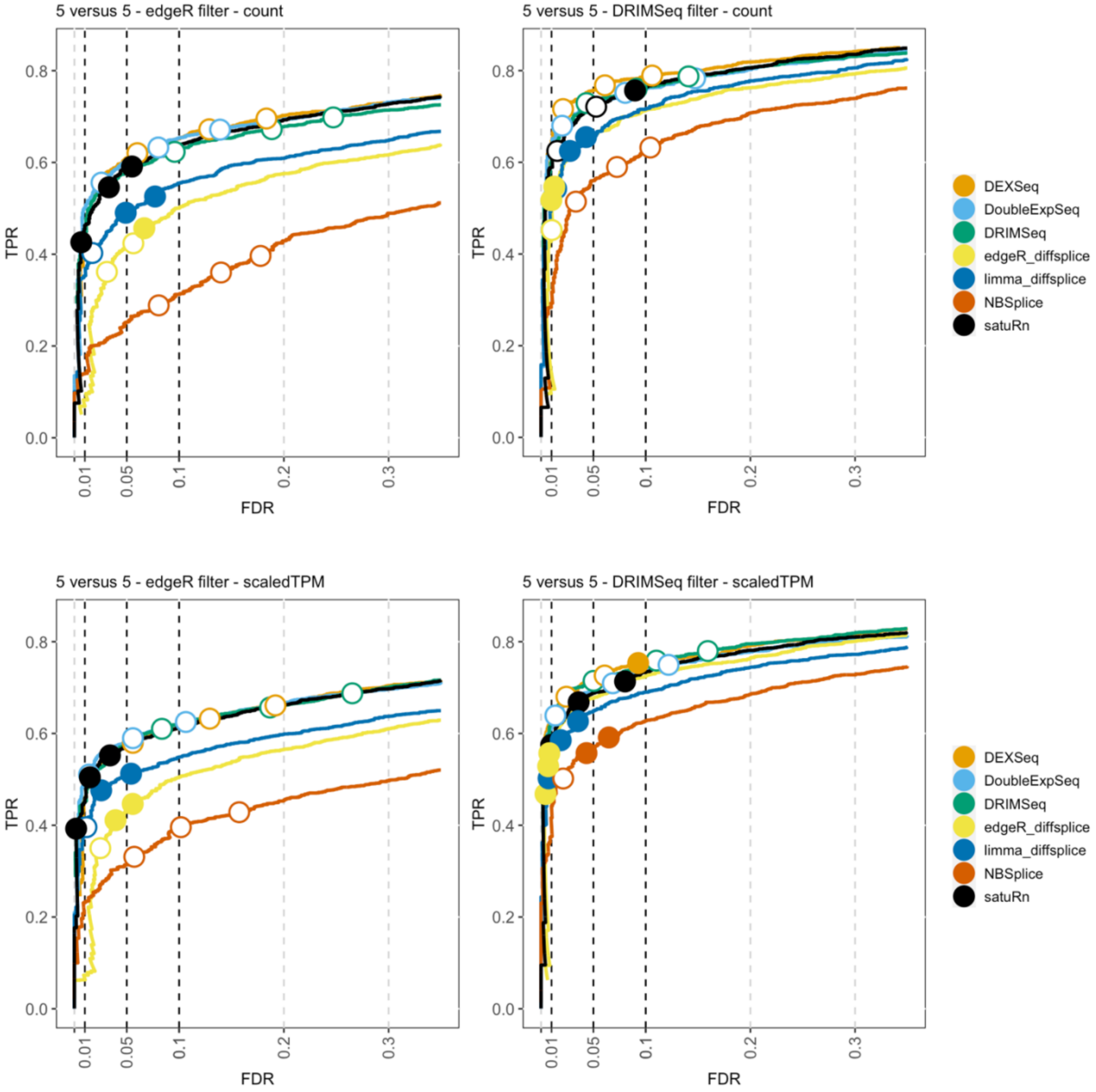
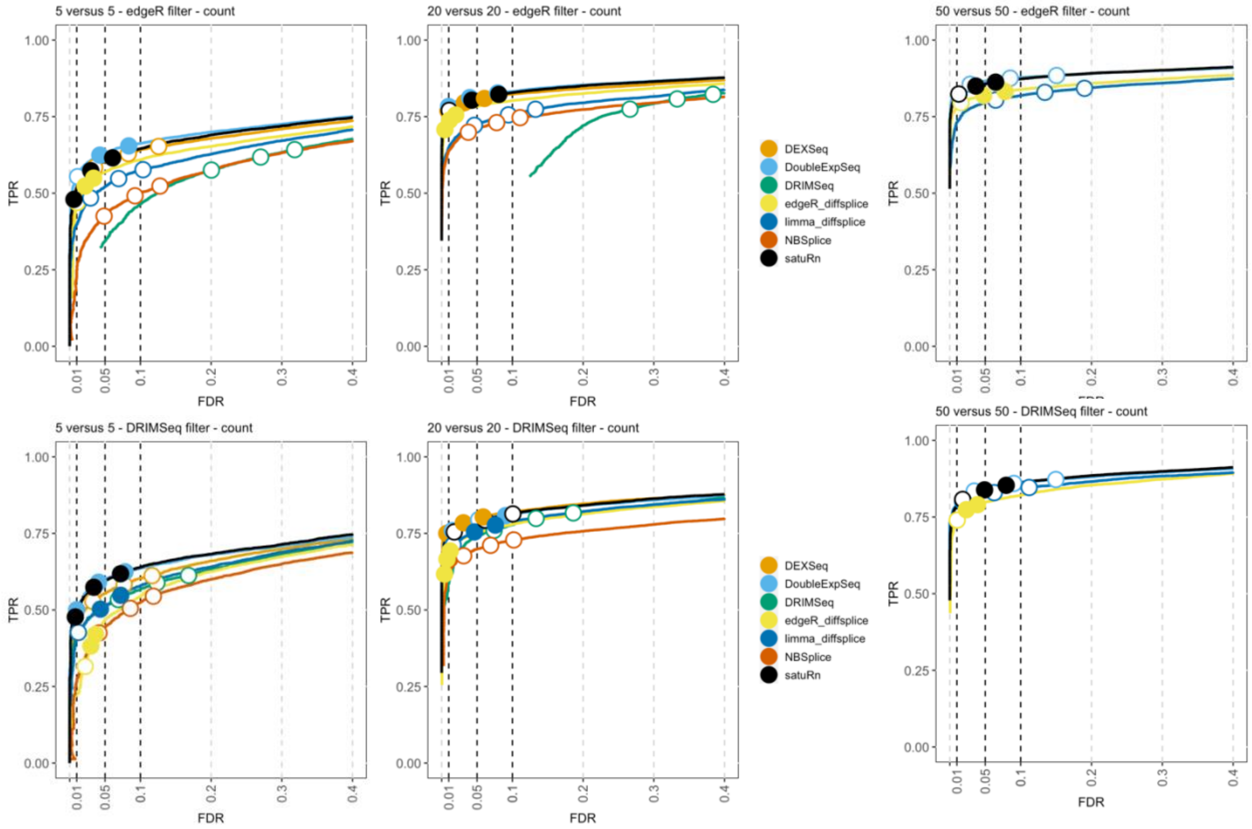


Figure S5: Performance evaluation of satuRn on the "Hsapiens" simulated bulk RNA-Seq dataset by Van den Berge et al. FDR-TPR curves visualize the performance of each method by displaying the sensitivity of the method (TPR) with respect to the false discovery rate (FDR). The three circles on each curve represent working points when the FDR level is set at nominal levels of 1%, 5% and 10%, respectively. The circles are filled if the empirical FDR is equal or below the imposed FDR threshold. The benchmark was performed both on the raw counts **(row 1)** and on scaled TPM **(row 2)** as imported with the Bioconductor R package tximport[1]. We additionally adopted two different filtering strategies; an edgeR-based filtering **(column 1)** and a DRIMSeq-based filtering **(column 2)**. Overall, the performance of satuRn is on par with those of the best tools in the literature, DEXSeq and DoubleExpSeq. In contrast to the performance evaluation on the dataset by Love et al. (Figures 1A and S2),), there is a limited difference in performances based on the data input type (i.e., counts versus scaled TPM), and DRIMSeq also performs well on these datasets.


5 versus 5 - edgeR filter - count
20 versus 20 - edgeR filter - count
50 versus 50 - edgeR filter - count
5 versus 5 - DRIMSeq filter - count
20 versus 20 - DRIMSeq filter - count
50 versus 50 - DRIMSeq filter - count
TPR
FDR
0.00
0.25
0.50
0.75
1.00
0.01
0.05
0.1
0.2
0.3
0.4
DEXSeq
DoubleExpSeq
DRIMSeq
edgeR_diffsplice
limma_diffsplice
NBSplice
satuRn

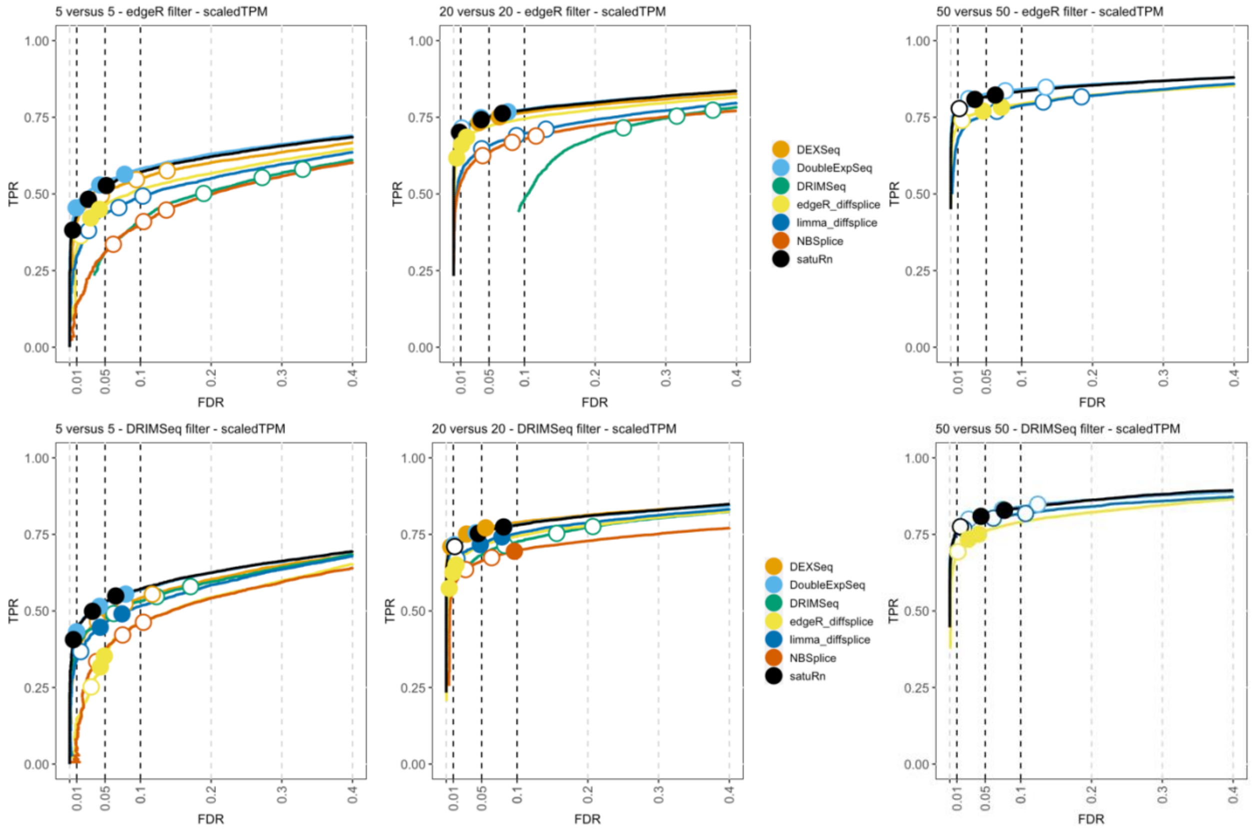

5 versus 5 - edgeR filter - scaledTPM
20 versus 20 - edgeR filter - scaledTPM
50 versus 50 - edgeR filter - scaledTPM
5 versus 5 - DRIMSeq filter - scaledTPM
20 versus 20 - DRIMSeq filter - scaledTPM
50 versus 50 - DRIMSeq filter - scaledTPM
TPR
FDR
0.00
0.25
0.50
0.75
1.00
0.01
0.05
0.1
0.2
0.3
0.4
DEXSeq
DoubleExpSeq
DRIMSeq
edgeR_diffsplice
limma_diffsplice
NBSplice
satuRn

Figure S6: Performance evaluation of satuRn on the GTEx bulk RNA-Seq dataset. FDR-TPR curves visualize the performance of each method by displaying the sensitivity (TPR) with respect to the false discovery rate (FDR). The three circles on each curve represent working points when the FDR level is set at nominal levels of 1%, 5% and 10%, respectively. The circles are filled if the empirical FDR is equal or below the imposed FDR threshold. The benchmark was performed both on the raw counts **(rows 1 and 2)** or on scaled transcripts-per-million (TPM) **(rows 3 and 4)** as imported with the Bioconductor R package tximport[1]. We additionally adopted two different filtering strategies; an edgeR-based filtering **(rows 1 and 3)** and a DRIMSeq-based filtering **(rows 2 and 4)**. The performance of satuRn is on par with the best tools from the literature, DEXSeq and DoubleExpSeq. In addition, satuRn consistently provides a stringent control of the FDR, while DoubleExpSeq becomes more liberal with increasing sample sizes. Note that DEXSeq, DRIMSeq and NBSplice were omitted from the largest comparison, as these methods do not scale to large datasets (Figure1).

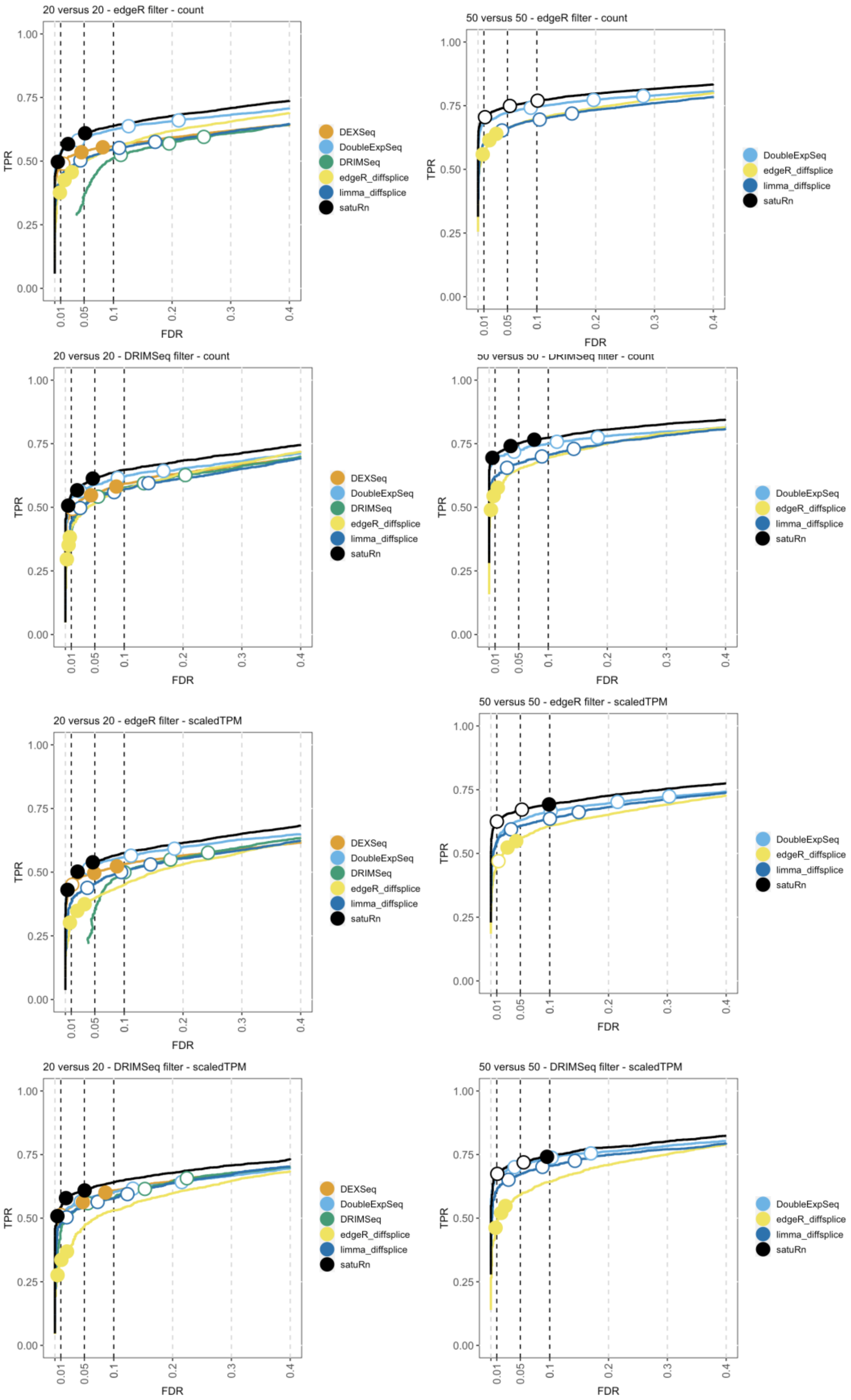
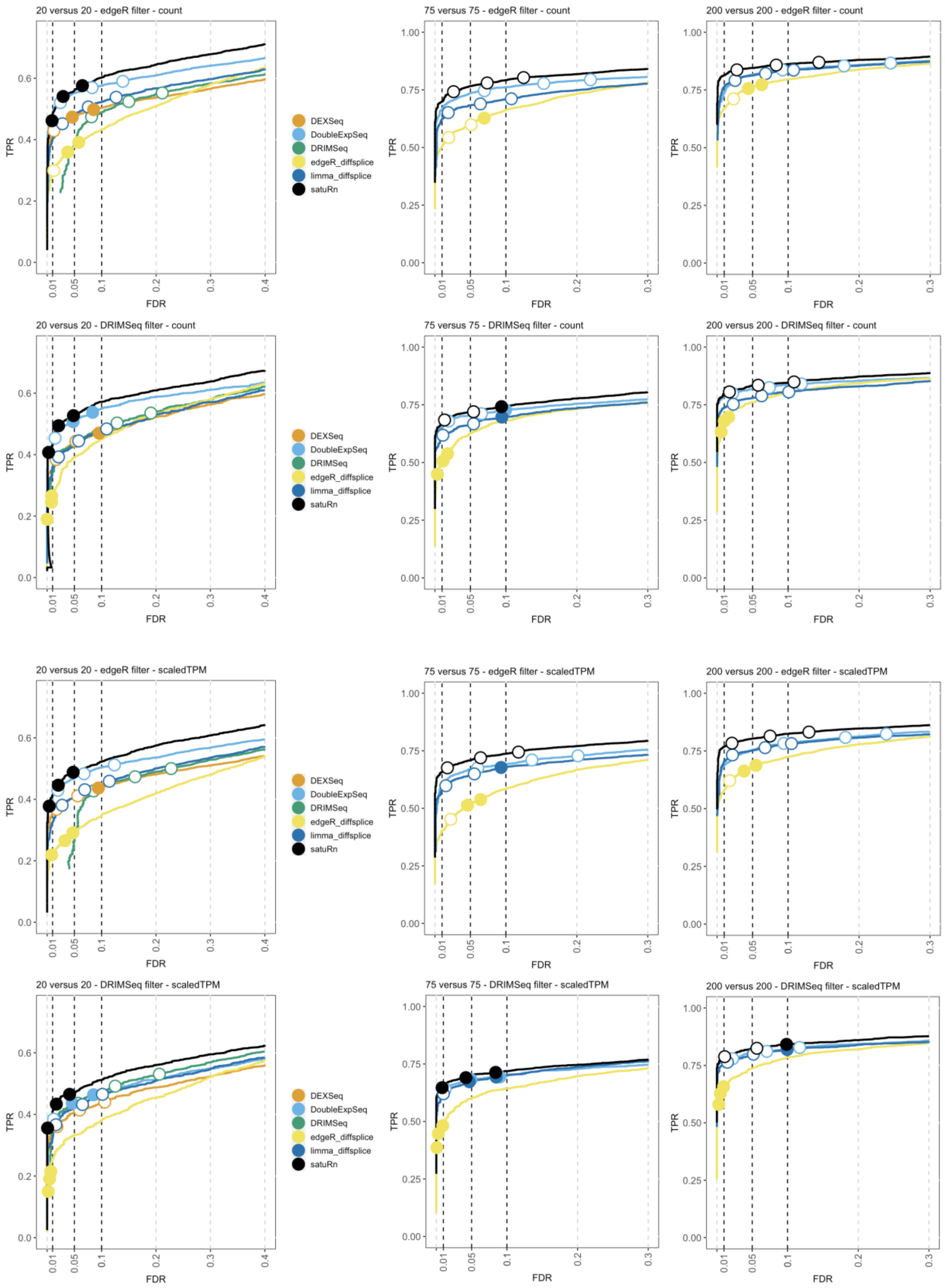

20 versus 20 - edgeR filter - count
50 versus 50 - edgeR filter - count
20 versus 20 - DRIMSeq filter - count
50 versus 50 - DRIMSeq filter - count
20 versus 20 - edgeR filter - scaledTPM
50 versus 50 - edgeR filter - scaledTPM
20 versus 20 - DRIMSeq filter - scaledTPM
50 versus 50 - DRIMSeq filter - scaledTPM
TPR
0.00
0.25
0.50
0.75
1.00
FDR
0.01
0.05
0.1
0.2
0.3
0.4
DEXSeq
DoubleExpSeq
DRIMSeq
edgeR_diffsplice
limma_diffsplice
satuRn

Figure S7: Performance evaluation of satuRn on the real scRNA-Seq dataset by Chen et al. FDR-TPR curves visualize the performance of each method by displaying the sensitivity of the method (TPR) with respect to the false discovery rate (FDR). The three circles on each curve represent working points when the FDR level is set at nominal levels of 1%, 5% and 10%, respectively. The circles are filled if the empirical FDR is equal or below the imposed FDR threshold. The benchmark was performed both on the raw counts **(rows 1 and 2)** or on scaled transcripts-per-million (TPM) **(rows 3 and 4)** as imported with the Bioconductor R package tximport[1]. We additionally adopted two different filtering strategies; an edgeR-based filtering **(rows 1 and 3)** and a DRIMSeq-based filtering **(rows 2 and 4)**. The performance of satuRn is at least on par with the best tools from the literature. Note that the performance of DEXSeq is clearly lower. In addition, our method consistently controls the FDR close to its imposed nominal FDR threshold, while DoubleExpSeq becomes more liberal with increasing sample sizes. DEXSeq and DRIMSeq were omitted from the largest comparison (two groups with 50 cells each), as these methods do not scale to large datasets (Figure 1). NBSplice was omitted from all comparisons, as it does not converge on datasets with many zeros, such as scRNA-Seq datasets.



20 versus 20 - edgeR filter - count
75 versus 75 - edgeR filter - count
200 versus 200 - edgeR filter - count
20 versus 20 - DRIMSeq filter - count
75 versus 75 - DRIMSeq filter - count
200 versus 200 - DRIMSeq filter - count
20 versus 20 - edgeR filter - scaledTPM
75 versus 75 - edgeR filter - scaledTPM
200 versus 200 - edgeR filter - scaledTPM
20 versus 20 - DRIMSeq filter - scaledTPM
75 versus 75 - DRIMSeq filter - scaledTPM
200 versus 200 - DRIMSeq filter - scaledTPM
TPR
0.0
0.2
0.4
0.6
0.00
0.25
0.50
0.75
1.00
FDR
0.01
0.05
0.1
0.2
0.3
0.4
DEXSeq
DoubleExpSeq
DRIMSeq
edgeR_diffsplice
limma_diffsplice
satuRn

Figure S8: Performance evaluation of satuRn on the real scRNA-Seq dataset by Tasic et al. FDR-TPR curves visualize the performance of each method by displaying the sensitivity of the method (TPR) with respect to the false discovery rate (FDR). The three circles on each curve represent working points when the FDR level is set at nominal levels of 1%, 5% and 10%, respectively. The circles are filled if the empirical FDR is equal or below the imposed FDR threshold. We generated three two-group comparisons of 20, 75 and 200 cells each (left, middle and right panel, respectively). The benchmark was performed both on the raw counts **(rows 1 and 2)** or on scaled transcripts-per-million (TPM) **(rows 3 and 4)** as imported with the Bioconductor R package tximport[1]. We additionally adopted two different filtering strategies; an edgeR-based filtering **(rows 1 and 3)** and a DRIMSeq-based filtering **(rows 2 and 4)**. Overall, satuRn slightly outperforms DoubleExpSeq, the best tools from the literature. Note that the performance of DEXSeq is clearly lower. In addition, our method consistently controls the FDR close to its imposed nominal FDR threshold, while DoubleExpSeq becomes more liberal with increasing sample sizes. DEXSeq and DRIMSeq were omitted from the largest comparison (two groups with 75 cells and 200 cells each, respectively), as these methods do not scale to large datasets (Figure 1). NBSplice was omitted from all comparisons, as it does not converge on datasets with many zeros, such as scRNA-Seq datasets.

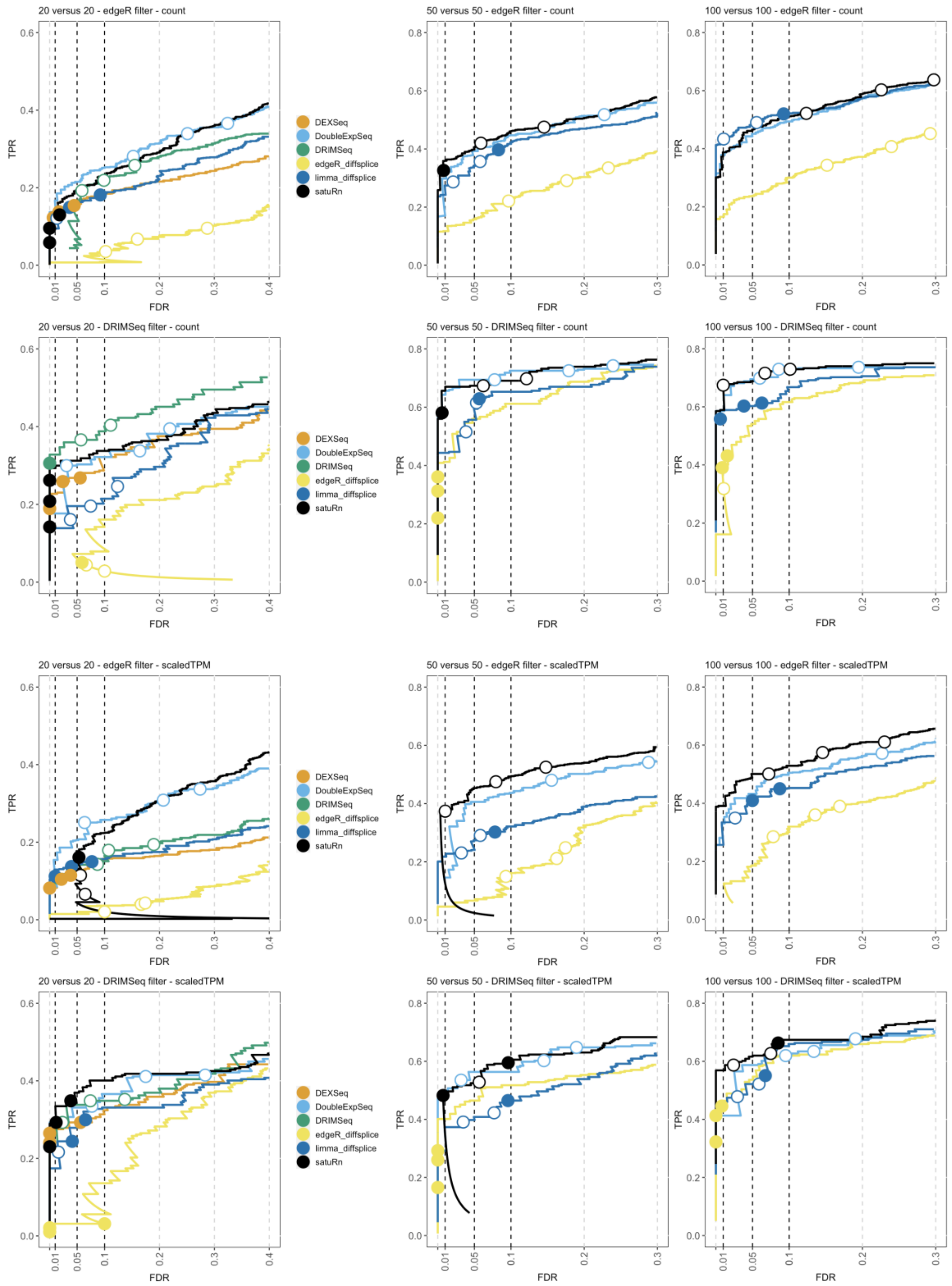

20 versus 20 - edgeR filter - count
50 versus 50 - edgeR filter - count
100 versus 100 - edgeR filter - count
20 versus 20 - DRIMSeq filter - count
50 versus 50 - DRIMSeq filter - count
100 versus 100 - DRIMSeq filter - count
20 versus 20 - edgeR filter - scaledTPM
50 versus 50 - edgeR filter - scaledTPM
100 versus 100 - edgeR filter - scaledTPM
20 versus 20 - DRIMSeq filter - scaledTPM
50 versus 50 - DRIMSeq filter - scaledTPM
100 versus 100 - DRIMSeq filter - scaledTPM
DEXSeq
DoubleExpSeq
DRIMSeq
edgeR_diffsplice
limma_diffsplice
satuRn
TPR
FDR
0.0
0.2
0.4
0.6
0.8
0.01
0.05
0.1
0.2
0.3
0.4

Figure S9: Performance evaluation of satuRn on the real scRNA-Seq dataset by Darmanis et al. FDR-TPR curves visualize the performance of each method by displaying the sensitivity of the method (TPR) with respect to the false discovery rate (FDR). The three circles on each curve represent working points when the FDR level is set at nominal levels of 1%, 5% and 10%, respectively. The circles are filled if the empirical FDR is equal or below the imposed FDR threshold. We generated three two-group comparisons of 20, 50 and 100 cells each (left, middle and right panel, respectively). The benchmark was performed both on the raw counts **(rows 1 and 2)** or on scaled transcripts-per-million (TPM) **(rows 3 and 4)** as imported with the Bioconductor R package tximport[1]. We additionally adopted two different filtering strategies; an edgeR-based filtering **(rows 1 and 3)** and a DRIMSeq-based filtering **(rows 2 and 4)**. Overall, the performance of satuRn is similar to DoubleExpSeq, the best tools from the literature. In addition, our method consistently controls the FDR close to its imposed nominal FDR threshold, while DoubleExpSeq becomes more liberal with increasing sample sizes. On the dataset with the smallest sample size, the FDR control of *satuRn* does become too strict.

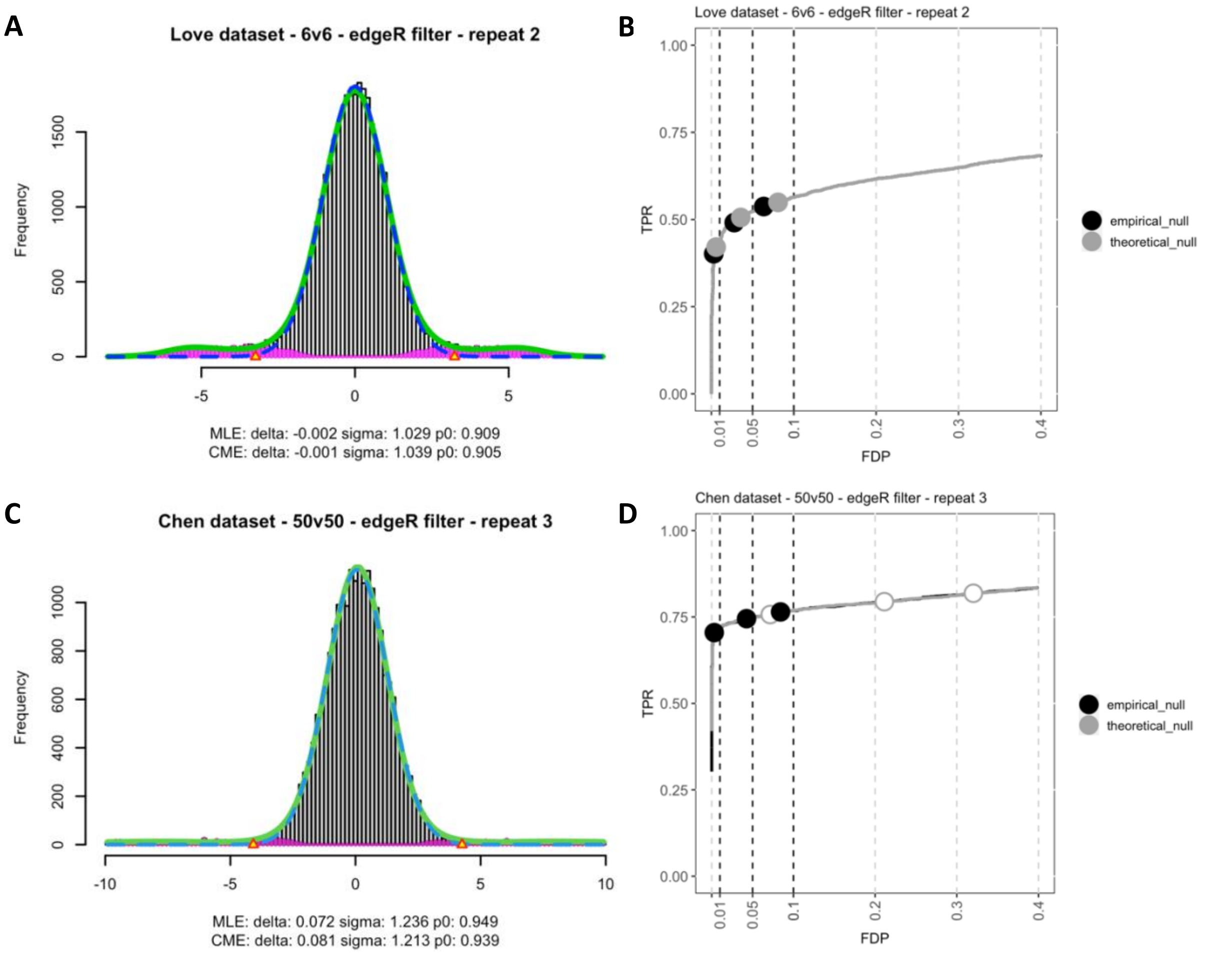


Figure S10: The effect of using an empirical null distribution on the false discovery control of satuRn. Panel A: Empirical distribution of the satuRn test statistics in one of the bulk transcriptomics benchmark datasets adapted from Love *et al***.** The test statistics are z-scores, calculated from satuRn p-values as described in formula 5 (see Methods). As this benchmark dataset is constructed to have 15% DTU transcripts and thus 85% non-DTU or null transcripts, most of these z-scores are expected to follow a standard normal distribution (mean = 0, standard deviation = 1). This is reflected in the maximum likelihood estimates for the mean and variance of the empirical null distribution (mean = -0.002, standard deviation = 1.029). **Panel B:** Corresponding FDP-TPR curve for the bulk transcriptomics benchmark dataset. As the theoretical null distribution and the empirical null distribution are virtually identical, we observe a negligible difference between both strategies, both in terms of performance and FDR control. **Panel C:** Empirical distribution of the satuRn test statistics in one of the single-cell benchmark datasets adapted from Chen *et al*. Again, most of these z-scores are expected to follow a standard normal distribution as this benchmark dataset is also constructed to have 15% DTU transcripts and thus 85% non-DTU or null transcripts. However, the empirical distribution is considerably wider than expected (standard deviation = 1.236). We additionally observe a small shift of the distribution (mean = 0.072). **Panel D:** Corresponding FDP-TPR curve for the single-cell benchmark dataset. While the inference for satuRn is overly liberal when working under the theoretical null, FDR control is restored by adopting the wider empirical null distribution. Note that the performance will only be affected when the empirical null distribution is strongly shifted with respect to the theoretical null (i.e., a large mean in absolute value), which was not the case in this example nor in any other dataset from our analyses.

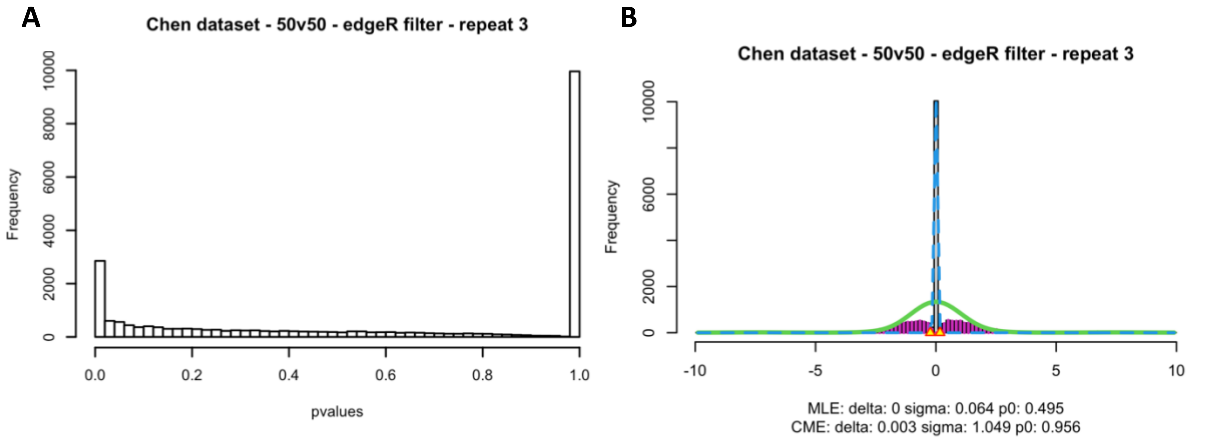


Figure S11: Adopting an empirical null distribution to improve FDR control is infeasible for DoubleExpSeq. Panel A: Distribution of the p-values from a DoubleExpSeq analysis in one of the single-cell benchmark datasets adapted from Chen *et al*. We immediately observe the large spike of p-values equal to 1, which distorts the p-value distribution. In addition, the p-values in the mid-range (e.g., from 0.1 to 0.9), which are expected to be uniformly distributed, are skewed towards smaller values, which underlies the overly liberal results of DoubleExpSeq in our single-cell benchmarks. **Panel B:** The corresponding empirical distribution of the DoubleExpSeq test statistics. The test statistics are z-scores, calculated from the original DoubleExpSeq p-values as described in formula 5 (see Methods). As all our benchmark datasets are constructed to have 15% DTU transcripts and thus 85% non-DTU or null transcripts, most of these z-scores are expected to follow a standard normal distribution (mean = 0, standard deviation =1). However, given the pathological distribution of the p-values it is not feasible to properly estimate the empirical null distribution, as also clearly shown by the widely different parameter estimates obtained using the two estimation frameworks implemented in the *locfdr* R package[2]; compare the estimates between MLE (maximum likelihood estimation) and CME (central matching estimation).

Comparison	Cell type 1 (ALM)	Cell type 2 (VISp)	DoubleExpSeq FDR	Limma FDR	Limma Empirical FDR
1	Cpa6 Gpr88	Batf3	2142	3602	169
2	Cbln4 Fezf2	Col27a1	644	468	297
3	Cpa6 Gpr88	Col6a1 Fezf2	335	1029	77
4	Gkn1 Pcdh19	Col6a1 Fezf2	1878	2861	58
5	Lypd1 Gpr88	Hsd11b1 Endou	829	1411	249
6	Tnc	Hsd11b1 Endou	4580	4819	341
7	Tmem163 Dmrtb1	Hsd11b1 Endou	3388	5603	176
8	Tmem163 Arhgap25	Whrn Tox2	455	1387	166

Figure S12: Number of differentially used transcripts as identified by DoubleExpSeq and limma diffsplice. The first three columns indicate the comparisons between ALM cell types (column 2) and VISp cell types (column 3), respectively. Column 4 indicates the number of differentially used transcripts as identified by DoubleExpSeq. Column 5 indicates the number of differentially used transcripts as identified by a limma diffsplice analysis with default settings. Column 6 displays the number of differentially used transcripts found by limma diffsplice after correcting for deviations between the theoretical and empirical null distributions.

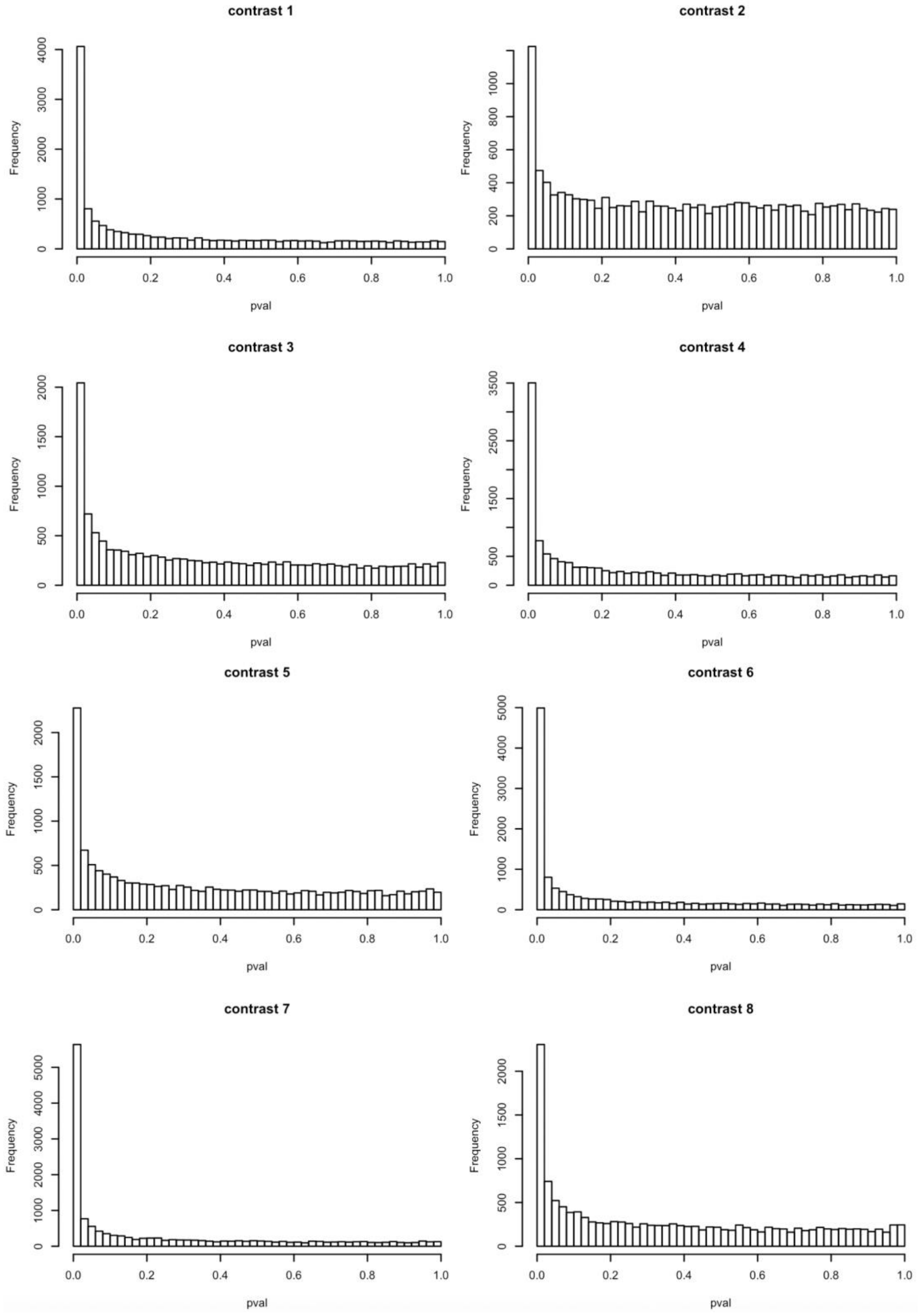


Figure S13: Histograms of the p-values from limma diffsplice. From these histograms, the huge number of DTU transcripts identified by limma diffsplice become apparent. Note that the general tendency of limma diffsplice for smaller p-values is better visible when converting the p-values into z-scores (see Figure S13)

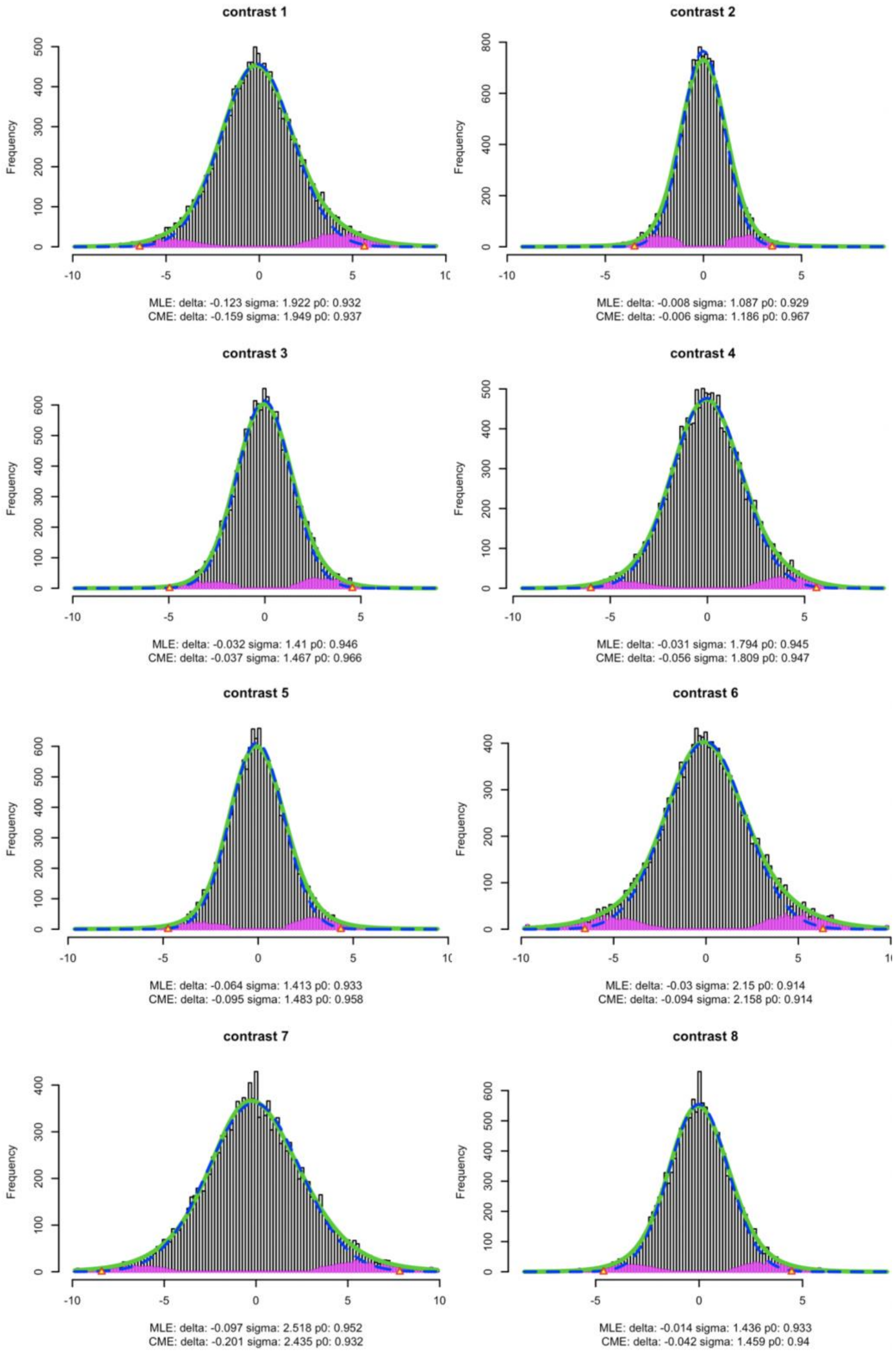


Figure S14: Empirical distribution of the limma diffsplice test statistics. The test statistics are z-scores, calculated from limma diffsplice p-values as described in formula 5. Theoretically, these z-scores are expected to follow a standard normal distribution (mean = 0, standard deviation =1). Here, however, the empirical distributions are considerably wider (standard deviation > 1), as indicated underneath the plots. This indicates that the results returned by limma diffsplice in this case study are overly liberal.

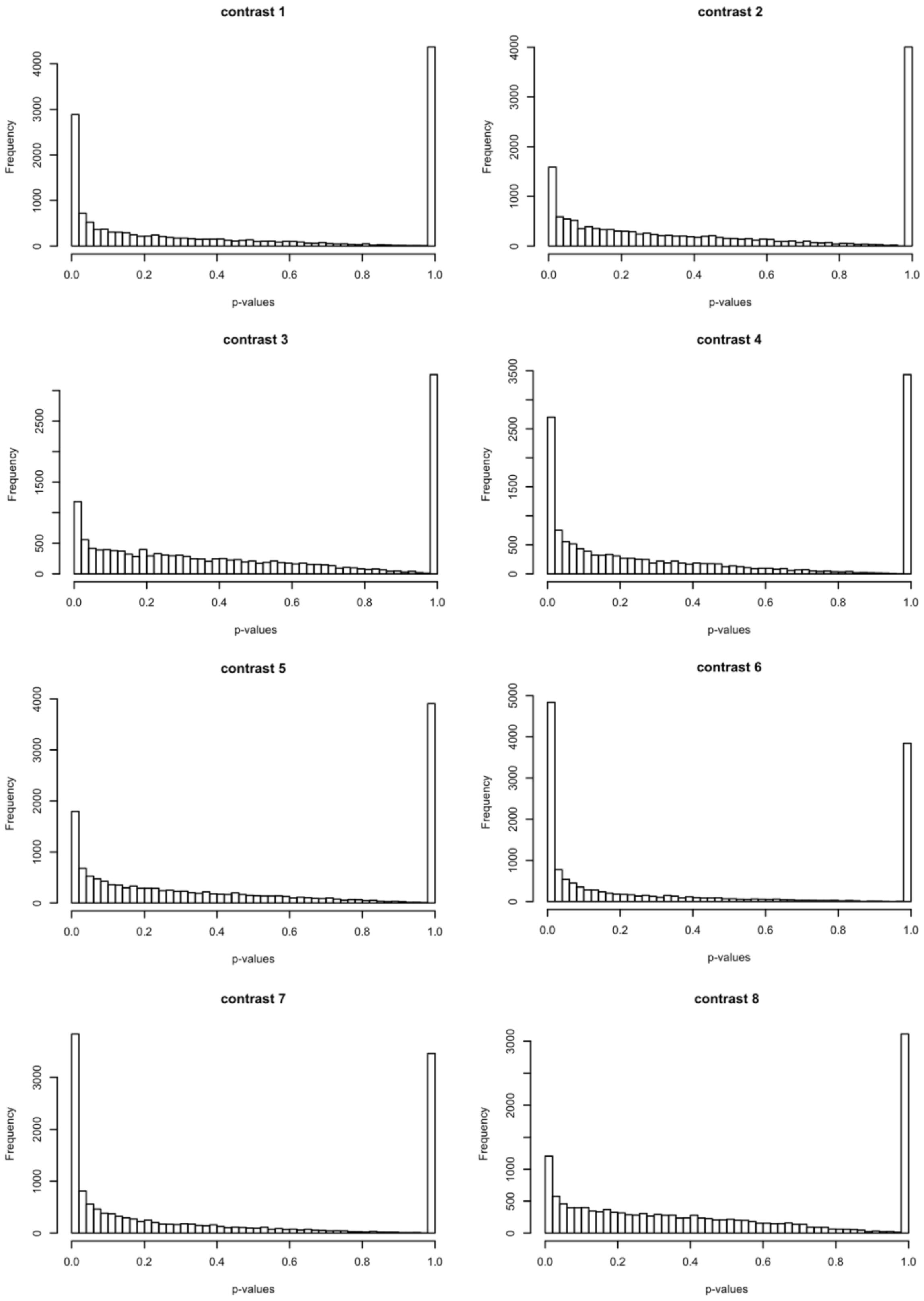


Figure S15: Histograms of the p-values from DoubleExpSeq. From these histograms, the huge number of DTU transcripts identified by limma diffsplice become apparent. In addition, we observe a gradual decrease of p-values over the interval [0.05 < p < 0.95], with a very large spike of p-values that are exactly 1 in all comparisons or contrasts of interest.

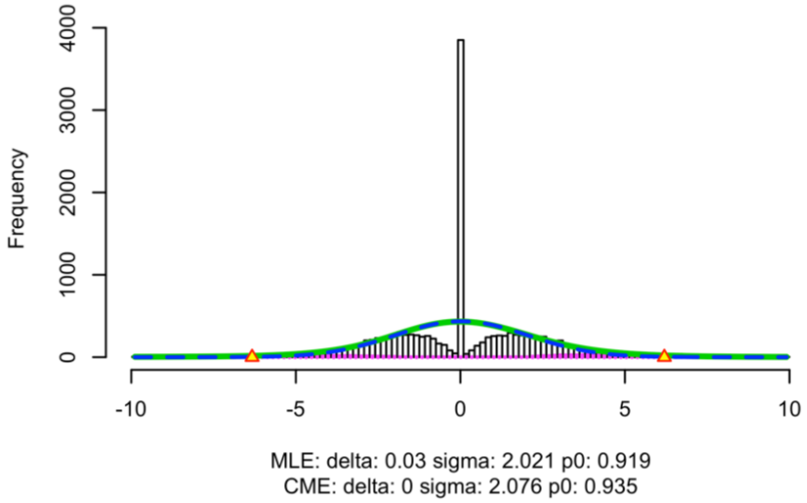


Figure S16: Empirical distribution of the test statistics in comparison #6 of the case study with DoubleExpSeq. The test statistics are z-scores, calculated from DoubleExpSeq p-values as described in formula 5 (see Methods). Theoretically, the bulk of these z-scores are expected to follow a standard normal distribution (mean = 0, standard deviation =1), i.e., assuming that most transcripts are not differentially used. However, the large spike of p-values equal to 1 (See Figure S14) results spike of z-scores equal to zero, which poses a problem when estimating the empirical null distribution (blue dashed curve).

References

1. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* **4**, 1521 (2016).
2. Efron, B., Turnbull, B. B. & Narasimhan, B. Locfdr: Computes Local False Discovery Rates. *R Packag.* **Version 1.**, http://CRAN.R-project.org/package=locfdr (2011).