

Supplementary Fig. 1 | A systematic benchmark of differential expression in single-cell transcriptomics.

a, Impact of varying the parameter k on the AUCC in the eighteen ground-truth datasets, as shown in **Fig. 1c** with $k = 500$.

b, Impact of varying the parameter k on the Δ AUCC in the eighteen ground-truth datasets, as shown in **Fig. 1d** with $k = 500$.

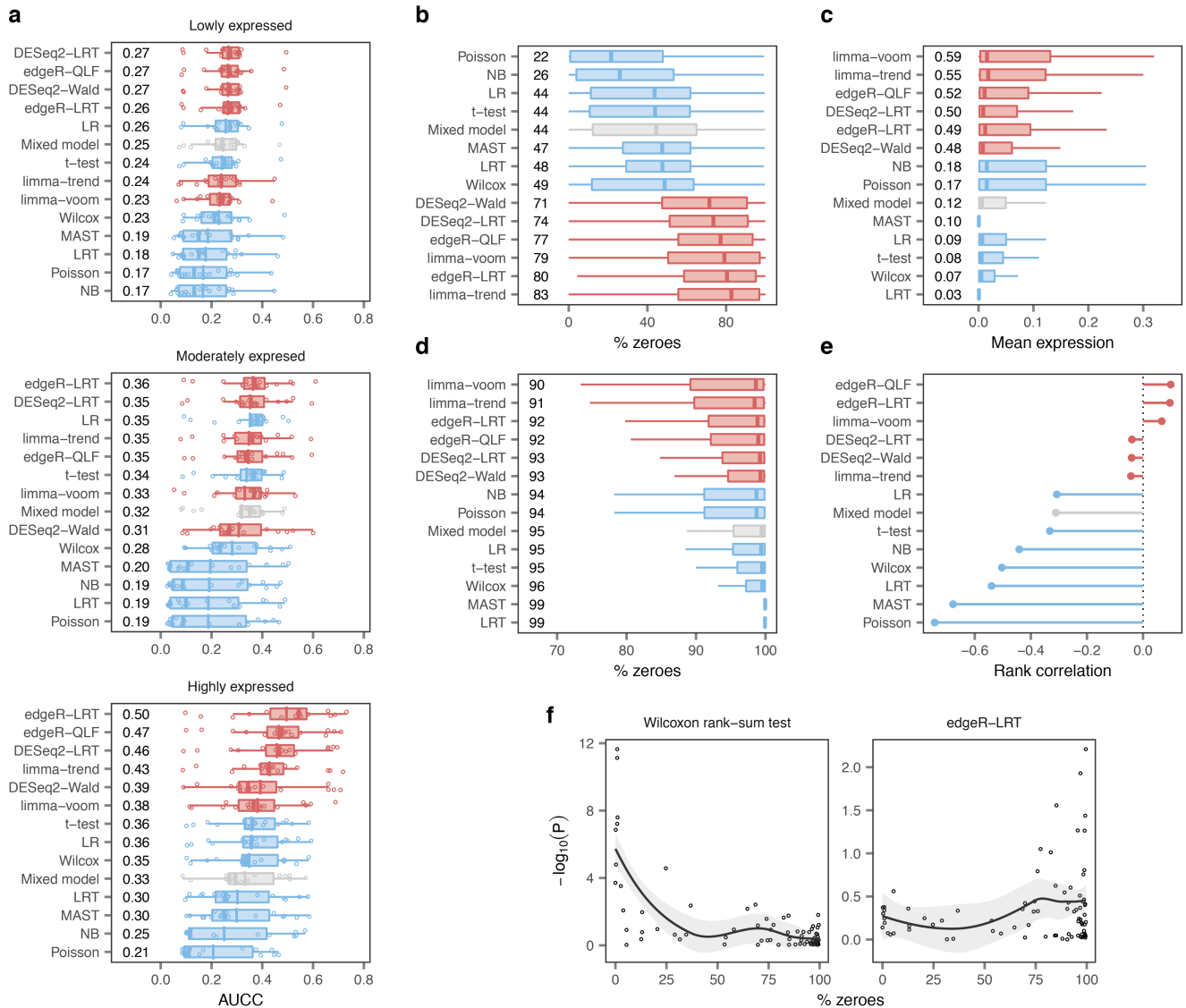
c, Transcriptome-wide rank correlation between single-cell and bulk RNA-seq in the eighteen ground-truth datasets shown in **a**.

d, Mean difference in the transcriptome-wide rank correlation (Δ correlation) between the fourteen DE methods shown in **c**. Asterisks indicate comparisons with a two-tailed t-test p-value less than 0.05.

e, AUCC in eight scRNA-seq datasets with matching bulk proteomics data¹.

f, Mean Δ AUCC between the fourteen DE methods shown in **e**. Asterisks indicate comparisons with a two-tailed t-test p-value less than 0.05.

g, Mean Δ AUCC of GO term enrichment between the fourteen DE methods shown in **Fig. 1e**. Asterisks indicate comparisons with a two-tailed t-test p-value less than 0.05.



Supplementary Fig. 2 | Single-cell DE methods are biased towards highly expressed genes.

a, AUCCs across eighteen ground-truth datasets after dividing the transcriptome into terciles of lowly (top), moderately (middle), or highly (bottom) expressed genes, as shown in **Fig. 2b**.

b, Mean proportion of zero gene expression measurements for the 100 top-ranked false-positive genes from each DE method.

c, Mean expression levels of the 100 top-ranked false-negative genes from each DE method.

d, Mean proportion of zero gene expression measurements for the 100 top-ranked false-negative genes from each DE method.

e, Spearman correlation between the mean proportion of zero gene expression measurements for 80 ERCC spike-ins expressed in at least three cells and the $-\log_{10}$ p-value of differential expression assigned by each DE method.

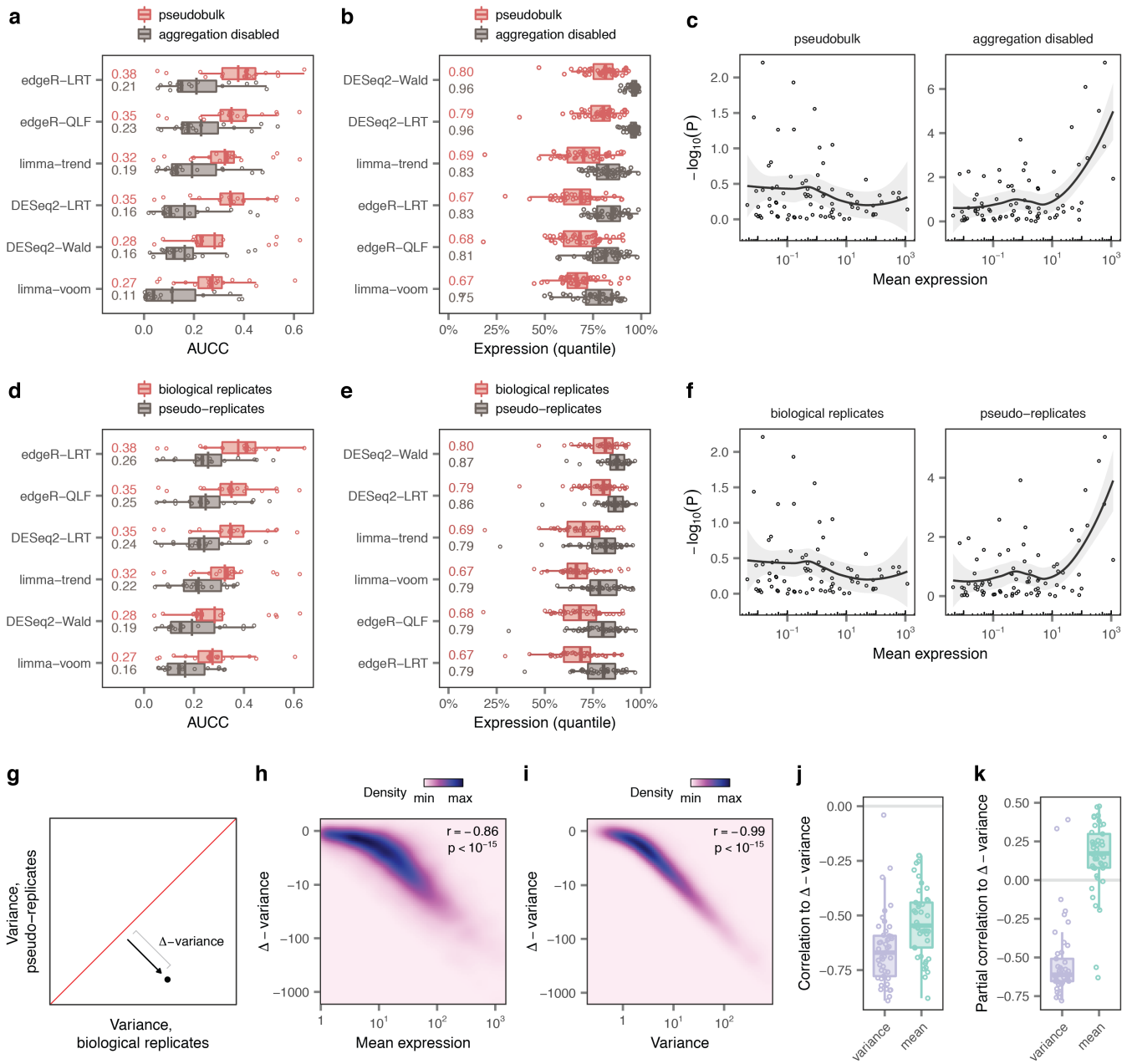
f, Scatterplots of mean proportion of zero gene expression measurements vs. $-\log_{10}$ p-value for exemplary single-cell and pseudobulk DE methods. Trend lines and shaded areas show local polynomial regression and the 95% confidence interval, respectively.



Supplementary Fig. 3 | Overview of single-cell transcriptomics datasets.

a, Overview of $n = 46$ published scRNA-seq datasets comparing two or more experimental conditions, used to systematically confirm the universality of the trends observed in analyses of individual datasets. Left, heatmap indicating the species of origin, the sequencing protocol, and whether cells or nuclei were sequenced. Right, properties of each dataset, including the total number of cell types identified in the original studies; the total number of cells sequenced; the number of cells per type (red bars indicate mean); and the mean number of reads for cells of each type. Datasets highlighted in grey contain matching bulk data and contributed to the 18 ground truth datasets shown in **b**.

b, Overview of $n = 18$ ground-truth datasets with matching scRNA-seq and bulk data, used to evaluate the biological accuracy of single-cell DE methods. Left, heatmap indicating the species of origin, the cell type under investigation and the perturbation to which it was exposed, the sequencing protocol, and whether cells or nuclei were sequenced. Right, properties of each dataset, including the total number of cells and and the number of reads per cell.



Supplementary Fig. 4 | DE analysis in single-cell data must account for biological replicates.

a, AUCC of the six pseudobulk methods applied to pseudobulks or individual cells in the eighteen ground-truth datasets.

b, Mean expression levels of the 200 top-ranked genes from six pseudobulk methods applied to pseudobulks or individual cells in a collection of 46 scRNA-seq datasets.

c, Scatterplots of mean ERCC expression vs. $-\log_{10}$ p-value for an exemplary pseudobulk method, edgeR-LRT, applied to pseudobulks (left) or individual cells (right). Trend lines and shaded areas show local polynomial regression and the 95% confidence interval, respectively.

d, AUCC of the six pseudobulk methods applied to pseudobulks or pseudo-replicates in the eighteen ground-truth datasets.

e, Mean expression levels of the 200 top-ranked genes from six pseudobulk methods applied to pseudobulks or pseudo-replicates in a collection of 46 scRNA-seq datasets.

f, Scatterplots of mean ERCC expression vs. $-\log_{10}$ p-value for an exemplary pseudobulk method, edgeR-LRT, applied to pseudobulks (left) or pseudo-replicates (right). Trend lines and shaded areas show local polynomial regression and the 95% confidence interval, respectively.

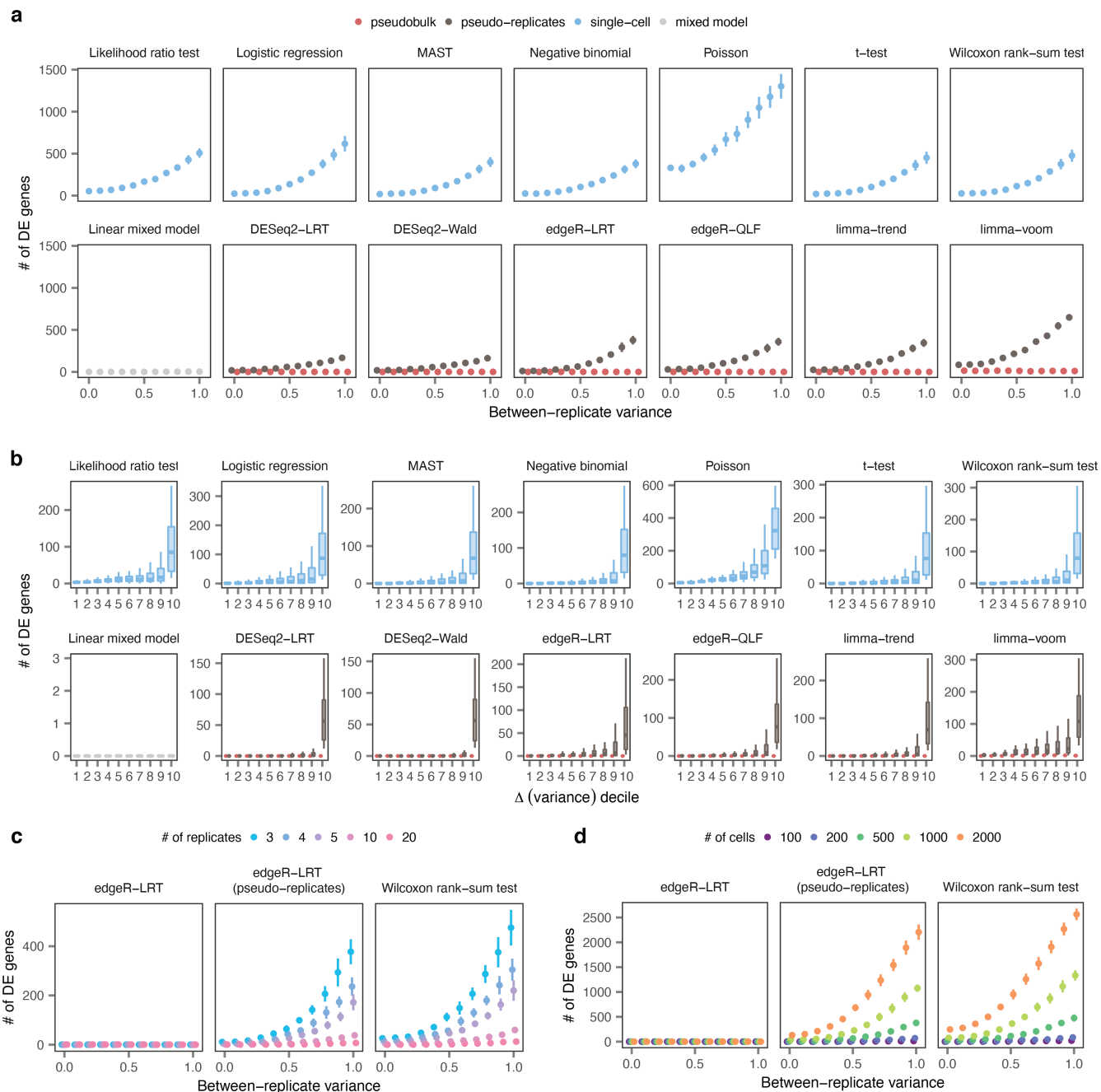
g, Schematic illustrating the calculation of the Δ -variance between biological replicates and pseudo-replicates.

h, Correlation between mean expression and Δ -variance for 10,448 genes with mean expression ≥ 1 CPM in the dataset of mouse bone marrow mononuclear cells stimulated with poly-I:C. Mean expression is strongly correlated with Δ -variance, such that the variance of highly expressed genes is disproportionately underestimated when ignoring information about biological replicates.

i, Correlation between expression variance and Δ -variance for 10,448 genes with mean expression ≥ 1 CPM in the dataset of mouse bone marrow mononuclear cells stimulated with poly-I:C. Variance is even more strongly correlated with Δ -variance than mean expression, such that the most variable genes are disproportionately underestimated when ignoring information about biological replicate.

j, Correlation between mean expression levels or expression variance and Δ -variance in 46 scRNA-seq datasets. Variance is even more strongly correlated with Δ -variance than mean expression across a large compendium of datasets, corroborating the trends shown in **h-i**.

k, Partial correlation between mean expression and Δ -variance, controlling for variance, or between variance and Δ -variance, controlling for mean expression. The variance of gene expression is the primary determinant of Δ -variance, implying that failing to account for biological replicates introduces a bias towards highly expressed genes because these genes are also more variable.



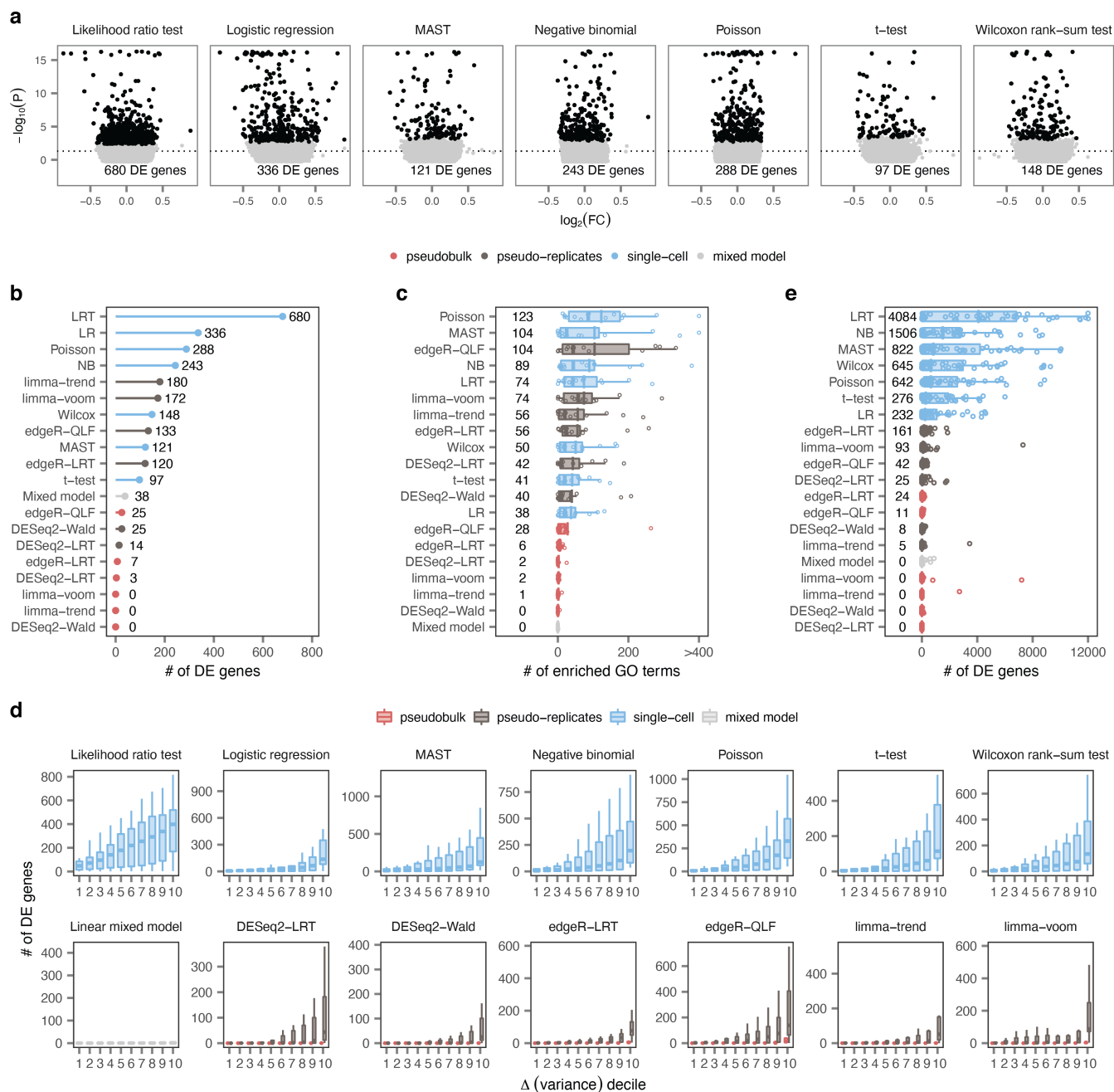
Supplementary Fig. 5 | Simulation studies expose false discoveries in single-cell DE.

a, Number of DE genes detected in stimulation experiments with varying degrees of heterogeneity between replicates by all DE methods. Points and error bars show the mean and standard deviation of ten independent simulations.

b, Number of DE genes detected by the tests shown in **a** for genes divided into deciles by the magnitude of the change in variance between biological replicates and pseudo-replicates (Δ -variance).

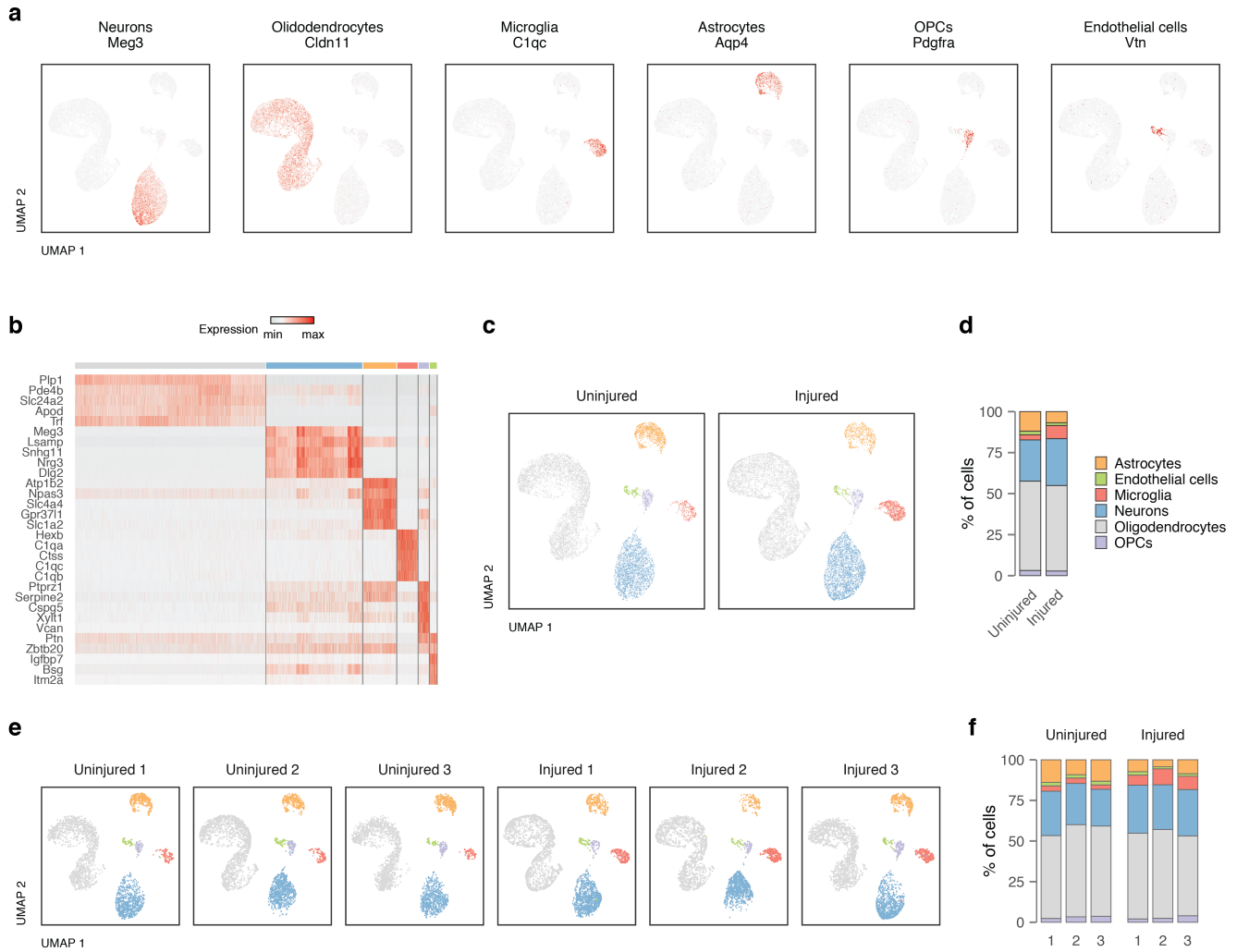
c, Number of DE genes detected by a representative single-cell DE method, a representative pseudobulk method, and the same pseudobulk method applied to pseudo-replicates, when varying the total number of replicates in the simulated dataset. Points and error bars show the mean and standard deviation of ten independent simulations.

d, Number of DE genes detected by a representative single-cell DE method, a representative pseudobulk method, and the same pseudobulk method applied to pseudo-replicates, when varying the total number of cells in the simulated dataset. Points and error bars show the mean and standard deviation of ten independent simulations.



Supplementary Fig. 6 | False discoveries in single-cell and spatial transcriptomics data.

- a**, Volcano plots showing DE between T cells from random groups of unstimulated controls drawn from Kang et al.² using seven single-cell DE methods.
- b**, Number of DE genes detected by all DE methods in unstimulated T cells.
- c**, Number of GO terms enriched at 5% FDR among DE genes identified in comparisons of random groups of unstimulated controls from fourteen scRNA-seq studies with at least six control samples.
- d**, Number of DE genes in comparisons of random groups of unstimulated controls from fourteen scRNA-seq studies with at least six control samples, as shown in Fig. 4e, for genes divided into deciles by the magnitude of the change in variance between biological replicates and pseudo-replicates (Δ -variance).
- e**, Number of DE genes detected by all DE methods within spinal cord regions from control mice profiled by spatial transcriptomics³.



Supplementary Fig. 7 | Single-nucleus RNA-seq of the injured mouse lumbar spinal cord.

a, Expression of key marker genes for the six major cell types of the lumbar spinal cord across 19,237 individual nuclei.

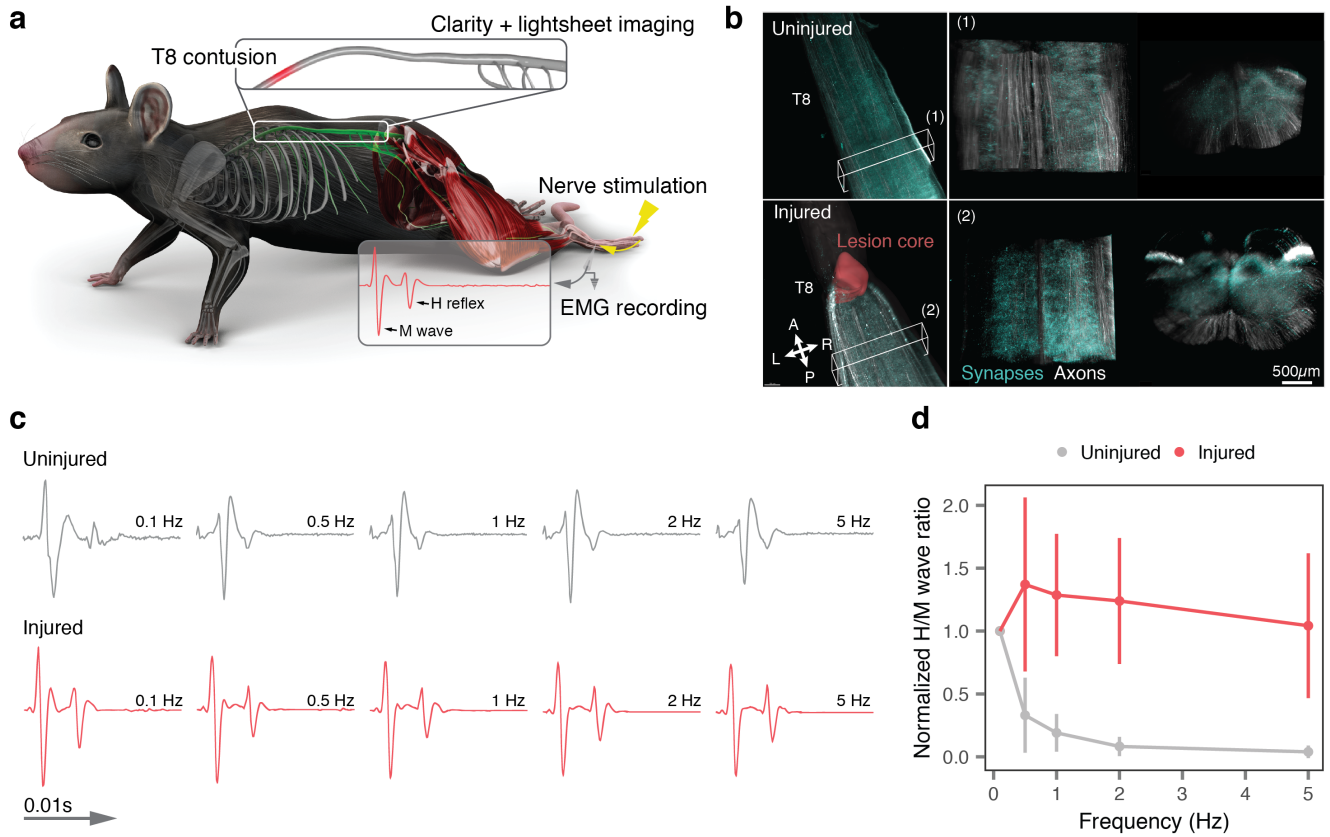
b, Top five marker genes for each major cell type of the lumbar spinal cord.

c, UMAP visualization of 19,237 nuclei, showing detection of the major cell types of the lumbar spinal cord across experimental conditions.

d, Proportion of cells of each major cell type detected in either experimental condition.

e, UMAP visualization of 19,237 nuclei, showing detection of the major cell types of the lumbar spinal cord across individual replicates.

f, Proportion of cells of each major cell type detected in each individual replicate.



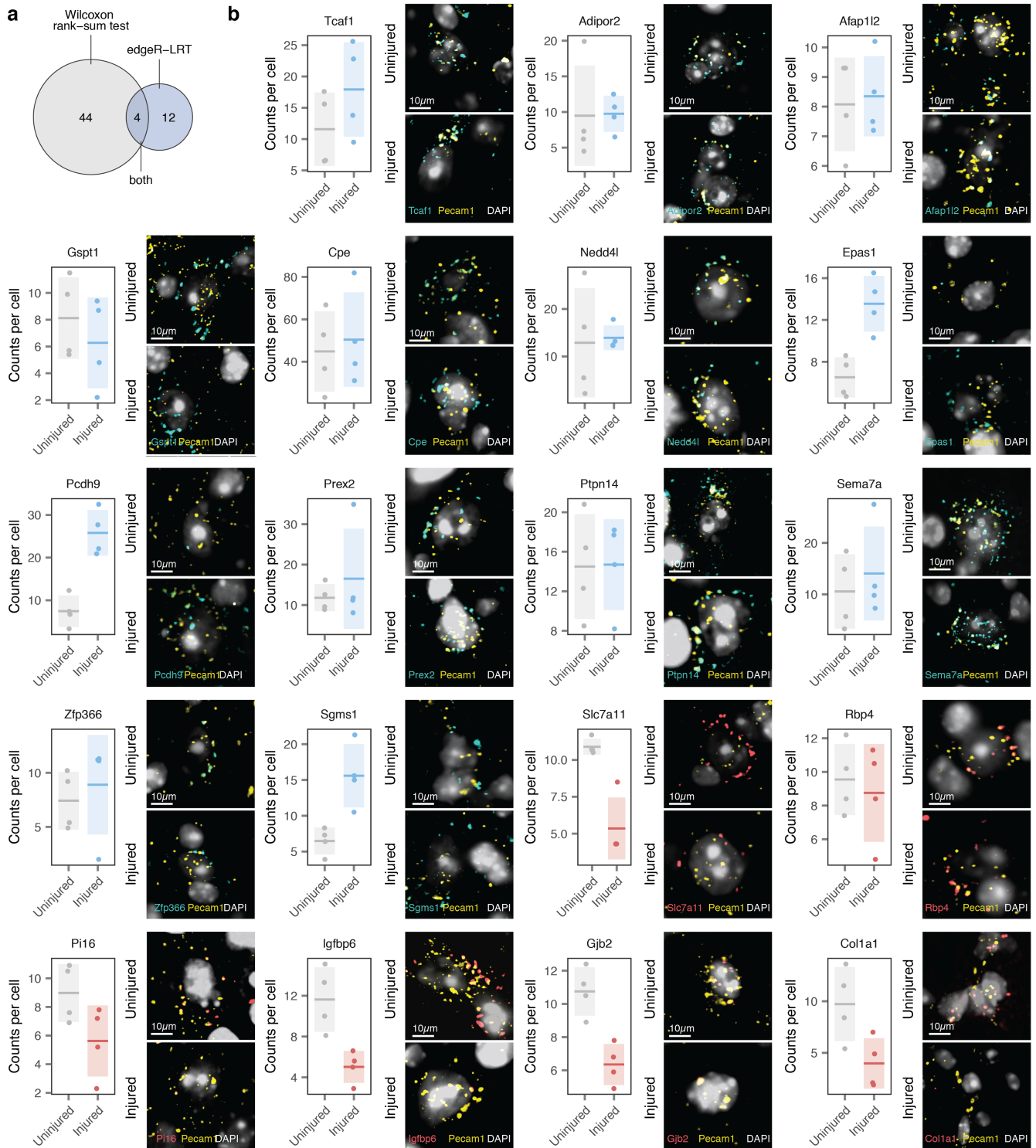
Supplementary Fig. 8 | Spasticity and hyperexcitability in the injured mouse lumbar spinal cord.

a, Schematic illustration of electrophysiology experimental design. Inset demonstrates persistent amplitude at increasing frequency of the H-reflex in response to plantar stimulation, reflecting hyperexcitability.

b, Synapses and projections of *Vglut2^{ON}* neurons in the mouse lumbar spinal cord before and after SCI.

c, Individual traces from representative injured and uninjured mice at each tested frequency, including 0.1, 0.5, 1, 2, and 5 Hz. We observed a persistent H-reflex in only injured animals, indicating hyperexcitability.

d, Quantification of electrophysiology traces as shown in **c**, demonstrating persistence of the H-reflex in the injured group. Points and error bars show mean and standard deviation, respectively, for $n = 6$ biologically independent animals examined with 12 repeated measures for each condition. Source data are provided as a Source Data file.

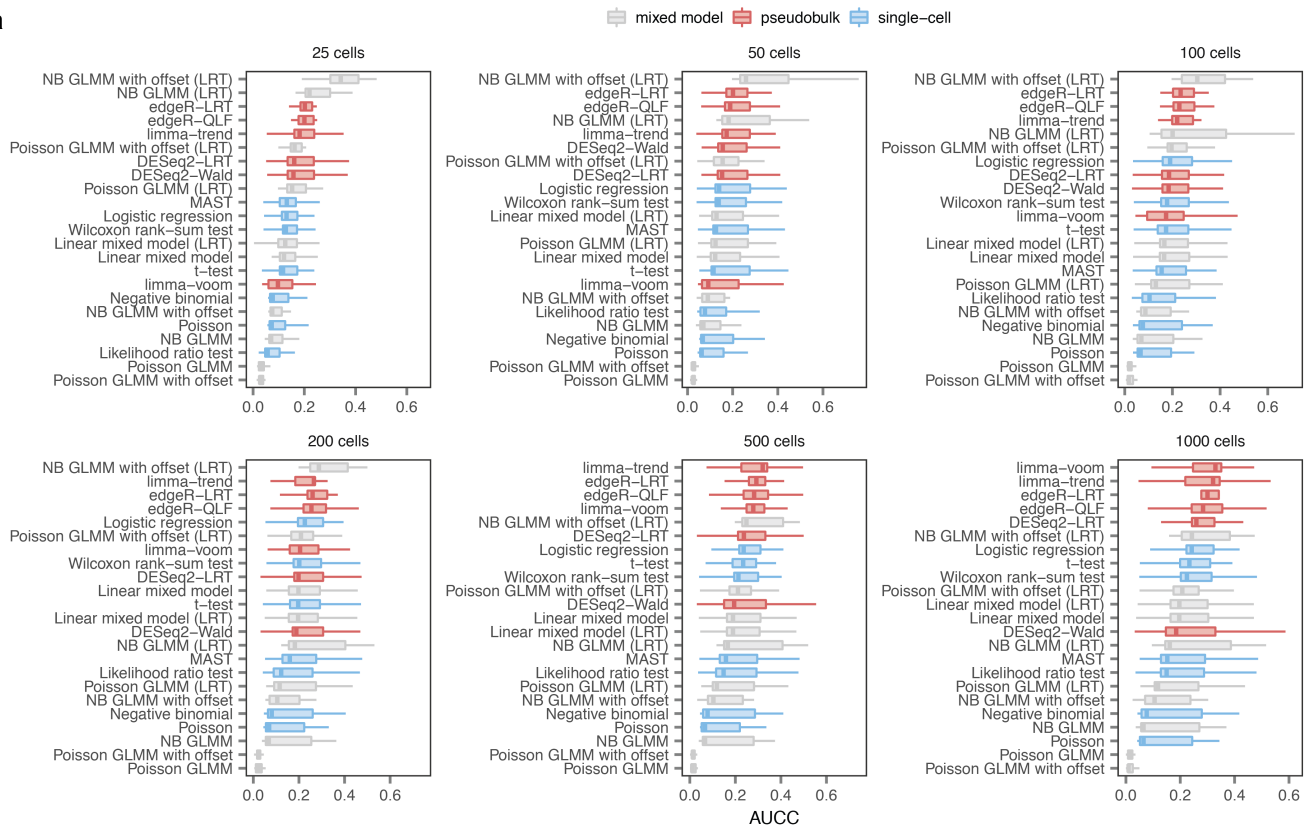


Supplementary Fig. 9 | *In vivo* validation of single-cell DE analysis by RNAscope.

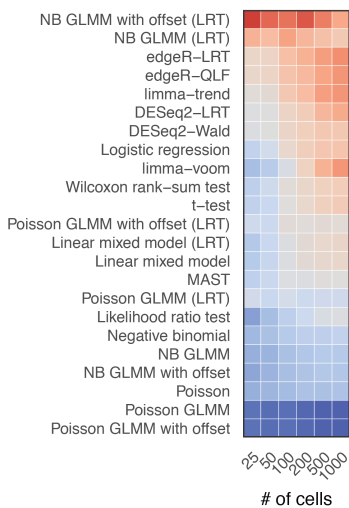
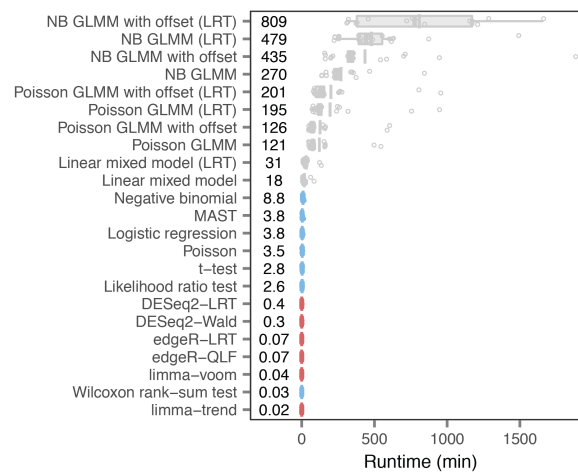
a, Overlap between genes identified as DE within endothelial cells of the injured mouse lumbar spinal cord by edgeR-LRT and the Wilcoxon rank-sum test.

b, RNAscope quantification, left, and representative images, right, for nineteen genes identified as DE exclusively by edgeR-LRT or the Wilcoxon rank-sum test. Horizontal line and shaded area show the mean and standard deviation, respectively.

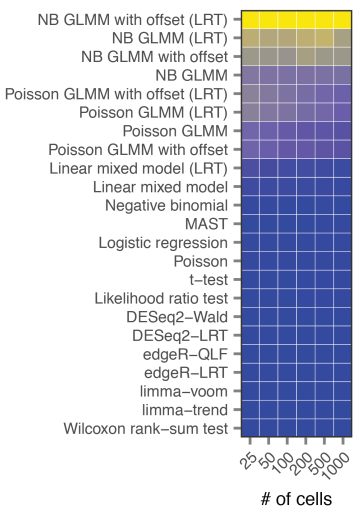
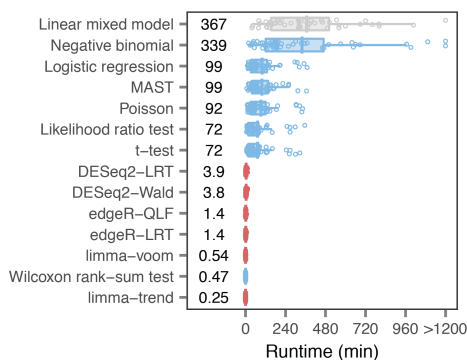
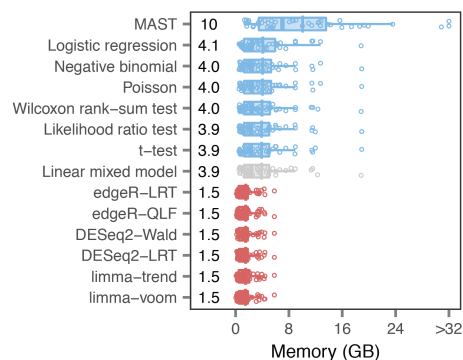
Source data are provided as a Source Data file.

a**b**

AUC 0.0 0.4

**c****d**

Runtime, % of max 0 100

**e****f**

Supplementary Fig. 10 | Single-cell DE analysis with generalized linear mixed models.

a, AUCC for ten different generalized linear mixed models (GLMMs), varying in the choice of link function (identity, Poisson, or negative binomial, NB); method used to evaluate statistical significance (Wald test or likelihood ratio test, LRT), and presence of an offset term, in samples of between 25 and 1,000 cells from the eighteen ground-truth datasets shown in **Fig. 1c**, and compared to the fourteen DE methods shown in the same panel.

b, As in **a**, but showing the mean AUCC as a function of the number of cells sampled for each DE method.

c, Runtime in minutes for the ten GLMMs shown in **a** in samples of 1,000 cells. The top-performing GLMM required a mean of 13.5 h per cell type to perform DE analysis.

d, Runtime of the ten GLMMs and the fourteen DE methods shown in **Fig. 1c**, shown as a percentage of the maximum runtime, as a function of the number of cells sampled.

e, Runtime in minutes of the fourteen DE methods shown in **Fig. 1c** across 46 scRNA-seq datasets.

f, Maximum memory required in gigabytes by the fourteen DE methods shown in **Fig. 1c** across 46 scRNA-seq datasets.

References

1. Cano-Gamez, E. *et al.* Single-cell transcriptomics identifies an effectorness gradient shaping the response of CD4+ T cells to cytokines. *Nat. Commun.* **11**, 1801 (2020).
2. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
3. Maniatis, S. *et al.* Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science* **364**, 89–93 (2019).