nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-25960-2 OPEN

Confronting false discoveries in single-cell differential expression

Jordan W. Squair[1,2,3], Matthieu Gautier[1,2], Claudia Kathe[1,2], Mark A. Anderson[1,2], Nicholas D. James[1,2], Thomas H. Hutson[1,2], Rémi Hudelle[1,2], Taha Qaiser[3], Kaya J. E. Matson[4], Quentin Barraud[1,2], Ariel J. Levine[4], Gioele La Manno[1], Michael A. Skinnider[1,2,5,6]✉ & Grégoire Courtine[1,2,6]✉

Differential expression analysis in single-cell transcriptomics enables the dissection of cell-type-specific responses to perturbations such as disease, trauma, or experimental manipulations. While many statistical methods are available to identify differentially expressed genes, the principles that distinguish these methods and their performance remain unclear. Here, we show that the relative performance of these methods is contingent on their ability to account for variation between biological replicates. Methods that ignore this inevitable variation are biased and prone to false discoveries. Indeed, the most widely used methods can discover hundreds of differentially expressed genes in the absence of biological differences. To exemplify these principles, we exposed true and false discoveries of differentially expressed genes in the injured mouse spinal cord.

[1] Center for Neuroprosthetics and Brain Mind Institute, Faculty of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. [2] NeuroRestore, Department of Clinical Neuroscience, Lausanne University Hospital (CHUV) and University of Lausanne (UNIL), Lausanne, Switzerland. [3] International Collaboration on Repair Discoveries (ICORD), University of British Columbia, Vancouver, BC, Canada. [4] Spinal Circuits and Plasticity Unit, National Institute of Neurological Disorders and Stroke, Bethesda, MD, USA. [5] Michael Smith Laboratories, University of British Columbia, Vancouver, BC, Canada. [6] These authors contributed equally: Michael A. Skinnider, Grégoire Courtine. ✉email: michaelskinnider@gmail.com; gregoire.courtine@epfl.ch

The abundance of RNA species informs on the past, present and future state of cells and tissues. By enabling the complete quantification of mRNA populations, RNA sequencing (RNA-seq) has provided unprecedented access to the molecular processes active in a biological sample[1]. Diseases, traumas, and experimental manipulations perturb these processes, which leads to changes in the expression of specific mRNAs. Historically, these altered mRNAs were identified using bulk RNA-seq in non-perturbed versus perturbed tissues[2]. However, biological tissues are composed of multiple cell types, whose responses to a perturbation can differ dramatically. Changes in mRNA abundance within multicellular tissues are confounded by different responses across cell types and changes in the relative abundance of these cell types[3]. Consequently, the resolution of bulk RNA-seq is insufficient to characterize the multifaceted responses to biological perturbations.

Single-cell RNA-seq (scRNA-seq) enables the quantification of RNA abundance at the resolution of individual cells[4]. The maturation of single-cell technologies now enables large-scale comparisons of cell states within complex tissues, thus providing the appropriate resolution to dissect cell-type-specific responses to perturbation[5,6]. The sparsity and heterogeneity of single-cell data initially encouraged the development of specialized statistical methods to identify differentially expressed mRNAs[7,8]. The proliferation of statistical methods for differential expression analysis prompted investigators to ask which methods produced the most biologically accurate results. To answer this question, investigators turned to simulations in an attempt to create a ground truth against which the various methods could be benchmarked. However, simulations require specifying a model from which synthetic patterns of differential expression are generated. Differences in the specification of this model have led investigators to contrasting conclusions[9,10].

These divergences emphasize the importance of developing a sound epistemological foundation for differential expression in single-cell data[11]. In this work, we reasoned that developing such a foundation would require quantifying the performance of the available methods across multiple datasets in which an experimental ground truth is known, and defining the principles that are responsible for differences in performance. We therefore first established a methodological framework that enabled us to curate a resource of ground-truth datasets. Using this resource, we conduct a definitive comparison of the various available methods for differential expression analysis. We find that differences in the performance of these methods reflect the failure of certain methods to account for intrinsic variation between biological replicates. Our understanding of this principle led us to discover that the most frequently used methods can identify differentially expressed genes even in the absence of biological differences. These false discoveries are poised to mislead investigators. However, we show that false discoveries can be avoided using statistical methodologies that account for between-replicate variation. In summary, we expose the principles that underlie valid differential expression analysis in single-cell data, and provide a toolbox to implement relevant statistical methods for single-cell users.

Results

A ground-truth resource to benchmark single-cell differential expression. We aimed to compare available statistical methods for differential expression (DE) analysis based on their ability to generate biologically accurate results. We reasoned that performing this comparison in real datasets where the experimental ground truth is known would faithfully reflect differences in the performance of these methods, while avoiding the shortcomings of simulated data. We posited that the closest possible approximation to this ground truth could be obtained from matched bulk and scRNA-seq performed on the same population of purified cells, exposed to the same perturbations, and sequenced in the same laboratories. An extensive survey of the literature identified a total of eighteen 'gold standard' datasets that met these criteria (Fig. 1a)[12–15]. This compendium allowed us to carry out a large-scale comparison of DE methods in experimental settings where the ground truth is known.

Pseudobulk methods outperform generic and specialized single-cell DE methods. We selected a total of fourteen DE methods, representing the most widely used statistical approaches for single-cell transcriptomics, to compare (Methods, "Differential expression analysis methods"). Together, these methods have been used by almost 90% of recent studies (Fig. 1b). We evaluated the relative performance of each method based on the concordance between DE results in bulk versus scRNA-seq datasets. To quantify this concordance, we calculated the area under the concordance curve (AUCC) between the results of bulk versus scRNA-seq datasets[16,17].

We compared the performance of the fourteen methods across the entire compendium of the eighteen gold standard datasets. This analysis immediately revealed that all six of the top-performing methods shared a common analytical property. These methods aggregated cells within a biological replicate, to form so-called 'pseudobulks', before applying a statistical test (Fig. 1c)[18]. In comparison, methods that compared individual cells performed poorly. The differences between pseudobulk and single-cell methods were highly significant (Fig. 1d), and robust to the methodology used to quantify concordance (Supplementary Fig. 1a-d). Moreover, comparisons to matching proteomics data[13] revealed that pseudobulk methods also more accurately predicted changes in protein abundance (Supplementary Fig. 1e-f).

We asked whether the differences between DE methods could also impact the functional interpretation of transcriptomic experiments. For this purpose, we compared Gene Ontology (GO) term enrichment analyses in bulk versus scRNA-seq DE. We found that pseudobulk methods again more faithfully reflected the ground truth, as captured in the bulk RNA-seq (Fig. 1e and Supplementary Fig. 1g). For example, single-cell methods failed to identify the relevant GO term when comparing mouse phagocytes stimulated with poly(I:C)[12], a synthetic double-stranded RNA (Fig. 1f).

Single-cell DE methods are biased towards highly expressed genes. The unexpected superiority of pseudobulk methods compelled us to study the mechanisms that are responsible for their ability to recapitulate biological ground truth. To investigate these mechanisms, we formulated and tested several hypotheses that could potentially explain these differences in performance.

Previous studies demonstrated that inferences about DE are generally more accurate for highly expressed genes[19,20]. Measurements of gene expression in single cells are inherently sparse. By aggregating cells within each replicate, pseudobulk methods dramatically reduce the number of zeros in the data, especially for lowly expressed genes (Fig. 2a). Consequently, we initially hypothesized that the difference in accuracy between pseudobulk and single-cell methods could be explained by superior performance of pseudobulk methods among lowly expressed genes.

To test this hypothesis, we allocated genes into three equally sized bins, comprising lowly, moderately, and highly expressed genes. We then re-calculated the concordance between bulk and scRNA-seq DE within each bin. Contrary to our prediction, we

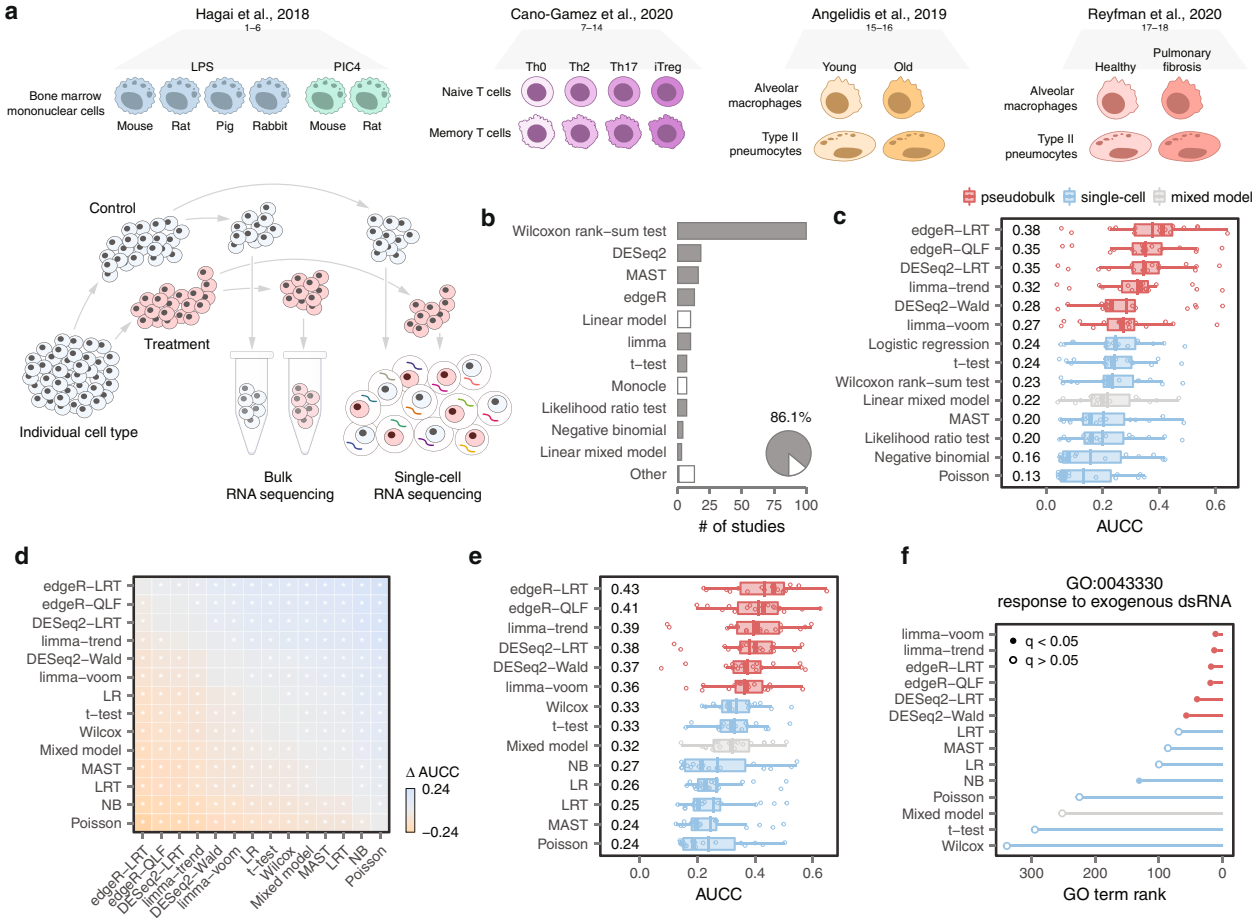

Fig. 1 A systematic benchmark of differential expression in single-cell transcriptomics. a Schematic overview of the eighteen ground-truth datasets analyzed in this study. **b** Statistical methods for DE analysis employed in 500 recent scRNA-seq papers. Grey bars represent DE analysis methods included in this study. "Other" includes methods used in two or fewer studies. Inset pie chart shows the total proportion of recent scRNA-seq papers that employed DE analysis methods included in this study. Source data are provided as a Source Data file. **c** Area under the concordance curve (AUCC) for fourteen DE methods in the eighteen ground-truth datasets shown in **a**. **d** Mean difference in the AUCC (Δ AUCC) 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 of GO term enrichment, as evaluated using gene set enrichment analysis[46], in the eighteen ground-truth datasets shown in **a**. **f** Rank and statistical significance of the GO term GO:0043330 ("response to exogenous dsRNA") in GSEA analyses of mouse bone marrow mononuclear cells stimulated with poly-I:C, a type of synthetic dsRNA, for four h, using the output of fourteen DE methods. Source data are provided as a Source Data file.

observed minimal differences between pseudobulk and single-cell methods for lowly expressed genes (Fig. 2b and Supplementary Fig. 2a). Instead, the most pronounced differences between pseudobulk and single-cell methods emerged among highly expressed genes.

This unexpected result led us to ask whether single-cell DE methods produce systematic errors for highly expressed genes. To explore this possibility, we scrutinized the bulk datasets to identify genes falsely called as DE by each method within scRNA-seq data. We found that false positives identified by single-cell DE methods were more highly expressed than those identified by pseudobulk methods (Fig. 2c and Supplementary Fig. 2b). Conversely, false-negatives overlooked by single-cell DE methods tended to be lowly expressed (Supplementary Fig. 2c-d). Together, these findings implied a systematic tendency for single-cell methods to identify highly expressed genes as DE, even when their expression remained unchanged.

To validate this conclusion experimentally, we analyzed a dataset in which a population of synthetic mRNAs were spiked into each well containing a single cell[12,21]. Each of these single-cell libraries therefore contained equal concentrations of each synthetic mRNA. We found that single-cell methods incorrectly identified many abundant spike-ins as DE (Fig. 2d-e and Supplementary Fig. 2e-f). In contrast, pseudobulk methods avoided this bias.

We then asked whether this bias was universal in single-cell transcriptomics. We assembled a compendium of 46 scRNA-seq datasets that encompassed disparate species, cell types, technologies, and biological perturbations (Supplementary Fig. 3). We found that single-cell DE methods displayed a systematic preference for highly expressed genes across the entire compendium (Fig. 2f).

Together, these experiments suggest that the inferior performance of single-cell methods can be attributed to their bias towards highly expressed genes.

DE analysis of single-cell data must account for biological replicates. These findings implied that pseudobulk methods possess a common analytical property that allows them to avoid this bias. We conducted a series of experiments to identify this mechanism.

The statistical tools applied to identify DE genes in pseudobulk data (i.e., edgeR, DESeq2, and limma) have been refined over

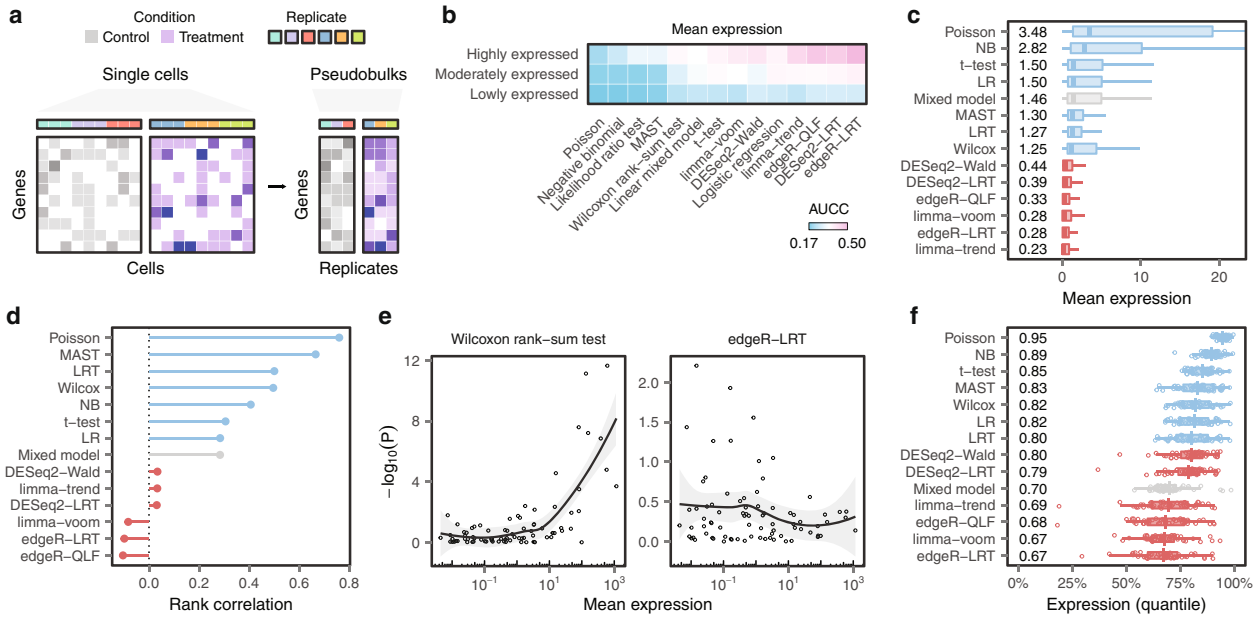

Fig. 2 Single-cell DE methods are biased towards highly expressed genes. a Schematic illustration of the creation of 'pseudobulks' from single-cell data. Top, biological replicate from which each cell was obtained. Bottom, simulated gene expression matrix. Read counts for each gene are aggregated across cells of a given type within each biological replicate. **b** Mean AUCCs across eighteen ground-truth datasets after dividing the transcriptome into terciles of lowly, moderately, or highly expressed genes. **c** Mean expression levels of the 100 top-ranked false-positive genes from each DE method. **d** Spearman correlation between the mean expression of 80 ERCC spike-ins expressed in at least three cells and the $-\log_{10}$ *p*-value of differential expression assigned by each DE method. **e** Scatterplots of mean ERCC expression 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. **f** Mean expression levels of the 200 top-ranked genes from each DE method in a collection of 46 scRNA-seq datasets.

many years of development. We therefore asked whether these methods incorporate inherent advantages that are independent from the procedure of aggregating gene expression across cells. To test this possibility, we disabled the aggregation procedure and applied the pseudobulk methods to individual cells (Fig. 3a). Strikingly, this procedure abolished the superiority of the pseudobulk methods (Fig. 3b and Supplementary Fig. 4a). The emergence of a bias towards highly expressed genes paralleled this decrease in performance (Fig. 3b and Supplementary Fig. 4b-c).

This result raised the possibility that the aggregation procedure itself was directly responsible for the superiority of pseudobulk methods. To evaluate this notion, we applied the aggregation procedure to random groups of cells, which produced a pseudobulk matrix composed of 'pseudo-replicates' (Fig. 3c). This experiment induced a similar decrease in the performance of pseudobulk methods, combined with the re-emergence of a bias towards highly expressed genes (Fig. 3d and Supplementary Fig. 4d–f).

We sought to understand the common factors that could explain the decreased performance of pseudobulk methods in these two experiments. We recognized that both experiments entailed a loss of information about biological replicates. Aggregating random groups of cells to form pseudo-replicates, or ignoring replicates altogether in comparisons of single cells, both introduced a bias towards highly expressed genes and a corresponding loss of performance.

Within the same experimental condition, replicates exhibit inherent differences in gene expression, which reflect both biological and technical factors[22]. We reasoned that failing to account for these differences could lead methods to misattribute the inherent variability between replicates to the effect of the perturbation. To study this potential mechanism, we compared the variance in the expression of each gene in pseudobulks and pseudo-replicates. Initially, we performed this comparison in a dataset of bone marrow mononuclear cells stimulated with poly-I:C[12]. We found that shuffling the replicates produced a systematic decrease in the variance of gene expression, affecting 98.2% of genes (Fig. 3e). We next tested whether this decrease in variance occurred systematically across our compendium of 46 datasets. Every comparison displayed the same decrease in the variance of gene expression (Fig. 3f).

The decrease in the variance of gene expression led statistical tests to attribute small changes in gene expression to the effect of the perturbation. For instance, in the poly-I:C dataset, failing to account for the variable expression of *Txnrd3* across replicates led to the spurious identification of this gene as differentially expressed (Fig. 3g). Moreover, we found that highly expressed genes exhibited the largest decrease in variance in pseudo-replicates, thus explaining the bias of single-cell methods towards highly expressed genes (Supplementary Fig. 4g-k).

Together, this series of experiments exposed the principle underlying the unexpected superiority of pseudobulk methods. Statistical methods for differential expression must account for the intrinsic variability of biological replicates to generate biologically accurate results in single-cell data. Accounting for this variability allows pseudobulk methods to correctly identify changes in gene expression caused by a biological perturbation. In contrast, failing to account for biological replicates causes single-cell methods to systematically underestimate the variance of gene expression. This underestimation of the variance biases single-cell methods towards highly expressed genes, compromising their ability to generate biologically accurate results.

False discoveries in single-cell DE. We realized that if failing to account for the variation between biological replicates could produce false discoveries in the presence of a real biological perturbation, then false discoveries might also arise in the absence

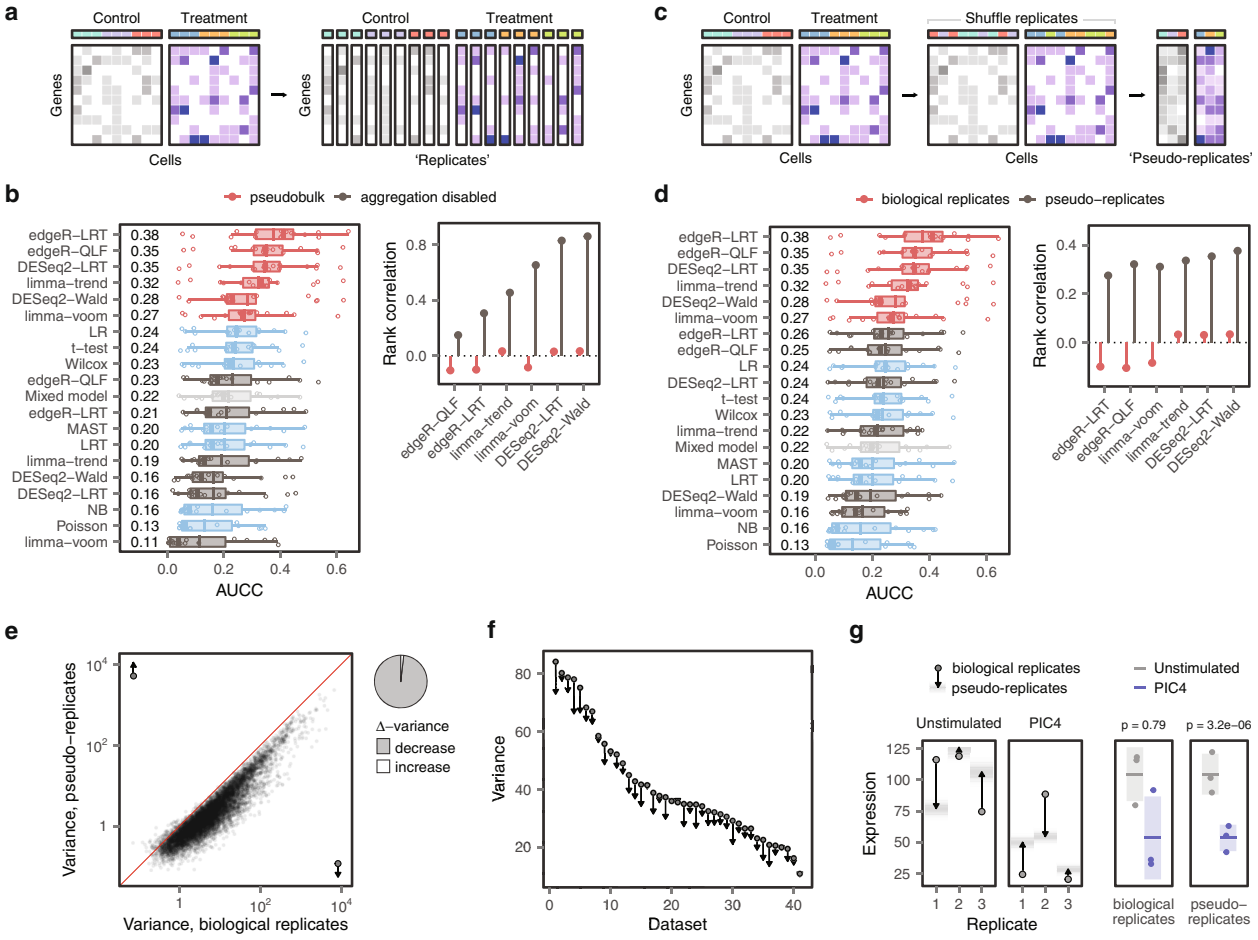

Fig. 3 DE analysis of single-cell data must account for biological replicates. a Schematic illustration of the experiment shown in **b**, in which the aggregation procedure was disabled and pseudobulk DE methods were applied to individual cells. **b** Left, AUCC of the original fourteen DE methods, plus six pseudobulk methods applied to individual cells, in the eighteen ground-truth datasets. Right, Spearman correlation between ERCC mean expression and $-\log_{10}$ p-value assigned by six pseudobulk DE methods, before and after disabling the aggregation procedure. **c** Schematic illustration of the experiment shown in **d**, in which the replicate associated with each cell was shuffled to produce 'pseudo-replicates.' **d** Left, AUCC of the original fourteen DE methods, plus six pseudobulk methods applied to pseudo-replicates, in the eighteen ground-truth datasets. Right, Spearman correlation between ERCC mean expression and $-\log_{10}$ p-value assigned by six pseudobulk DE methods, before and after shuffling replicates to produce pseudo-replicates. **e** Variance of gene expression in pseudobulks formed from biological replicates and pseudo-replicates in mouse bone marrow mononuclear cells stimulated with poly-I:C. Shuffling the replicate associated with each cell produced a systematic decrease in the variance of gene expression. Right, pie chart shows the proportion of genes with increased or decreased variance in pseudo-replicates, as compared to biological replicates. **f** Decreases in the variance of gene expression in pseudo-replicates as compared to biological replicates across 46 scRNA-seq datasets. Points show the mean variance in biological replicates; arrowheads show the mean variance in pseudo-replicates. **g** Left, expression of the gene c in biological replicates (points) and pseudo-replicates (arrowheads) from unstimulated cells and cells stimulated with poly-I:C, with the range of possible pseudo-replicate expression values shown as a density. Right, mean (horizontal line) and variance (shaded area) of Txnrd3 expression in biological replicates (left) and pseudo-replicates (right). P-values were calculated by edgeR-LRT.

of any biological difference. To test this possibility, we simulated single-cell data with different degrees of heterogeneity between replicates in the absence of any difference between groups (Fig. 4a). We randomly assigned each replicate to an artificial 'control' or 'treatment' group, and tested for DE between the two conditions. Strikingly, single-cell methods identified hundreds of DE genes in the absence of any perturbation (Fig. 4b and Supplementary Fig. 4a). Moreover, in line with our understanding of the mechanisms underlying the failure of single-cell DE methods, the genes that were falsely called as DE were those whose expression was most variable between replicates (Fig. 4c and Supplementary Fig. 4b). Pseudobulk methods abolished the false detection of DE genes. However, creating pseudo-replicates led to the reappearance of spurious DE genes (Fig. 4b-c and Supplementary Fig. 4a-b), further corroborating the requirements for accurate DE analyses. The number of false discoveries was reduced when additional replicates were introduced to the dataset (Supplementary Fig. 4c). In contrast, introducing additional cells to the simulated data only exacerbated the underlying problem (Supplementary Fig. 4d).

These findings compelled us to investigate whether similar false discoveries could arise in real single-cell data. To explore this possibility, we initially analyzed a dataset of human peripheral blood mononuclear cells (PBMCs) exposed to interferon[5]. We extracted the control samples that had not been exposed to interferon, and split them randomly into two groups. We then performed DE analysis. Failing to account for the intrinsic variability of biological replicates produced hundreds of DE genes between randomly assigned replicates (Fig. 4d and Supplementary Fig. 6a, b).

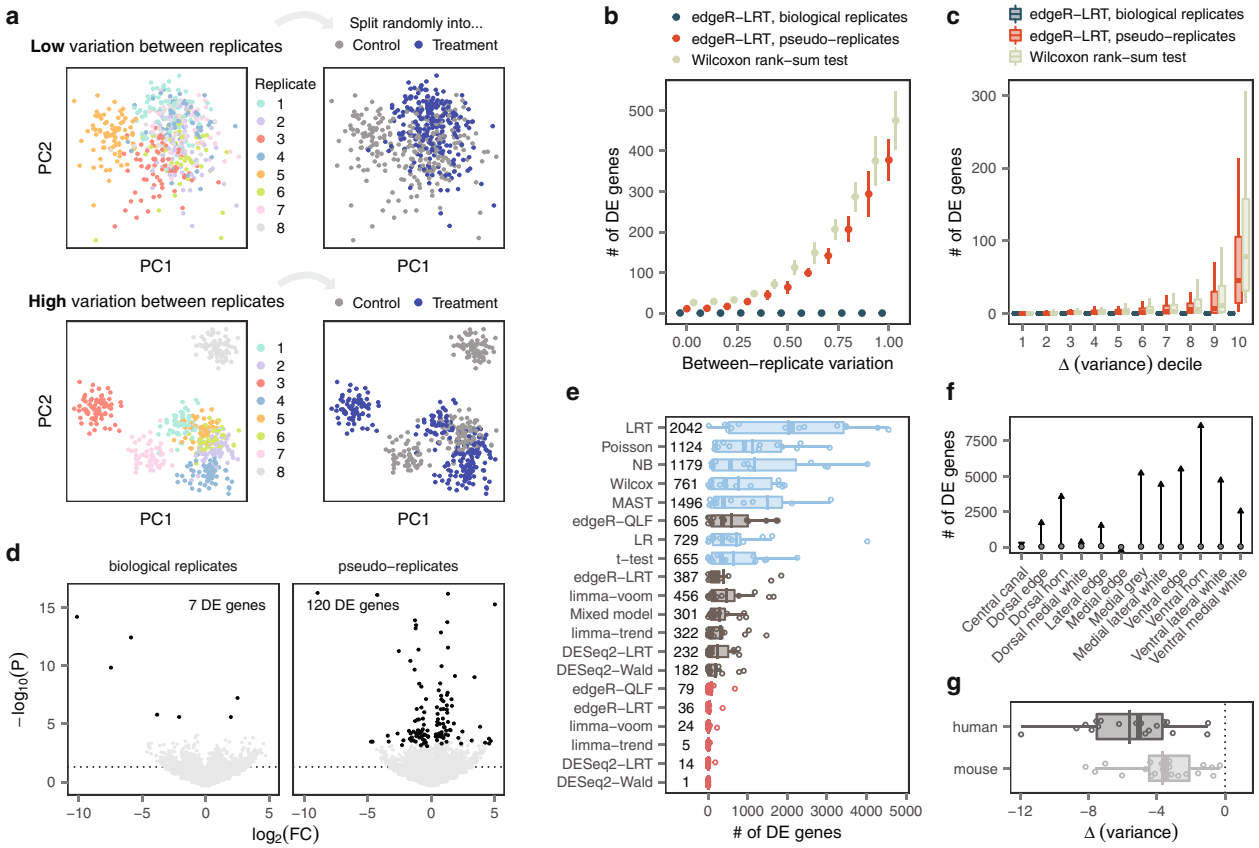

Fig. 4 False discoveries in single-cell DE. a Schematic illustration of simulation experiments. Single-cell RNA-seq datasets were simulated with varying degrees of heterogeneity between replicates. Replicates were then randomly assigned to either a 'treatment' or 'control' group, and DE analysis was performed between groups. **b** Number of DE genes detected in stimulation experiments with varying degrees of heterogeneity between replicates by a representative single-cell DE method, a representative pseudobulk method, and the same pseudobulk method applied to pseudo-replicates. Points and error bars show the mean and standard deviation across ten independent simulations. **c** Number of DE genes detected by the tests shown in **b** for genes divided into deciles by the magnitude of the change in variance between biological replicates and pseudo-replicates (Δ-variance). **d** Volcano plots showing DE between T cells from random groups of unstimulated controls drawn from Kang et al.[5] using a representative pseudobulk method, edgeR-LRT, applied to biological replicates or pseudo-replicates. Discarding information about biological replicates leads to the appearance of false discoveries. **e** Number of DE genes detected in comparisons of random groups of unstimulated controls from 14 scRNA-seq studies with at least six control samples. **f** Number of DE genes detected within spinal cord regions from control mice profiled by spatial transcriptomics[24] using a representative pseudobulk method, edgeR-LRT (points), or a representative single-cell method, the Wilcoxon rank-sum test (arrowheads). **g** Mean change in variance between biological replicates and pseudo-replicates for 18 human and 20 mouse scRNA-seq datasets.

Unsettled by this appearance of false discoveries, we asked whether this observation reflected a universal pitfall. To address this concern comprehensively, we identified a total of fourteen datasets that included at least six replicates in the control condition. As in the previous experiment, we split these unperturbed samples randomly into synthetic control and treatment groups, before conducting DE analyses between these two groups. This systematic analysis confirmed that single-cell methods produced a systematic excess of false positives compared to pseudobulk methods (Fig. 4e). The resulting DE genes were enriched for hundreds of Gene Ontology (GO) terms, despite a complete absence of biological perturbation (Supplementary Fig. 6c). Moreover, we again confirmed that the genes falsely identified as DE corresponded to those with the highest variability between replicates (Supplementary Fig. 6d).

Together, these experiments exposed a fundamental pitfall for DE analysis in single-cell transcriptomics. We intuited, however, that this pitfall could afflict any technology in which many observations are obtained from each biological replicate. For example, we anticipated that false discoveries would also emerge in spatial transcriptomics data[23]. To test this prediction, we analyzed a spatial transcriptomics dataset that profiled spinal cords from a model of amyotrophic lateral sclerosis (ALS)[24]. We randomly partitioned data from control mice into two groups, and performed DE within each region of the spinal cord. Statistical methods that failed to account for variability between biological replicates identified thousands of DE genes within each region (Fig. 4f and Supplementary Fig. 6e). In contrast, pseudobulk methods abolished these false discoveries.

These experiments demonstrated that the variability between biological replicates can confound the identification of genes affected by a biological perturbation. Many of the factors that produce this variability between replicates can be minimized in animal models, including the genetic background, environment, intensity and timing of the biological perturbation, and sample processing. In contrast, these sources of variation are inherently more difficult to control in experiments involving human subjects. This distinction raised the possibility that single-cell studies of human tissue would exhibit greater variability between biological replicates, and consequently, would be more vulnerable to false discoveries. To evaluate this possibility systematically, we calculated the variability between replicates within 41 human and

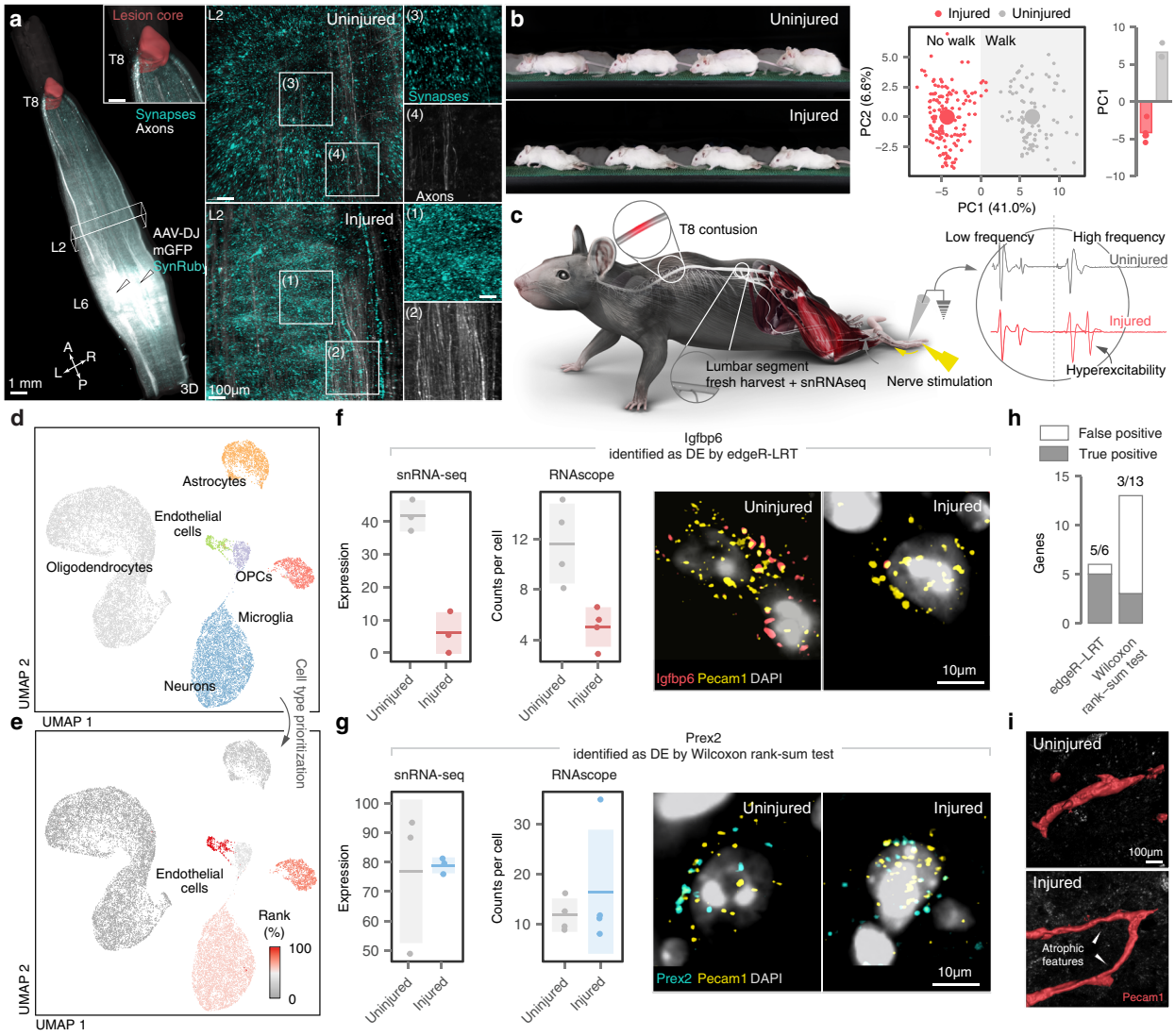

Fig. 5 True and false discoveries in the injured mouse spinal cord. a Synapses and projections of *Vglut2*^{ON} neurons in the mouse lumbar spinal cord before and after SCI. Experiments were repeated in $n = 3$ mice. **b** Top, chronophotography of mice before and after SCI. Bottom left, principal component analysis of gait parameters for each condition (small circles). Large circles show the average per group. Bottom right, bar plot showing the average scores on principal component 1 (PC1), which quantify the locomotor performance of injured and uninjured mice. **c** snRNA-seq experimental design. Inset demonstrates persistent amplitude at increasing frequency of the H-reflex in response to plantar stimulation, reflecting hyperexcitability. **d** Uniform manifold approximation and projection (UMAP) visualization of 19,237 nuclei, revealing the major cell types of the mouse lumbar spinal cord. **e** UMAP visualization colored by the AUC of cell type prioritizations, as calculated by Augur[27]. Endothelial cells are highlighted as the cell type with the highest AUC. **f** Pseudobulk expression levels in lumbar spinal cord endothelial cells by snRNA-seq (in counts per 10,000), left, and mean counts per *Pecam1*^{ON} cell by RNA-scope, center, for *Igfbp6*, a gene identified as DE only by edgeR-LRT. Horizontal line and shaded area show the mean and standard deviation, respectively. Right, colocalization of *Igfbp6* with the endothelial marker gene *Pecam1* in RNAscope in situ hybridizations from injured and uninjured mouse spinal cords. **g** As in **f**, but for *Prex2*, a gene identified as DE only by the Wilcoxon rank-sum test. **h** Summary of in vivo screen results. Filled bars represent genes validated by RNAscope. **i**, Immunohistochemistry reveals *Pecam1*^{ON} cells with atrophic features after SCI. Experiments were repeated in $n = 3$ mice. Source data are provided as a Source Data file.

mouse scRNA-seq datasets. In agreement with our hypothesis, we detected significantly more variability between replicates in the human datasets (Fig. 4g). While we show that accounting for biological replicates is critical for any DE analysis, this result stresses the paramount importance of addressing this issue in single-cell studies of human tissue.

True and false discoveries in the injured mouse spinal cord. We finally sought to demonstrate the extent to which DE analyses can produce true and false discoveries in previously unexplored biological tissues. For this purpose, we characterized the impact of a spinal cord injury (SCI) on gene expression in cells located below the injury. We specifically focused on the lumbar spinal cord, since this region undergoes multifaceted changes that lead to the irreversible degradation of neuronal function[25,26].

We performed experiments in mice that received a severe contusion of the mid-thoracic spinal cord (Fig. 5a-c). Multifactorial quantification of whole-body kinematics revealed profound impairments in the ability of the mice to produce locomotion (Fig. 5b and Supplementary Fig. 8a). We found that the injury triggered the aberrant growth of new synapses throughout lumbar segments, combined with the emergence of abnormal segmental reflexes (Fig. 5a and Supplementary Fig. 8a). This chaotic reorganization of circuits below the SCI has been

linked to spasticity and neuronal dysfunction (Fig. 5b and Supplementary Fig. 8b-c)[25,26].

We then harvested the lumbar spinal cords of mice with chronic SCI and uninjured controls, and performed single-nucleus RNA-seq (snRNA-seq) of these tissues[27]. We sequenced a total of 19,237 cells that encompassed all the major cell types of the lumbar spinal cord (Fig. 5d).

We initially aimed to identify the cell types in which transcription was most perturbed by the injury. To answer this question, we performed cell type prioritization using Augur[27,28]. This unbiased analysis indicated that endothelial cells underwent the most profound transcriptional changes in the spinal cord below the injury (Fig. 5e).

This unexpected finding spurred us to investigate the specific transcriptional changes underlying this prioritization, and the capacity of different statistical methods to reveal these changes. For this purpose, we performed DE analyses between injured and uninjured endothelial cells using representative single-cell and pseudobulk methods. We selected the Wilcoxon rank-sum test as a single-cell method, since this test has been the most widely used approach in the field of single-cell transcriptomics (Fig. 1b), and edgeR-LRT[29] as a pseudobulk method due to its high level of performance (Fig. 1c). These methods identified largely distinct sets of DE genes, with only four genes overlapping between the two methods. Conversely, the Wilcoxon rank-sum test and edgeR-LRT each nominated an additional 44 and 12 genes as DE, respectively (Supplementary Fig. 9a).

Our results thus far have demonstrated that failing to account for variation between replicates can lead single-cell DE methods to produce false discoveries. We therefore suspected that some of the additional genes identified by the Wilcoxon rank-sum test in this dataset could represent false positives. To clarify the ground truth expression of these genes in the injured spinal cord, we carried out a systematic in vivo screen. We obtained RNAscope probes for nineteen putatively DE genes identified by only one of the two methods, and quantified the expression of these genes in endothelial cells from injured and uninjured mice[30] (Supplementary Fig. 9b). RNAscope validated five of the six genes called as DE by edgeR-LRT. In marked contrast, only three of thirteen genes called as DE by the Wilcoxon rank-sum test could be corroborated ($p < 0.05$, χ^2 test; Fig. 5f-h). Several of the validated edgeR-LRT genes, including *Slc7a11* and *Igfbp6*, are involved in the response to hypoxia within endothelial cells, supporting the establishment of a chronically hypoxic state in the lumbar spinal cord[31–33]. In line with the expected consequences of chronic hypoxia, we detected the presence of numerous atrophic blood vessels below the level of injury (Fig. 5i).

Together, these observations illustrate the potential for single-cell DE methods to produce false discoveries. Conversely, valid single-cell DE analysis that accounted for variation between biological replicates yielded reproducible conclusions that could be validated in vivo.

DE analysis with mixed models. Our experiments established that accounting for variation between biological replicates dictated the performance of single-cell DE methods. We were therefore puzzled by the unsatisfying performance of a linear mixed model. By explicitly modeling variation both within and between biological replicates, mixed models should benefit from increased statistical power compared to pseudobulk methods[9]. To clarify this discrepancy, we evaluated eight additional Poisson or negative binomial generalized linear mixed models (GLMMs; Supplementary Fig. 10a-b). In datasets of 25-50 cells, GLMMs could produce accurate results under very specific parameter combinations. However, in datasets comprising 500 or more cells, their performance converged to that of pseudobulk DE methods. Moreover, the computational resources required to fit the best-performing GLMMs were enormous. Even in downsampled datasets, DE analysis of a single cell type took an average of 13.5 h (Supplementary Fig. 10c-d). In contrast, pseudobulk methods required only minutes per cell type in our compendium of 46 datasets (Supplementary Fig. 10e-f). These observations suggest that, in practice, pseudobulk approaches provide an excellent trade-off between speed and accuracy for single-cell DE analysis.

Discussion

Accurate DE analysis in single-cell transcriptomics is required to dissect the transcriptional programs underlying the multifaceted responses to disease, trauma, and experimental manipulations. Despite the importance of statistical methods for DE analyses, the principles that determine their performance have remained elusive. Here, we demonstrate that the central principle underlying valid DE analysis is the ability of statistical methods to account for the intrinsic variability of biological replicates. Accounting for this variability dictates the biological accuracy of statistical methods. Conversely, methods that fail to account for the variability of biological replicates can produce hundreds of false discoveries in the absence of any biological difference.

Investigators study single cells to understand more general principles underlying the response to a biological perturbation. Clarifying these principles requires statistical inferences that generalize beyond the individual cells that constitute any particular dataset. Our results demonstrate that by performing a statistical inference at the level of individual cells, single-cell DE methods conflate variability between biological replicates with the effect of a biological perturbation. The presence of variability between replicates is unavoidable, and can be attributed to both technical factors and intrinsic biological differences[22]. The possibility that conflating variability between replicates with the biological effect of interest can lead to spurious findings has previously been recognized[18,34]. However, these studies relied almost entirely on synthetic data, supplemented by a few illustrative case studies. Consequently, the pervasiveness of false discoveries in published analyses of single-cell data and the propensity for these false discoveries to affect the biological conclusions of a study have remained unclear.

Here, we show that the appearance of false discoveries is a universal phenomenon. Leveraging a collection of 18 single-cell datasets with an experimental ground truth, we demonstrate that the use of inappropriate statistical methodology can produce false discoveries that compromise the biological interpretation of a single-cell experiment. These false discoveries have the potential to squander time, effort, and financial resources in pursuit of misleading hypotheses. For example, we show through a systematic in vivo screen of the injured mouse spinal cord that most DE genes identified by the most commonly used statistical method are false discoveries. Moreover, we elucidate the progression of mechanisms by which failing to account for biological and technical variability makes certain genes disproportionately likely to be spuriously identified as DE. We demonstrate the universality of these mechanisms in multifaceted datasets from an additional 46 single-cell RNA-seq studies. Understanding these mechanisms led us to discover that the same fundamental issues affect other high-dimensional assays, including spatial transcriptomics, and are most likely to manifest in studies of human tissue, suggesting that inference at the level of biological replicates is critical to understand the cellular and molecular basis for human disease.

Our results demonstrate that single-cell DE methods are poised to produce false discoveries. This understanding uncovers an enormous risk for the field. Our findings suggest that many

published findings may be false. Moreover, if left unresolved, substantial research funding may be allocated to follow up on these false discoveries, to the detriment of science. However, this concerning possibility is straightforward to correct with the use of DE methods that account for variability between replicates. Among these, we found that pseudobulk methods achieve the highest fidelity to the experimental ground truth at the levels of the transcriptome, proteome, and functional interpretation. Consequently, we contend that there is an urgent need for a paradigm shift in the statistical methods that are used for DE analysis of single-cell data. The need for such a shift is underscored by our observation that most studies published in the past two years have used inappropriate statistical methods for DE analysis. Moreover, the most widely used analysis packages in the field currently employ DE methods prone to false discoveries by default[35,36]. The increasing prevalence of multi-condition datasets stresses the importance of employing appropriate statistical methodologies to prevent a proliferation of false discoveries. To catalyze this transition, we implement all of the methods tested here in an R package (Supplementary Software 1).

Methods

Literature review. To identify which statistical methods for DE analysis have been most commonly used within the field, we conducted an extensive literature review. We annotated the statistical method used to perform DE analysis across experimental conditions within cell types for each publication included in a large, curated database of scRNA-seq studies[37]. The database was accessed on November 4, 2020. Because the single-cell studies catalogued in this database span a long period of time, and we aimed to establish which methods for DE analysis are currently in wide use, we limited our analysis to the 500 most recently published studies. Accordingly, the inclusion criteria for our review were (i) studies present in the curated database as of November 4, 2020, and (i) studies within the 500 most recent entries in this database at the time it was accessed. Each of these 500 studies were then manually reviewed to determine the statistical methodology used to compare cells of the same type between experimental conditions. We did not annotate methods used to identify genes differentially expressed between cell types (i.e., marker gene identification), as this problem presents a distinct set of statistical challenges[10,38]. In total, 205 of the 500 studies conducted DE analysis between biological conditions. The complete list of all 500 studies is provided as Source Data.

Ground-truth datasets. Previous benchmarks of DE analysis methods for single-cell transcriptomics have relied heavily on simulated data, or else have compared the results of different methods in scenarios where no ground truth was available[10,17]. We reasoned that the best possible approximation to the biological ground truth in a scRNA-seq experiment would consist of a matched bulk RNA-seq dataset in the same purified cell type, exposed to the same perturbation under identical experimental conditions, and sequenced in the same laboratory. We surveyed the literature to identify such matching single-cell and bulk RNA-seq datasets, which led us to compile a resource of eighteen ground truth datasets from four publications[12–15]. Datasets of mouse, rat, pig, and rabbit bone marrow-derived mononuclear phagocytes stimulated with either lipopolysaccharide or poly-I:C for 4 h were obtained from Hagai et al.[12] Datasets of naive or memory T cells stimulated for 5 d with anti-CD3/anti-CD28 coated beads in the presence or absence of various combinations of cytokines (Th0: anti-CD3/anti-CD28 alone; Th2: IL-4, anti-IFNγ; Th17: TGFβ, IL6, IL23, IL1β, anti-IFNγ, anti-IL4; iTreg: TGFβ, IL2) were obtained from Cano-Gamez et al.[13] We additionally obtained label-free quantitative proteomics data for the same comparisons from this study. Datasets of alveolar macrophages and type II pneumocytes from young (3 m) and old (24 m) mice were obtained from Angelidis et al.[14] Datasets of alveolar macrophages and type II pneumocytes from patients with pulmonary fibrosis and control individuals were obtained from Reyfman et al.[15]

Differential expression analysis methods. We compared fourteen statistical methods for DE analysis of single-cell transcriptomics data on their ability to recover ground-truth patterns of DE, as established through bulk RNA-seq analysis of matching cell populations. These fourteen methods comprised seven statistical tests that compared gene expression in individual cells ("single-cell methods"); six tests that aggregated cells within a biological replicate to form pseudobulks before performing statistical analysis ("pseudobulk methods"); and a linear mixed model.

The seven single-cell methods analyzed here included a t-test, a Wilcoxon rank-sum test, logistic regression[39], negative binomial and Poisson generalized linear models, a likelihood ratio test[40], and the two-part hurdle model implemented by MAST[7]. The implementation provided in the Seurat function 'FindMarkers' was used for all seven tests, with all filters ('min.pct', 'min.cells.feature', and 'logfc.threshold') disabled. In addition, we implemented a linear mixed model within Seurat, using the 'lmerTest' R package to optimize the restricted maximum likelihood and obtain p-values from the Satterthwaite approximation for degrees of freedom. We observed that some statistical tests returned a large number of p-values below the double precision limit in R (approximately 2×10^{-308}), potentially confounding the calculation of the concordance metrics described below. To avoid this pitfall, we modified the Seurat implementation to also return the value of the test statistic from which the p-value was derived. The modified version of Seurat 3.1.5 used to perform all single-cell DE analyses reported in this study is available from http://github.com/jordansquair/Seurat.

The pseudobulk methods employed the DESeq2[41], edgeR[29], and limma[42] packages for analysis of aggregated read counts. Briefly, for cells of a given type, we first aggregated reads across biological replicates, transforming a genes-by-cells matrix to a genes-by-replicates matrix using matrix multiplication. For DESeq2, we used both a Wald test of the negative binomial model coefficients (DESeq2-Wald) as well as a likelihood ratio test compared to a reduced model (DESeq2-LRT) to compute the statistical significance. For edgeR, we used both the likelihood ratio test (edgeR-LRT)[43] as well as the quasi-likelihood F-test approach (edgeR-QLF)[44]. For limma, we compared two modes: limma-trend, which incorporates the mean-variance trend into the empirical Bayes procedure at the gene level, and voom (limma-voom), which incorporates the mean-variance trend by assigning a weight to each individual observation[45]. Log-transformed counts per million values computed by edgeR were provided as input to limma-trend.

DE analysis of bulk RNA-seq datasets was performed with six methods (DESeq2-LRT, DESeq2-Wald, edgeR-LRT, edgeR-QLF, limma-trend, and limma-voom), except for the two pulmonary fibrosis datasets[15]; for these datasets, the raw bulk RNA-seq data from sorted cells could not be obtained, so only the results of the bulk DE analysis performed by the authors of the original publication were used. The AUCC and rank correlation were calculated for each bulk DE analysis method separately, and subsequently averaged over all six methods. DE analysis of normalized bulk proteomics data was performed using the moderated t-test implemented within limma, as in the original publication.

Measuring concordance between single-cell and bulk RNA-seq. To evaluate the concordance between DE analyses of matched single-cell and bulk RNA-seq data, we computed two metrics, designed to evaluate the concordance between only the most highly ranked subset of DE genes and across the entire transcriptome, respectively. To calculate the first of these metrics, the area under the concordance curve (AUCC)[16,17], we ranked genes in both the single-cell and bulk datasets in descending order by the statistical significance of their differential expression. Then, we created lists of the top-ranked genes in each dataset of matching size, up to some maximum size k. For each of these lists (that is, for the top-1 genes, top-2 genes, top-3 genes, and so on), we computed the size of the intersection between the single-cell and bulk DE genes. This procedure yielded a curve relating the number of shared genes between datasets to the number of top-ranked genes considered. The area under this curve was computed by summing the size of all intersections, and normalized to the range [0, 1] by dividing it by its maximum possible value, $k \times (k + 1) / 2$. To evaluate the concordance of DE analysis, we used $k = 500$ except where otherwise noted, but found our results were insensitive to the precise value of k. To compute the second metric, the transcriptome-wide rank correlation, we multiplied the absolute value of the test statistic for each gene by the sign of its log-fold change between conditions, and then computed the Spearman correlation over genes between the single-cell and bulk datasets.

In addition to evaluating the consistency of DE analyses at the gene level, we also asked whether each DE method yielded broader patterns of functional enrichment that were similar between the single-cell and bulk datasets, allowing for some divergence in the rankings of individual genes. To address this question, we performed gene set enrichment analysis[46] using the 'fgsea' R package[47]. GO term annotations for human and mouse (2019-12-09 release) were obtained from the Gene Ontology Consortium website. GO terms annotated to less than 10 genes or more than 1,000 genes within each dataset were excluded in order to mitigate the influence of very specific or very broad terms. Genes were ranked in descending order by the absolute value of the test statistic, and 10^6 permutations were performed. To evaluate the concordance of GO term enrichment, we used $k = 100$, on the basis that fewer top-ranked GO terms are generally of interest than are top-ranked genes.

Impact of mean expression. We initially hypothesized that differences between single-cell DE analysis methods could be attributed to their differing sensitivities towards lowly expressed genes. To explore this hypothesis, we performed the following analyses. First, we divided genes from the eighteen gold standard datasets into three equally sized bins on the basis of their mean expression, then re-calculated the AUCC as described above within each bin separately. Second, we inspected the properties of genes falsely called as DE in the single-cell data (false positives) or incorrectly inferred to be unchanging in the single-cell data (false negatives). To identify false positive genes, we used the bulk DE analysis to exclude genes called as DE at a false discovery rate of 10% from the matched single-cell results, then retained the 100 top-ranked remaining genes in the single-cell data. To identify false negative genes, we used the bulk DE analysis to identify genes called as DE at a false discovery rate of 10%, but with a false discovery rate exceeding 10%

in the matched single-cell results, again retaining the 100 top-ranked such genes. For each of these genes, we computed both the mean expression level and the proportion of zero gene expression measurements. Third, we analyzed a Smart-seq2 dataset of human dermal fibroblasts stimulated with interferon-β, in which a mixture of synthetic RNAs was spiked into each individual cell[12]. We performed DE analysis on the synthetic spike-ins, then calculated the Spearman correlation between the mean expression level of each spike-in and the statistical significance of differential expression, as assigned by each single-cell DE method. Fourth, we assembled a compendium of 46 published scRNA-seq datasets, and asked whether the genes called as DE by each method tended to be more or less highly expressed across the entire compendium. Complete details on the preprocessing of these 46 datasets are provided below. Because each of these datasets were sequenced to different depths, and captured different total numbers of genes (depending on both the sequencing depth and the biological system under study), mean expression values were not directly comparable across datasets. To enable such a comparison, we first calculated the mean expression for each gene, then converted this value into the quantile of mean expression using the empirical cumulative distribution function. We then calculated the mean expression quantile of the 200 top-ranked genes from each method in each of the 46 datasets.

Dissecting pseudobulk DE methods. To understand the principles underlying the improved performance of the six pseudobulk DE methods, we performed the following analyses. First, we disabled the aggregation procedure that led to the creation of pseudobulks (that is, we treated each individual cell as its own replicate), then performed an identical DE analysis of individual cells. For each DE method, we then re-calculated both the AUCC and the bias towards highly expressed genes, as quantified by (i) the rank correlation to mean-spike in expression, and (ii) the expression quantile across 46 scRNA-seq datasets. Second, we aggregated random groups of cells into 'pseudo-replicates' by randomizing the replicate associated with each cell. We then again re-calculated both the AUCC and the bias towards highly expressed genes.

These experiments led us to suspect that discarding information about the inherent variability of biological replicates caused both the bias towards highly expressed genes and the attendant decrease in performance. To test this hypothesis, we compared the variance of gene expression in pseudobulks and pseudo-replicates. For each gene, we calculated the difference in variance (Δ-variance) between pseudobulks and pseudo-replicates. We initially visualized the Δ-variance in an exemplary dataset, consisting of mouse bone marrow mononuclear cells stimulated with poly-I:C12. Subsequently, we calculated the mean Δ-variance across all genes in each of the 46 datasets in our scRNA-seq compendium, observing a decrease in the variance in all 46 cases. To clarify the relationship between the Δ-variance and mean gene expression, we computed the correlation between Δ-variance and mean expression, first in the poly-I:C dataset and then across all 46 datasets in the compendium. We observed a significant negative correlation, confirming that the variance of highly expressed genes is disproportionately underestimated when discarding information about biological replicates. We performed a similar analysis correlating the original variance of gene expression to the Δ-variance, demonstrating that the variance of the most variable genes is disproportionately underestimated when discarding information about biological replicates. However, in partial correlation analyses, only gene expression variance remained correlated with Δ-variance, implying that failing to account for biological replicates induces a bias towards highly expressed genes because these genes are also more variably expressed. Supplementary Fig. 4h-i employ the signed pseudo-logarithm transformation from the 'ggallin' R package to visualize the Δ-variance.

Simulation studies. Our understanding of the importance of accounting for variability between biological replicates led us to ask whether failing to account for biological replication could lead to the appearance of false discoveries in the absence of a perturbation. To test this hypothesis, we simulated scRNA-seq data with no biological effect, in which we systematically varied the degree of heterogeneity between replicates. Simulations were performed using the 'Splatter' R package[48], with simulation parameters estimated from the Kang et al. dataset[5] using the 'splatEstimate' function. Populations of between 100 and 2,000 cells were simulated, with between 3 and 20 replicates per condition. DE of varying magnitudes was simulated between replicates by varying the location parameter of the DE factor log-normal distribution ('de.facLoc') between 0 and 1, treating each replicate as its own group, and the total proportion of DE genes ('de.prob') set to 0.5. Then, half of the replicates were randomly assigned to an artificial 'treatment' condition and the remaining half to a 'control' condition, and DE analysis was performed between the treatment and control groups. Except where otherwise noted, plots show results from a simulated population of 500 cells, with three replicates per condition.

Analysis of published scRNA-seq control groups. To confirm that the trends observed in simulation studies were reflective of experimental datasets, we performed a similar analysis using published scRNA-seq data. Within our compendium, we identified a total of fourteen studies with control groups that included six or more samples[5,6,15,49–59]. Details on the preprocessing of each of these datasets are provided below. For each of these studies, we split the control group randomly into artificial 'control' and 'treatment' groups, and performed DE analysis. In addition to computing the total number of DE genes, we identified GO terms enriched among DE genes using a hypergeometric test. We also performed a similar analysis for one spatial transcriptomics dataset24, identifying DE genes between random groups of control mice with barcodes grouped by spinal cord region rather than cell type. Spatial transcriptomics data was downloaded from the supporting website at https://als-st.nygenome.org. Only data from wild-type mice was retained for the analysis. Last, we hypothesized that scRNA-seq studies of human tissues would display more heterogeneity between replicates than studies of animal models, where factors such as genotype, environment, and perturbation can be precisely controlled. To test this hypothesis, we computed the mean Δ-variance across all genes in the 38 human or mouse scRNA-seq datasets in our compendium ($n = 18$ human datasets and 20 mouse datasets).

Application to spinal cord injury. To demonstrate the relevance of our findings to the discovery of new biological mechanisms, we collected scRNA-seq data of the mouse lumbar spinal cord after SCI, and performed DE analysis.

Animal model. Experiments were conducted on adult male or female C57BL/6 mice (15-35 g body weight, 12-30 weeks of age). Vglut2:Cre (Jackson Laboratory 016963) transgenic mice were used and maintained on a mixed genetic background (129/C57BL/6). Housing, surgery, behavioral experiments and euthanasia were performed in compliance with the Swiss Veterinary Law guidelines. Animal care, including manual bladder voiding, was performed twice daily for the first 3 weeks after injury and once daily for the remaining post-injury period. All procedures and surgeries were approved by the Veterinary Office of the Canton of Geneva (Switzerland; GE/57/20 A).

Surgical procedures and post-surgical care. Surgical procedures were performed as previously described[25,60–62]. Briefly, a laminectomy was made at the mid-thoracic level (T9 vertebra). We performed a contusion injury using a force-controlled spinal cord impactor (IH-0400 Impactor, Precision Systems and Instrumentation LLC, USA[63]), as previously described[60,64]. The applied force was set to 90 kdyn. Analgesia (buprenorphine, Essex Chemie AG, Switzerland, 0.01–0.05 mg per kg, s.c.) was provided for three days after surgery.

Kinematic recordings. Kinematic recordings were performed as previously described[25,60,61,65–67]. Bilateral leg kinematics were captured using a 12-camera infrared (200 Hz) Vicon high-speed motion capture system (Vicon Motion Systems, UK). We attached reflective markers bilaterally at the iliac crest, the greater trochanter (hip joint), the lateral condyle (knee joint), the lateral malleolus (ankle), and the distal end of the fifth metatarsophalangeal joint.

Kinematic analysis. For each leg, 15 step cycles were extracted for each mouse. A total of 75 parameters quantifying kinematic and kinetic features were computed for each gait cycle accordingly. To evaluate differences between conditions we implemented a multistep statistical procedure based on principal component analysis, as previously described[25,60,61,65–67].

Electrophysiology. Mice were anaesthetised using a ketamine/xylazine anesthesia mixture. Stainless steel needle electrodes (30 G) were inserted through the posterior surface of the ankle for nerve stimulation and into the lateral, plantar surface of the foot for digital electromyographic recordings. Responses were recorded at a stimulation intensity of 2 x threshold for evoking an H-wave. Signals were amplified and filtered (1000x and 300 Hz–5 kHz, AM Systems differential amplifier) then digitised (PowerLab, AD instruments) for acquisition. Twenty recordings were made at each of 5 different stimulation frequencies (0.1, 0.5, 1, 2, and 5 Hz) with a one minute break between each frequency setting. Peak to peak amplitudes for at least three responses were measured for both M and H waves at each frequency, for each animal. Response amplitudes were first normalized to the amplitude of the M wave at each frequency, and then normalized to the H/M ratio at 0.1 Hz for comparisons across animals.

Single-nucleus RNA sequencing. Single-nucleus dissociation of the mouse spinal cord was performed as previously described[27,51]. Animals were first euthanized by isoflurane inhalation and cervical dislocation. The lumbar spinal cord site was rapidly dissected and frozen on dry ice. Spinal cords were dounced in 500 µl sucrose buffer (0.32 M sucrose, 10 mM HEPES [pH 8.0], 5 mM $CaCl_2$, 3 mM Mg acetate, 0.1 mM EDTA, 1 mM DTT) and 0.1% Triton X-100 with the Kontes Dounce Tissue Grinder. 2 mL of sucrose buffer was added and filtered through a µm cell strainer. The lysate was centrifuged at 3200 g for 10 min at 4 °C. The supernatant was decanted, and 3 mL of sucrose buffer added to the pellet and incubated for 1 min. The pellet was homogenized using an Ultra-Turrax and 12.5 mL of density buffer (1 M sucrose, 10 mM HEPES [pH 8.0], 3 mM Mg acetate, 1 mM DTT) was added below the nuclei layer. The tube was centrifuged at 3200 g at 4 °C and supernatant poured off. Nuclei on the bottom half of the tube wall were collected with 100 µl PBS with 0.04% BSA and 0.2 U/µl RNase inhibitor.

Resuspended nuclei were filtered through a 30 µm strainer, and adjusted to 1000 nuclei/µl.

Library preparation. Library preparation was carried out using the 10x Genomics Chromium Single Cell Kit Version 2. The nuclei suspension was added to the Chromium RT mix to achieve loading numbers of 5,000. For downstream cDNA synthesis (13 PCR cycles), library preparation and sequencing, the manufacturer's instructions were followed.

Read alignment. Reads were aligned to the most recent Ensembl release (GRCm38.93) using Cell Ranger, and a matrix of unique molecular identifier (UMI) counts, including both intronic and exonic reads, was obtained using velocyto[68]. Seurat[35] was then used to calculate quality control metrics for each cell barcode, including the number of genes detected, number of UMIs, and proportion of reads aligned to mitochondrial genes. Low-quality cells were filtered by removing cells expressing less than 200 genes or with more than 5% mitochondrial reads. Genes expressed in less than 3 cells were likewise removed, yielding a count matrix consisting of 22,806 genes and 19,237 cells.

Clustering and integration. Prior to clustering analysis, we first performed batch effect correction and data integration across the two different experimental conditions as previously described[27]. Gene expression data were normalized using regularized negative binomial models[69], then integrated across batches using the data integration workflow within Seurat. The normalized and integrated gene expression matrices were then subjected to clustering to identify cell types in the integrated dataset, again using the default Seurat workflow. Cell types were manually annotated on the basis of marker gene expression, guided by previous studies of the mouse spinal cord[27,51,70].

Viral tract tracing. All surgeries on mice were performed at EPFL under general anaesthesia with isoflurane in oxygen-enriched air using an operating microscope, and rodent stereotaxic apparatus (David Kopf). We identified plasticity of excitatory neurons in the lumbar spinal cord after SCI using AAV-DJ-hSyn Flex mGFP 2 A synaptophysin mRuby (Stanford Vector Core Facility, reference AAV DJ GVVC-AAV-100, titer 1.15E12 genome copies per ml[71]) injections on each side of the cord of Vglut2:Cre mice at the L6 spinal level, 0.25 µl 0.6 mm below the surface at 0.1 µl per minute using glass micropipettes (ground to 50 to 100 µm tips) connected via high-pressure tubing (Kopf) to 10-µl syringes under the control of microinfusion pumps.

Immunohistochemistry. After terminal anaesthesia by barbiturate overdose, mice were perfused transcardially with 4% paraformaldehyde and spinal cords processed for immunofluorescence as previously described[60,72]. Primary antibodies were: rat anti-Pecam1 (BD Biosciences 550274, 1:200). Secondary antibodies were: Alexa Fluor 555 Goat Anti Rat (1:200, Life Technologies, A21434). Immunofluorescence was imaged digitally on a slide scanner [Olympus VS-120 Slide scanner] or confocal microscope [Zeiss LSM880 + Airy fast module with ZEN 2 Black software (Zeiss, Oberkochen, Germany)]. Images were processed using ImageJ (NIH) or Imaris (Bitplane, version 9.0.0).

Tissue clearing. Mouse spinal cords were cleared using CLARITY[73] four weeks after injection of AAV-DJ-hSyn-flex-mGFP-2A-Synaptophysin-mRuby[71]. Mice were perfused transcardially first with 0.1 M PBS followed by 4% PFA (in 0.1 M PBS, pH 7.4) at 4 °C. The spinal cords were dissected and post-fixed in 4% PFA (in 0.1 M PBS) for 24 h at 4 °C. The dura was removed from the samples prior to clearing. Samples were incubated in A4P0 hydrogel solution (4% acrylamide in 0.001 M PBS with 0.25% of the photoinitiator 2,2′-azobis[2-(2-imidazolin-2-yl)propane] dihydrochloride (Wako Pure Chemical, Osaka, Japan)) for 24 h at 4 °C with gentle nutation. Samples were degassed by bubbling nitrogen gas through the tube for 3 m. Hydrogel polymerization was initiated by incubating the sample in a 37 °C water bath for 2 h. Tissue was washed in 0.001 M PBS for 5 m at room temperature. Samples were then placed in the X-CLARITY Tissue Clearing System I (Logos Biosystems Inc., South Korea) set to 1.2 A, 100 RPM, 37 °C. Clearing solution was made in-house from 40 g of sodium dodecyl sulfate (SDS), 200 mM boric acid, and filled to a total volume of 1 L with dH2O (pH adjusted to 8.5). Samples cleared after ~10–15 h. The sample was then washed for at least 24 h at room temperature with shaking in 1x PBS and 0.1% Triton-X (with 0.02% sodium azide) to remove excess SDS. The sample was then incubated in RIMS (40 g of Histodenz dissolved in 30 mL of 0.02 M PB, pH 7.5, 0.01% sodium azide, refractive index 1.465) for at least 24 h at room temperature with gentle shaking prior to imaging. Imaging was performed using a custom-built MesoSPIM[74] and CLARITY-optimized light-sheet microscope (COLM) as described[73]. A customized sample holder was used to secure the spinal cords in a chamber filled with RIMS. Samples were imaged using a 2.5× objective at the MesoSPIM and a 4x objective at the COLM with two lightsheets illuminating the sample from the left and the right sides. The pixel resolution was 2.6 µm × 2.6 µm × 3 µm for the 2,5x acquisition; and 1.4 µm by 1.4 µm by 5 µm for the 4x acquisition in the x-, y-, and z-directions. Images were acquired as 16-bit TIFF files. 3D reconstructions of the raw images were produced using Arivis (v3.4) and Imaris softwares (Bitplane, v.9.0.0).

RNAscope. To corroborate the results suggested by DE analysis of scRNA-seq data, we analyzed the in situ co-localization of putatively DE genes and cell type marker genes using RNAscope (Advanced Cell Diagnostics)[30]. Lists of putatively DE genes were obtained for representative single-cell and pseudobulk DE methods (the Wilcoxon rank-sum test and edgeR-LRT, respectively), and cross-referenced against a list of validated probes designed and provided by Advanced Cell Diagnostics, Inc. In total, probes were obtained for 13 genes identified as DE by the Wilcoxon rank-sum test (Sgms1, catalog no. 538561; Pcdh9, catalog no. 524921; Epas1, catalog no. 314371; Tcaf1, catalog no. 466921; Gspt1, catalog no. 530471; Prex2, catalog no. 432481; Sema7a, catalog no. 437261; Zfp366, catalog no. 443301; Cpe, catalog no. 454091; Afap1l2, catalog no. 556251; Nedd4l, catalog no. 491981; Adipor2, catalog no. 452861; Ptpn14, catalog no. 493181) and 7 genes identified by edgeR-LRT (Slc7a11, catalog no. 422511; Gjb2, catalog no. 518881; Pi16, catalog no. 451311; Rbp4, catalog no. 508501; Col1a1, catalog no. 319371; Igfbp6, catalog no. 425721). In addition, we obtained probes for Pecam1 (catalog no. 316721), a classic endothelial cell marker gene. We then obtained 16 µm cryosections from fixed-frozen spinal cords as previously described[27] and performed staining for each probe according to the manufacturer's instructions, using the RNAscope Fluorescent Multiplex Reagent Kit (cat. no. 323133). For each biological replicate ($n = 4$ per condition for both injured and uninjured mice), we analyzed ten cells positive for *Pecam1* and tallied the number of speckles for each gene of interest. The mean expression of each gene was then calculated for each biological replicate, and a one-tailed t-test was conducted based on the directional change in the snRNA-seq data.

Mixed models. Having established that the performance of DE methods is contingent on their ability to account for biological replicates, we asked why mixed models failed to match the performance of pseudobulk methods. In addition to the linear mixed model described above, we implemented generalized linear mixed models (GLMMs) based on the negative binomial or Poisson distributions, adapting implementations provided in the 'muscat' R package[10]. For each of these models, we evaluated the impact of incorporating the library size factors as an offset term, and compared the Wald test of model coefficients to a likelihood ratio test against a reduced model, yielding a total of four GLMMs from each distribution. The enormous computational requirements of the GLMMs prevented us from evaluating these models in the full ground truth datasets; instead, we analyzed a series of downsampled datasets, each containing between 25 and 1,000 cells. To quantify the computational resources required by each DE method, we monitored peak memory usage using the 'peakRAM' R package, and the base R function 'system.time' to record wall time.

Preprocessing and analysis of published single-cell datasets. We assembled a compendium of 46 published single-cell or single-nucleus RNA-seq studies (Supplementary Fig. 3), and performed DE analyses across this compendium to establish the generality of our conclusions. For publications containing more than one comparison, only a single comparison was retained, as described in detail below. We retained the comparison involving the greatest number of cells, and used the most fine-grained cell type annotations provided by the authors of the original studies. When count matrices did not use gene symbols, the provided identifiers were mapped to gene symbols, and counts summed across genes mapping to identical symbols. Only cell types with at least three cells in each condition were subjected to DE analysis, and genes detected in less than three cells were removed.

Angelidis et al., 2019[14]. scRNA-seq data from young and aged mouse lung (3 m and 24 m, respectively), as well as matching bulk data from two purified cell types, was obtained from GEO (GSE124872). Metadata was obtained from GitHub (https://github.com/theislab/2018_Angelidis). Cells with missing cell type annotations were removed from the single-cell data. DE analysis was performed by comparing cells from young and old mice.

Arneson et al., 2018[75]. scRNA-seq data from the hippocampus of mice after a mild traumatic brain injury (mTBI), delivered using a mild fluid percussion injury model, and matched controls was obtained from GEO (GSE101901). Metadata, including cell type annotations, were provided by the authors. DE analysis was performed by comparing cells from mTBI and control mice.

Avey et al., 2018[76]. scRNA-seq data from the nucleus accumbens of mice treated with morphine for 4 h and saline-treated controls was obtained from GEO (GSE118918). Cells identified as doublets and non-unique barcodes were removed. Metadata, including cell type annotations, were provided by the authors. DE analysis was performed by comparing cells from morphine- and saline-treated mice.

Aztekin et al., 2019[77]. scRNA-seq data from regeneration-competent (NF stage 40-41) Xenopus laevis tadpoles was obtained from ArrayExpress (E-MTAB-7716). DE analysis was performed by comparing cells from tadpoles at 1 d post-amputation to control tadpoles.

Bhattacherjee et al., 2019[78]. scRNA-seq data from the prefrontal cortex of mice exposed to a cocaine withdrawal paradigm was obtained from GEO (GSE124952). DE analysis was performed by comparing cells at the 15 d post-withdrawal timepoint from cocaine- or saline-treated mice.

Brenner et al., 2020[79]. snRNA-seq data from the prefrontal cortex of alcoholic and control individuals was obtained from GEO (GSE141552). Metadata, including cell type annotations, were provided by the authors. DE analysis was performed by comparing nuclei from alcoholic and control individuals.

Cano-Gamez et al., 2020[13]. scRNA-seq data from naive and memory T cells, stimulated with anti-CD3/anti-CD28 coated beads in the presence or absence of various combinations of cytokines, was obtained from the supporting website (https://www.opentargets.org/projects/effectorness). Matching bulk RNA-seq and proteomics data was obtained from the same source. For the analyses presented as part of the compendium of 46 datasets, DE analysis was performed by comparing iTreg and control cells.

Cheng et al., 2019[80]. scRNA-seq data from intestinal crypt cells in wild-type and Hmgcs2 knockout mice was obtained directly from the authors of the original publication. DE analysis was performed by comparing wild type and KO mice.

Co et al., 2020[81]. scRNA-seq data of sorted cells from Drd1a-tdTomato+ control and Foxp2 KO mice was obtained from GEO (GSE130653). Cell type annotations were provided by the authors. Cell types annotated as 'Low quality' were removed prior to further analysis. DE analysis was performed by comparing WT and Foxp2 KO mice.

Crowell et al., 2020[10]. snRNA-seq data from the prefrontal cortex of mice peripherally stimulated with lipopolysaccharide (LPS) and control mice was obtained from the Bioconductor package 'muscData', using the 'Crowell19_4vs4' function. DE analysis was performed by comparing nuclei from LPS-treated and control mice.

Davie et al., 2018[82]. scRNA-seq data from the brains of flies of varying ages, sexes, and genotypes was obtained from the supporting website (http://scope.aertslab.org, file 'Aerts_Fly_AdultBrain_Filtered_57k.loom'). Cells marked as 'Unannotated' were removed. DE analysis was performed by comparing cells from DGRP-551 and W^{1118} flies.

Denisenko et al., 2020[83]. scRNA-seq data from human kidneys subjected to varying dissociation methods and cell fixation techniques was obtained from GEO (GSE141115). Metadata, including cell type annotations, was obtained from the supporting information files accompanying the published manuscript. DE analysis was performed by comparing cells fixed with methanol and freshly dissociated cells, both at −20 °C.

Der et al., 2019[84]. scRNA-seq data of skin samples from patients with lupus nephritis (LN) and healthy controls was obtained from ImmGen (SDY997, EXP15077). Cell type annotations were obtained from the authors of the original manuscript. Other metadata, including biological replicate and experimental condition annotations for each individual cell, was obtained from the supporting information files accompanying the published manuscript. DE analysis was performed by comparing cells from patients with LN and healthy controls.

Goldfarbmuren et al., 2020[56]. scRNA-seq data of tracheal epithelial cells from smokers and non-smokers was obtained from GEO (GSE134174). Patients designated as 'excluded' were removed prior to downstream analysis. DE analysis was performed by comparing cells from smokers and non-smokers.

Grubman et al., 2019[52]. snRNA-seq data from the entorhinal cortex of patients with Alzheimer's disease and matched controls was obtained from the supporting website (http://adsn.ddnetbio.com). Nuclei annotated as 'undetermined' or 'doublet' were removed. DE analysis was performed by comparing nuclei from patients with Alzheimer's disease and controls.

Gunner et al., 2019[85]. scRNA-seq data from the mouse barrel cortex before or after whisker lesioning was obtained from GEO (GSE129150). Cell types not included in Supplementary Fig. 10 of the original publication were removed. DE analysis was performed by comparing cells from lesioned and control mice.

Haber et al., 2017[86]. scRNA-seq data from epithelial cells of the mouse small intestine in healthy mice and after ten days of infection with the parasitic helminth *Heligmosomoides polygyrus* was obtained from GEO (GSE92332), using the Drop-seq data collected by the original publication. DE analysis was performed by comparing cells from infected and uninfected mice.

Hagai et al., 2018[12]. scRNA-seq data of bone marrow-derived mononuclear phagocytes from four different species (mouse, rat, pig, and rabbit) exposed to lipopolysaccharide (LPS) or poly-I:C for two, four, or six h was obtained from ArrayExpress (E-MTAB-6754). Matching bulk RNA-seq data was also obtained from ArrayExpress (E-MTAB-6773). Finally, scRNA-seq data from human dermal fibroblasts stimulated with interferon-β for two or six h, in which the ERCC mixture of synthetic mRNAs was spiked in alongside every cell, was obtained from ArrayExpress (E-MTAB-7051). Counts were summed across technical replicates of the same biological samples. For the analyses presented as part of the compendium of 46 datasets, DE analysis was performed by comparing rabbit cells stimulated with LPS for 2 h and control cells. DE analysis of the spike-in dataset was performed by comparing cells stimulated for 2 h and 6 h.

Hashimoto et al., 2019[87]. scRNA-seq data of peripheral blood mononuclear cells from human supercentenarians and younger controls was obtained from the supporting website (http://gerg.gsc.riken.jp/SC2018). Metadata, including cell type annotations, were provided by the authors. DE analysis was performed by comparing cells from supercentenarians and younger controls.

Hrvatin et al., 2018[50]. scRNA-seq data from the visual cortex of mice housed in darkness, then exposed to light for 0 h, 1 h, or 4 h was obtained from GEO (GSE102827). Cell types labeled as 'NA' were removed from downstream analyses. DE analysis was performed by comparing cells from mice stimulated with light for 4 h to control mice.

Hu et al., 2017[88]. snRNA-seq data from the cerebral cortex of mice after pentylenetetrazole (PTZ)-induced seizure and saline-treated controls was obtained from the Google Drive folder accompanying the original publication (https://github.com/wulabupenn/Hu_MolCell_2017). DE analysis was performed by comparing cells from PTZ- and saline-treated mice.

Huang et al., 2019[58]. scRNA-seq data from the colon of pediatric patients with colitis and inflammatory bowel disease and matched controls was obtained from the supporting website (https://zhanglaboratory.com/research-data/). DE analysis was performed by comparing cells from patients with colitis and healthy controls.

Jaitin et al., 2019[89]. scRNA-seq data from white adipose tissue of mice fed either a high-fat diet or normal chow for six weeks were obtained from the Bitbucket repository accompanying the original publication (https://bitbucket.org/account/user/amitlab/projects/ATIC). Metadata, including cell type annotations, were provided by the authors. DE analysis was performed by comparing cells from high-fat diet and normal chow-fed mice.

Jakel et al., 2019[90]. snRNA-seq data of oligodendrocytes from patients with multiple sclerosis and matched controls was obtained from GEO (GSE118257). DE analysis was performed by comparing nuclei from individuals with multiple sclerosis versus matched controls.

Kang et al., 2018[5]. scRNA-seq data from peripheral blood mononuclear cells (PBMCs) stimulated with recombinant IFN-β for 6 h and unstimulated PBMCs was obtained from GEO (GSE96583). Doublets and unclassified cells were removed. DE analysis was performed by comparing IFN-stimulated and unstimulated cells.

Kim et al., 2019[91]. scRNA-seq data from the ventromedial hypothalamus of mice exposed to a range of behavioral stimuli and control mice was obtained from the Mendeley repository accompanying the original publication. Cell type annotations were provided directly by the authors. DE analysis was performed by comparing cells from animals engaging in aggressive behaviour to the common population of control animals.

Kotliarov et al., 2020[92]. scRNA-seq data of peripheral blood mononuclear cells from subjects who were subsequently given an influenza vaccination were obtained from Figshare (https://doi.org/10.35092/yhjc.c.4753772). DE analysis was performed by comparing cells from high and low responders to the influenza vaccination, as categorized by the authors.

Madissoon et al., 2020[93]. scRNA-seq data from esophagus, lung, and spleen samples after varying durations of cold storage was obtained from the study website (https://cellgeni.cog.sanger.ac.uk/tissue-stability/). DE analysis was performed by comparing cells from samples preserved for 12 h and fresh samples.

Mathys et al., 2019[6]. snRNA-seq data from the prefrontal cortex of patients with Alzheimer's disease and matched controls was obtained from Synapse (syn18681734). Patient data and additional metadata were also obtained from Synapse (syn3191087 and syn18642926, respectively). DE analysis was performed by comparing nuclei from patients with Alzheimer's disease and controls.

Nagy et al., 2020[57]. snRNA-seq data from the dorsolateral prefrontal cortex of patients with major depressive disorder (MDD) and matched controls was obtained from GEO (GSE144136). DE analysis was performed by comparing nuclei from patients with MDD and controls.

Nault et al., 2021[94]. snRNA-seq data from the livers of mice gavaged with 2,3,7,8-tetrachlorodibenzo-p-dioxin or sesame oil vehicle was obtained from GEO (GSE148339). DE analysis was performed by comparing nuclei from treated and vehicle livers.

Ordovas-Montanes et al., 2018[95]. scRNA-seq data from ethmoid sinus cells of patients with chronic rhinosinusitis (CRS), with and without nasal polyps, from Supplementary Table 2 of the original publication. DE analysis was performed by comparing cells from patients with polyposis and non-polyposis CRS.

Reyes et al., 2020[96]. scRNA-seq data of peripheral blood mononuclear cells from patients with sepsis and healthy controls was obtained from the Broad Institute's Single Cell Portal (SCP548). DE analysis was performed by comparing cells from individuals with bacterial sepsis and controls.

Reyfman et al., 2019[15]. scRNA-seq data from the lungs of patients with pulmonary fibrosis and healthy controls was obtained from GEO (GSE122960). Metadata, including cell type annotations, was provided by the authors. One sample ("Cryobiopsy_01") was removed as it was sequenced separately from the rest of the experiment. The results of DE analysis of bulk RNA-seq data, comparing purified AT2 cells or alveolar macrophages from patients with pulmonary fibrosis and healthy controls, were obtained from the supporting information accompanying the original publication. DE analysis was performed by comparing cells from patients with pulmonary fibrosis and controls.

Rossi et al., 2019[53]. scRNA-seq data from the hypothalamus of mice fed either a high-fat diet or normal chow for between 9-16 weeks was obtained directly from the authors, in the form of a processed Seurat object. Cells annotated as 'unclassified' were removed. DE analysis was performed by comparing cells from high-fat diet and normal chow-fed mice.

Sathyamurthy et al., 2018[51]. snRNA-seq data from the spinal cord parenchyma of adult mice exposed to formalin or matched controls was obtained from GEO (GSE103892). Cell types with blank annotations, or annotated as 'discarded', were

removed. DE analysis was performed by comparing cells from mice exposed to formalin and control animals.

Schafflick et al., 2020[97]. scRNA-seq data of peripheral blood mononuclear cells from individuals with multiple sclerosis and matched controls was obtained from GEO (GSE138266). Metadata, including cell type annotations, was obtained from Github (https://github.com/chenlingantelope/MSscRNAseq2019). DE analysis was performed by comparing cells from individuals with multiple sclerosis and controls.

Schirmer et al., 2019[98]. snRNA-seq data from cortical and subcortical areas from the brains of patients with multiple sclerosis and control tissue from unaffected individuals was obtained from the web browser accompanying the original publication (https://cells.ucsc.edu/ms). DE analysis was performed by comparing cells from multiple sclerosis and control patients.

Skinnider et al., 2021[27]. snRNA-seq data from the spinal cords of mice with a spinal cord injury, some of which were exposed to epidural electrical stimulation to restore locomotion after paralysis, was obtained from GEO (GSE142245). DE analysis was performed by comparing nuclei from paralyzed and walking mice.

Tran et al., 2019[55]. scRNA-seq data from the retinal ganglion of mice at various timepoints after an optic nerve crush injury, as well as uninjured controls, was obtained from GEO "GSE137398 ". Metadata, including cell type annotations, was obtained from the Broad Institute's Single-Cell Portal (SCP509). DE analysis was performed by comparing cells from mice at 12 h post-injury and uninjured mice.

Wagner et al., 2018[99]. scRNA-seq data from zebrafish embryos between 14-16 h post-fertilization, with either the chordin locus or a control locus (tyrosinase) disrupted by CRISPR-Cas9 knock- out, was obtained from GEO (GSE112294). DE analysis was performed by comparing cells from chordin- or tyrosinase-targeted embryos.

Wang et al., 2020[100]. scRNA-seq data from the ovaries of young and old cynomolgus monkeys was obtained from GEO (GSE130664). Metadata, including cell type annotations, was obtained from the supporting information accompanying the original publication. Spike-ins were removed. DE analysis was performed by comparing cells from young and old primates.

Wilk et al., 2020[59]. scRNA-seq data of peripheral blood mononuclear cells from patients with COVID-19 and healthy controls was obtained from the supporting website (https://www.covid19cellatlas.org/). DE analysis was performed by comparing patients with COVID-19 and controls.

Wirka et al., 2019[101]. scRNA-seq data from the aortic root of mice fed a high-fat diet or normal chow for eight weeks from GEO (GSE131776). Metadata, including cell type annotations, was provided by the authors, and unannotated cells were removed. DE analysis was performed by comparing cells from high-fat diet and normal chow-fed mice.

Wu et al., 2017[49]. scRNA-seq data from the amygdala of mice subjected to 45 min of immobilization stress and control mice was obtained from GEO (GSE103976). DE analysis was performed by comparing cells from stressed and control mice.

Ximerakis et al., 2019[102]. scRNA-seq data from whole brains of young (2–3 mo) and old (21–23 mo) mice from the Broad Institute's Single Cell Portal (SCP263). DE analysis was performed by comparing cells from young and old mice.

Visualization. Throughout the manuscript, box plots show the median (horizontal line), interquartile range (hinges) and smallest and largest values no more than 1.5 times the interquartile range (whiskers), and the error bars show the standard deviation.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw sequencing data and count matrices have been deposited to the Gene Expression Omnibus under accession code GSE165003. The 18 'ground truth' datasets, including single-cell RNA-seq, bulk RNA-seq and proteomics data, are available from Zenodo at https://doi.org/10.5281/zenodo.5048449. All other relevant data supporting the key findings of this study are available within the article and its Supplementary Information files or from the corresponding author upon reasonable request. The complete list of all 500 studies is provided as Source Data. Source data are provided with this paper.

Code availability

We provide an R package, Libra, implementing all methods for DE analysis discussed in this study within a consistent interface. The Libra package is available from GitHub (https://github.com/neurorestore/Libra) and as Supplementary Software 1. In addition, the R source code used to perform data analysis is available from GitHub at https://github.com/neurorestore/DE-analysis.

Received: 18 May 2021; Accepted: 6 September 2021;
Published online: 28 September 2021

References

1. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
2. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat. Rev. Genet.* **20**, 631–656 (2019).
3. Srinivasan, K. et al. Untangling the brain's neuroinflammatory and neurodegenerative transcriptional responses. *Nat. Commun.* **7**, 11295 (2016).
4. Chen, X., Teichmann, S. A. & Meyer, K. B. From tissues to cell types and back: single-cell gene expression analysis of tissue architecture. *Annu. Rev. Biomed. Data Sci.* **1**, 29–51 (2018).
5. Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
6. Mathys, H. et al. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).
7. Finak, G. et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
8. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).
9. Zimmerman, K. D., Espeland, M. A. & Langefeld, C. D. Pseudoreplication bias in single-cell studies; a practical solution. *BioRxiv* (2020) https://doi.org/10.1101/2020.01.15.906248.
10. Crowell, H. L. et al. muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nat. Commun.* **11**, 6077 (2020).
11. Mehta, T., Tanik, M. & Allison, D. B. Towards sound epistemological foundations of statistical methods for high-dimensional biology. *Nat. Genet.* **36**, 943–947 (2004).
12. Hagai, T. et al. Gene expression variability across cells and species shapes innate immunity. *Nature* **563**, 197–202 (2018).
13. 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).
14. Angelidis, I. et al. An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nat. Commun.* **10**, 963 (2019).
15. Reyfman, P. A. et al. Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* **199**, 1517–1536 (2019).
16. Irizarry, R. A. et al. Multiple-laboratory comparison of microarray platforms. *Nat. Methods* **2**, 345–350 (2005).
17. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15**, 255–261 (2018).
18. Lun, A. T. L. & Marioni, J. C. Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data. *Biostatistics* **18**, 451–464 (2017).
19. Rapaport, F. et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.* **14**, R95 (2013).
20. Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. & Conesa, A. Differential expression in RNA-seq: a matter of depth. *Genome Res.* **21**, 2213–2223 (2011).
21. Jiang, L. et al. Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21**, 1543–1551 (2011).
22. Tung, P.-Y. et al. Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* **7**, 39921 (2017).
23. Ståhl, P. L. et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).
24. Maniatis, S. et al. Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science* **364**, 89–93 (2019).
25. van den Brand, R. et al. Restoring voluntary control of locomotion after paralyzing spinal cord injury. *Science* **336**, 1182–1185 (2012).
26. Beauparlant, J. et al. Undirected compensatory plasticity contributes to neuronal dysfunction after severe spinal cord injury. *Brain* **136**, 3347–3361 (2013).
27. Skinnider, M. A. et al. Cell type prioritization in single-cell data. *Nat. Biotechnol.* **39**, 30–34 (2021).
28. Squair, J. W., Skinnider, M. A., Gautier, M., Foster, L. J. & Courtine, G. Prioritization of cell types responsive to biological perturbations in single-cell data with Augur. *Nat. Protoc.* **16**, 3836–3873 (2021).
29. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
30. Wang, F. et al. RNAscope: a novel in situ RNA analysis platform for formalin-fixed, paraffin-embedded tissues. *J. Mol. Diagn.* **14**, 22–29 (2012).
31. Samanta, D. & Semenza, G. L. Maintenance of redox homeostasis by hypoxia-inducible factors. *Redox Biol.* **13**, 331–335 (2017).
32. Zhang, C. et al. IGF binding protein-6 expression in vascular endothelial cells is induced by hypoxia and plays a negative role in tumor angiogenesis. *Int. J. Cancer* **130**, 2003–2012 (2012).

33. Li, Y. et al. Pericytes impair capillary blood flow and motor function after chronic spinal cord injury. *Nat. Med.* **23**, 733–741 (2017).
34. Zimmerman, K. D., Espeland, M. A. & Langefeld, C. D. A practical solution to pseudoreplication bias in single-cell studies. *Nat. Commun.* **12**, 738 (2021).
35. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
36. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
37. Svensson, V., da Veiga Beltrame, E. & Pachter, L. A curated database reveals trends in single-cell transcriptomics. *Database (Oxford)* **2020**, (2020).
38. Zhang, J. M., Kamath, G. M. & Tse, D. N. Valid post-clustering differential analysis for single-cell RNA-Seq. *Cell Syst.* **9**, 383–392.e6 (2019).
39. Ntranos, V., Yi, L., Melsted, P. & Pachter, L. A discriminative learning approach to differential expression analysis for single-cell RNA-seq. *Nat. Methods* **16**, 163–166 (2019).
40. McDavid, A. et al. Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics* **29**, 461–467 (2013).
41. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
42. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
43. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).
44. Lun, A. T. L., Chen, Y. & Smyth, G. K. It's DE-licious: a recipe for differential expression analyses of RNA-seq experiments using quasi-likelihood methods in edgeR. *Methods Mol. Biol.* **1418**, 391–416 (2016).
45. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
46. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
47. Sergushichev, A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *BioRxiv* (2016) https://doi.org/10.1101/060012.
48. Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* **18**, 174 (2017).
49. Wu, Y. E., Pan, L., Zuo, Y., Li, X. & Hong, W. Detecting activated cell populations using single-cell RNA-Seq. *Neuron* **96**, 313–329.e6 (2017).
50. Hrvatin, S. et al. Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nat. Neurosci.* **21**, 120–129 (2018).
51. Sathyamurthy, A. et al. Massively parallel single nucleus transcriptional profiling defines spinal cord neurons and their activity during behavior. *Cell Rep.* **22**, 2216–2225 (2018).
52. Grubman, A. et al. A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. *Nat. Neurosci.* **22**, 2087–2097 (2019).
53. Rossi, M. A. et al. Obesity remodels activity and transcriptional state of a lateral hypothalamic brake on feeding. *Science* **364**, 1271–1274 (2019).
54. Smillie, C. S. et al. Intra- and inter-cellular rewiring of the human colon during ulcerative Colitis. *Cell* **178**, 714–730.e22 (2019).
55. Tran, N. M. et al. Single-cell profiles of retinal ganglion cells differing in resilience to injury reveal neuroprotective genes. *Neuron* **104**, 1039–1055.e12 (2019).
56. Goldfarbmuren, K. C. et al. Dissecting the cellular specificity of smoking effects and reconstructing lineages in the human airway epithelium. *Nat. Commun.* **11**, 2485 (2020).
57. Nagy, C. et al. Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. *Nat. Neurosci.* **23**, 771–781 (2020).
58. Huang, B. et al. Mucosal profiling of pediatric-onset Colitis and IBD reveals common pathogenics and therapeutic pathways. *Cell* **179**, 1160–1176.e24 (2019).
59. Wilk, A. J. et al. A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nat. Med.* **26**, 1070–1076 (2020).
60. Asboth, L. et al. Cortico-reticulo-spinal circuit reorganization enables functional recovery after severe spinal cord contusion. *Nat. Neurosci.* **21**, 576–588 (2018).
61. Wenger, N. et al. Spatiotemporal neuromodulation therapies engaging muscle synergies improve motor control after spinal cord injury. *Nat. Med.* **22**, 138–145 (2016).
62. Anderson, M. A. et al. Required growth facilitators propel axon regeneration across complete spinal cord injury. *Nature* **561**, 396–400 (2018).
63. Scheff, S. W., Rabchevsky, A. G., Fugaccia, I., Main, J. A. & Lumpp, J. E. Experimental modeling of spinal cord injury: characterization of a force-defined injury device. *J. Neurotrauma* **20**, 179–193 (2003).
64. Squair, J. W. *et al.* Integrated systems analysis reveals conserved gene networks underlying response to spinal cord injury. *elife* **7**, (2018).
65. Courtine, G. et al. Transformation of nonfunctional spinal circuits into functional states after the loss of brain input. *Nat. Neurosci.* **12**, 1333–1342 (2009).
66. Takeoka, A., Vollenweider, I., Courtine, G. & Arber, S. Muscle spindle feedback directs locomotor recovery and circuit reorganization after spinal cord injury. *Cell* **159**, 1626–1639 (2014).
67. Dominici, N. et al. Versatile robotic interface to evaluate, enable and train locomotion and balance after neuromotor disorders. *Nat. Med.* **18**, 1142–1147 (2012).
68. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
69. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
70. Zeisel, A. et al. Molecular architecture of the mouse nervous system. *Cell* **174**, 999–1014.e22 (2018).
71. Grimm, D. et al. In vitro and in vivo gene therapy vector evolution via multispecies interbreeding and retargeting of adeno-associated viruses. *J. Virol.* **82**, 5887–5911 (2008).
72. Anderson, M. A. et al. Astrocyte scar formation aids central nervous system axon regeneration. *Nature* **532**, 195–200 (2016).
73. Tomer, R., Ye, L., Hsueh, B. & Deisseroth, K. Advanced CLARITY for rapid and high-resolution imaging of intact tissues. *Nat. Protoc.* **9**, 1682–1697 (2014).
74. Voigt, F. F. et al. The mesoSPIM initiative: open-source light-sheet microscopes for imaging cleared tissue. *Nat. Methods* **16**, 1105–1108 (2019).
75. Arneson, D. et al. Single cell molecular alterations reveal target cells and pathways of concussive brain injury. *Nat. Commun.* **9**, 3894 (2018).
76. Avey, D. et al. Single-cell RNA-Seq uncovers a robust transcriptional response to morphine by Glia. *Cell Rep.* **24**, 3619–3629.e4 (2018).
77. Aztekin, C. et al. Identification of a regeneration-organizing cell in the Xenopus tail. *Science* **364**, 653–658 (2019).
78. Bhattacherjee, A. et al. Cell type-specific transcriptional programs in mouse prefrontal cortex during adolescence and addiction. *Nat. Commun.* **10**, 4169 (2019).
79. Brenner, E. et al. Single cell transcriptome profiling of the human alcohol-dependent brain. *Hum. Mol. Genet.* **29**, 1144–1153 (2020).
80. Cheng, C.-W. et al. Ketone body signaling mediates intestinal stem cell homeostasis and adaptation to diet. *Cell* **178**, 1115–1131.e15 (2019).
81. Co, M., Hickey, S. L., Kulkarni, A., Harper, M. & Konopka, G. Cortical foxp2 supports behavioral flexibility and developmental dopamine D1 receptor expression. *Cereb. Cortex* **30**, 1855–1870 (2020).
82. Davie, K. et al. A single-cell transcriptome Atlas aging Drosophilla brain. *Cell* **174**, 982–998.e20 (2018).
83. Denisenko, E. et al. Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol.* **21**, 130 (2020).
84. Der, E. et al. Tubular cell and keratinocyte single-cell transcriptomics applied to lupus nephritis reveal type I IFN and fibrosis relevant pathways. *Nat. Immunol.* **20**, 915–927 (2019).
85. Gunner, G. et al. Sensory lesioning induces microglial synapse elimination via ADAM10 and fractalkine signaling. *Nat. Neurosci.* **22**, 1075–1088 (2019).
86. Haber, A. L. et al. A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).
87. Hashimoto, K. et al. Single-cell transcriptomics reveals expansion of cytotoxic CD4 T cells in supercentenarians. *Proc. Natl Acad. Sci. USA* **116**, 24242–24251 (2019).
88. Hu, P. et al. Dissecting cell-type composition and activity-dependent transcriptional state in mammalian brains by massively parallel single-nucleus RNA-seq. *Mol. Cell* **68**, 1006–1015.e7 (2017).
89. Jaitin, D. A. et al. Lipid-associated macrophages control metabolic homeostasis in a Trem2-dependent manner. *Cell* **178**, 686–698.e14 (2019).
90. Jäkel, S. et al. Altered human oligodendrocyte heterogeneity in multiple sclerosis. *Nature* **566**, 543–547 (2019).
91. Kim, D.-W. et al. Multimodal analysis of cell types in a hypothalamic node controlling social behavior. *Cell* **179**, 713–728.e17 (2019).
92. Kotliarov, Y. et al. Broad immune activation underlies shared set point signatures for vaccine responsiveness in healthy individuals and disease activity in patients with lupus. *Nat. Med.* **26**, 618–629 (2020).
93. Madissoon, E. et al. scRNA-seq assessment of the human lung, spleen, and esophagus tissue stability after cold preservation. *Genome Biol.* **21**, 1 (2019).
94. Nault, R., Fader, K. A., Bhattacharya, S. & Zacharewski, T. R. Single-nuclei RNA sequencing assessment of the hepatic effects of 2,3,7,8-Tetrachlorodibenzo-p-dioxin. *Cell. Mol. Gastroenterol. Hepatol.* **11**, 147–159 (2021).
95. Ordovas-Montanes, J. et al. Allergic inflammatory memory in human respiratory epithelial progenitor cells. *Nature* **560**, 649–654 (2018).
96. Reyes, M. et al. An immune-cell signature of bacterial sepsis. *Nat. Med.* **26**, 333–340 (2020).

97. Schafflick, D. et al. Integrated single cell analysis of blood and cerebrospinal fluid leukocytes in multiple sclerosis. *Nat. Commun.* **11**, 247 (2020).
98. Schirmer, L. et al. Neuronal vulnerability and multilineage diversity in multiple sclerosis. *Nature* **573**, 75–82 (2019).
99. Wagner, D. E. et al. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).
100. Wang, S. et al. Single-cell transcriptomic atlas primate ovarian aging. *Cell* **180**, 585–600.e19 (2020).
101. Wirka, R. C. et al. Atheroprotective roles of smooth muscle cell phenotypic modulation and the TCF21 disease gene as revealed by single-cell analysis. *Nat. Med.* **25**, 1280–1289 (2019).
102. Ximerakis, M. et al. Single-cell transcriptomic profiling of the aging mouse brain. *Nat. Neurosci.* **22**, 1696–1708 (2019).

Acknowledgements

We thank L. Adlung, I. Amit, D. Anderson, C. Antelope, D. Arneson, D. Avey, M. Basiri, E. Brenner, G. Chew, M. Co, E. Der, A. Haber, K. Hashimoto, D. Kim, G. Konopka, A. Misharin, R. Mitra, J. Polo, M. Reyes, T. Quertermous, R. Wirka, and O. Yilmaz for providing data and/or cell type annotations. We acknowledge the Advanced Lightsheet Imaging Center (ALICe) at the Wyss Center for Bio and Neuroengineering, Geneva and Katia Galan for their tissue clearing and imaging support. This work was supported by a Consolidator Grant from the European Research Council [ERC-2015-CoG HOW2-WALKAGAIN 682999] (to G.C.) and the Swiss National Science Foundation (to G.C.; subside 310030_192558). This work was also supported in part by the Intramural Research Program of the NIH, NINDS (to K.J.E.M. and A.L.). Computational resources that supported this work were provided by the Swiss National Supercomputing Center, WestGrid, Compute Canada, and Advanced Research Computing at the University of British Columbia. M.A.S. acknowledges support from the Wings for Life Spinal Cord Research Foundation, the Canadian Institutes of Health Research (CIHR) (Vanier Canada Graduate Scholarship, Michael Smith Foreign Study Supplement), a Vancouver Coastal Health–CIHR–UBC MD/PhD Studentship, and an IUBMB Wood-Whelan Fellowship. J.W.S. is supported by a CIHR Banting Postdoctoral fellowship and a Marie Skłodowska-Curie Individual Fellowship (no. 842578). M.A.A. is supported by a SNF Ambizione fellowship (PZ00P3_185728) and Wings for Life. G.L.M. was supported by the CZI seed network grant HCA3-0000000081 and Swiss National Science Foundation grants CRSK-3_190495 and PZ00P3_193445. We are grateful to Jimmy Ravier for artistic contributions to Fig. 5c and Supplementary Fig. 8a. All other figures were created by the authors.

Author contributions

J.W.S. and M.A.S. performed computational analysis, with contributions from M.G. J.W.S., C.K., M.A.A., N.D.J., T.H. and Q.B. performed experimental validation, including RNAscope, immunohistochemistry, CLARITY, microscopy, and electrophysiology. J.W.S., M.G. and Q.B. analyzed experimental validation data. C.K., K.J.E.M. and A.J.L. performed nucleus extraction and single-nucleus RNA-seq. J.W.S., M.G., R.H., T.Q. and M.A.S. carried out the literature review. G.L.M. contributed to study design. M.A.S. and G.C. supervised the work. J.W.S., M.A.S. and G.C. wrote the manuscript. All authors contributed to its editing.

Competing interests

G.C. is a founder and shareholder of Onward Medical, a company with no direct relationships with the present work. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-25960-2.

Correspondence and requests for materials should be addressed to Michael A. Skinnider or Grégoire Courtine.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021