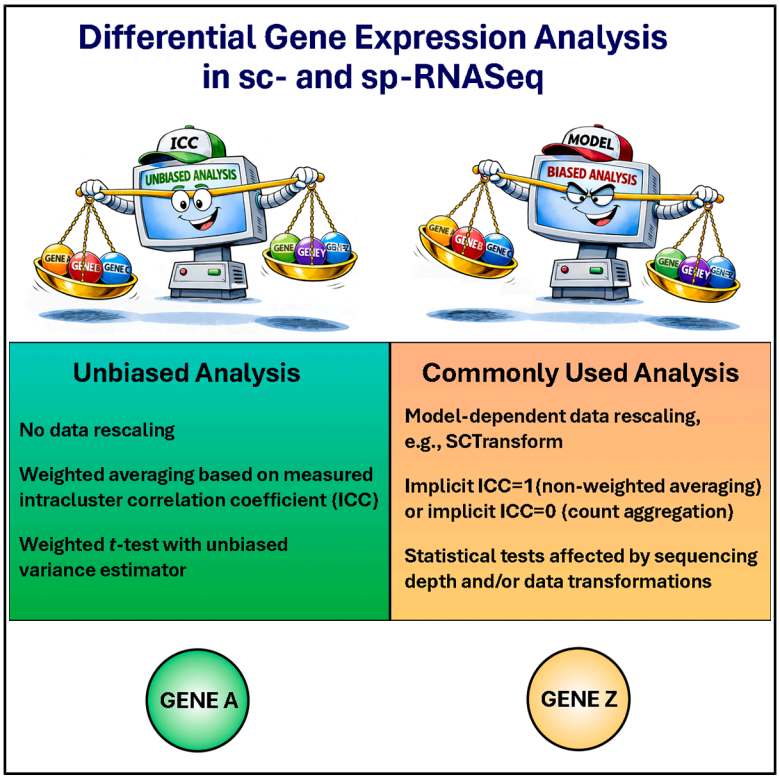
Cell Reports
Methods

Article

Differential expression analysis in single-cell and spatial RNA-seq without model assumptions

Graphical abstract



Highlights

- We propose methods for mitigating bias in differential gene expression analysis
- Intracluster, within-cell correlations between transcripts are derived from data
- Weighted data analyses based on measured intracluster correlations are utilized
- Artifacts caused by bias of common implicit and model assumptions are analyzed

Authors

Gennady Margolin, Andrew Tang, Sergey Leikin

Correspondence

leikins@mail.nih.gov

In brief

Differential expression analysis is key for identifying genetic checkpoints in biomedical research and diagnostics. Margolin et al. introduce simple tools for more unbiased expression analysis in single-cell and spatial transcriptomics. This analysis relies only on the data and eliminates biases and artifacts caused by unnecessary implicit and model assumptions.

Margolin et al., 2026, Cell Reports Methods 6, 101383
May 18, 2026 Published by Elsevier Inc.
https://doi.org/10.1016/j.crmeth.2026.101383

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

Differential expression analysis in single-cell and spatial RNA-seq without model assumptions

Gennady Margolin,[1] Andrew Tang,[1] and Sergey Leikin[1,2,*]
[1]Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD 20892, USA
[2]Lead contact
*Correspondence: leikins@mail.nih.gov
https://doi.org/10.1016/j.crmeth.2026.101383

MOTIVATION Widely accepted data analysis approaches for single-cell and spatial transcriptomics often incorporate model-dependent data pre-processing and/or additional implicit and explicit assumptions. This raises the possibility that model assumptions are introducing biases into these analyses. We propose a simpler analysis method that involves no data pre-processing and no assumptions beyond random, independent transcript sampling. This method is closely related to well-established approaches to resolving similar problems in cluster-randomized experiments.

SUMMARY

Gene up(down)regulation findings in single-cell and spatial RNA sequencing can be inconsistent despite remarkable progress in technology. Inconsistent findings in high-quality samples raise concerns about assumptions behind widely accepted data analysis approaches. We, therefore, propose a weighted averaging approach for data analysis without assuming anything besides randomness of technical noise. This approach is closely related to prior work on statistics of cluster-randomized experiments. We show that weighing transcript counts based on measured noise variances and utilizing weighted rather than standard unweighted tests reduce both false-positive and false-negative findings. Our approach eliminates the need for parametrizing data distributions and/or rescaling transcript counts, which may cause artifacts by distorting and biasing the data. The resulting analysis is less complex and produces more consistent differential gene expression estimates.

INTRODUCTION

Identification of up- and downregulated genes from single-cell and spatial transcriptomics is a key element of many biological studies. Statistical analysis of this differential gene expression (DGE) is increasingly important, yet false findings are a known problem in single-cell RNA sequencing (scRNA-seq).[1,2] Many different approaches to data analysis designed to increase statistical power while reducing false findings have been proposed. However, all commonly used[3–7] and more recent models and methods[8–27] for this analysis still rely on unnecessary simplifications, data fitting, and/or parametrization that may cause false findings. For instance, the Wilcoxon test does not take into account up to 100-fold variation in the number of detected transcripts from cell to cell. A commonly used "pseudo-bulk" (PB) approach involves an assumption that cells contribute proportionally to the number of transcripts detected in them, like in bulk RNA-seq. We show that such implicit assumptions in common DGE analysis approaches are not consistent with experimental data. In scRNA-seq, each cell is an independent experimental cluster, the number of detected transcripts is the cluster size, and effects of the cluster size and its variability have been well established before in statistics of cluster-randomized experiments.[28–31]

We also find that parametrization of the transcript distribution (e.g., with negative binomial one to rescale the counts in SCTransform [SCT][7]) is not beneficial for the analysis of good quality scRNA-seq or spatial RNA sequencing (spRNA-seq) data. *In silico* modeling of the ever-present sequencing depth variation and other tests show that altering the original data based on parametrizations and widely used logarithmic normalization may produce significant artifacts in DGE analysis.

Here, we propose a new approach to DGE analysis in scRNA- and spRNA-seq, which makes no assumptions beyond random technical noise and sampling to more accurately describe the data. We use known features of scRNA-seq physics to derive the needed data distribution moments directly from the

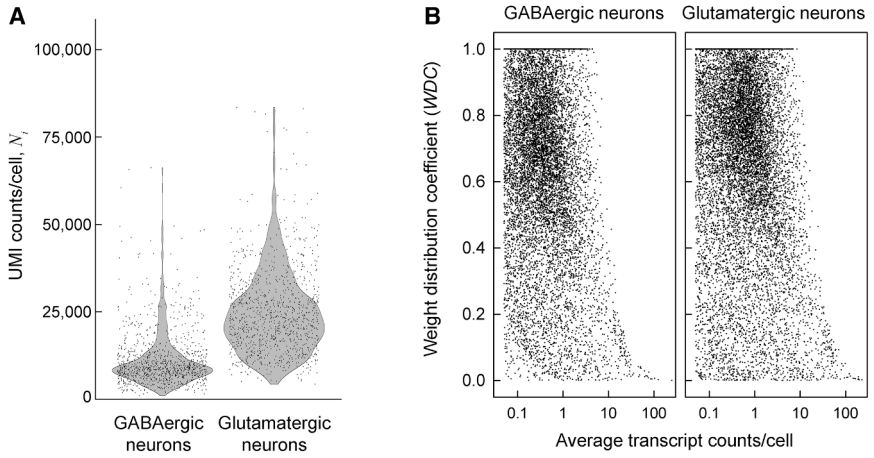


Figure 1. Measured statistical weights of cells in E18 mouse brain GABAergic and glutamatergic neurons do not match assumptions of common DGE analysis models

(A) Total transcript (UMI) counts per cell N_i in a combined E18 mouse brain dataset from 10× Genomics (see STAR Methods). Each data point represents an individual cell.

(B) Weight distribution coefficient (*WDC*) for the statistical weight of different mRNA transcripts. Each point shows WDC_Z for a gene Z calculated directly from the measured transcript counts as described in STAR Methods. Standard, unweighted averaging of transcript counts utilized in most DGE analysis models assumes $WDC_Z = 0$. Utilization of aggregate counts in pseudo-bulk models assumes $WDC_Z = 1$. Neither reflects the reality.

measured transcript counts without assuming a specific form of data distribution or modeling it. We do not assume the same transcript counting accuracy in different cells because the observed cell-to-cell variations in the total counts suggest highly variable accuracy. Instead, we deduce relative contributions of different cells to the mean gene expression and its variance directly from the data by utilizing well-established methods of unbiased weighted averaging.[32]

Comparison of the results produced by our method with commonly used approaches revealed significant differences in the results. We, therefore, benchmarked all methods with simulated data, experimental data modified by well-defined changes in expression, and simpler count subsampling in experimental data that mimics variations in the sequencing depth. Based on the known ground truth, this statistical testing revealed many more false-positive and false-negative findings in the commonly used approaches compared with those in our method. We invite interested readers to examine our findings and judge the reliability of the output produced by the commonly used methods on their own. The annotated R code for our tests and simulations is available at https://github.com/sergeyleikin/sc-sp-RNASeq.

Understanding the basic principles behind our approach and its practical usage does not require any special expertise. We define all concepts and explain their meaning in the main text, while focusing on theoretical proofs and derivations in STAR Methods. We implemented all procedures for our data analysis as a simple set of functions compatible with data objects generated by the popular Seurat package for R programming language (https://satijalab.org/seurat/). Benchmarking of their computational performance vs. commonly used Seurat functions is provided in STAR Methods. Our set of functions and their source codes can be downloaded at https://github.com/sergeyleikin/sc-sp-RNASeq. For simplicity and consistency, we compare the results produced by our method with the standard models within the Seurat workflow, using publicly available 10× Genomics scRNA-seq datasets (https://www.10xgenomics.com/datasets) and our spRNA-seq data collected with the 10× Genomics Visium HD assay. We first formulate our arguments and tools for scRNA-seq and then adapt them to spRNA-seq at the end of the article.

RESULTS

Statistical weights of cells significantly affect the average expression of many genes

scRNA-seq data represent a matrix $N_{Z,i}$ of unique gene Z transcript counts in a cell i (unique molecular identifier [UMI] counts). Even functionally similar cells exhibit large (up to ~100-fold) variation in the total number of counts per cell N_i, as illustrated in Figure 1A for GABAergic and glutamatergic neurons from embryonic mouse brain. This variation may be affected by biological differences between cells and technical variability of the assay. Therefore, relative gene expression

$$n_{Z,i} = \frac{N_{Z,i}}{N_i}, \quad N_i = \sum_Z N_{Z,i} \qquad \text{(Equation 1)}$$

is recalculated from the data and further used in all statistical models.

The large variation in N_i, however, necessitates not only normalizing the counts but also accounting for the statistical weight $w_{Z,i}$ of each cell i when calculating the average relative expression $\overline{n}_Z$. In general,

$$\overline{n}_Z = \sum_{i=1}^{N} w_{Z,i} n_{Z,i}, \qquad \text{(Equation 2)}$$

where $\sum_i w_{Z,i} = 1$, and N is the number of cells, as has been well established in the statistics of experiments with nonuniform accuracy of data measurement.[32]

In scRNA-seq, $w_{Z,i}$ accounts for the measurement precision of gene Z counts in cell i and, therefore, may be gene dependent. The most accurate unbiased estimate of the mean value (here $\overline{n}_Z$) for a random variable ($n_{Z,i}$) is obtained when $w_{Z,i}$ is proportional to the inverse variance of the variable,[32] $1/Var(n_{Z,i})$. Highly variable N_i leads to $w_{Z,i}$ that changes from cell to cell and gene to gene.

The value of $w_{Z,i}$ can be calculated directly from the experimental data without any additional assumptions as follows:

$$w_{Z,i} = \frac{N_i}{1+\rho_Z(N_i-1)} \Big/ \sum_{i=1}^{N} \frac{N_i}{1+\rho_Z(N_i-1)} \qquad \text{(Equation 3)}$$

where $0 \leq \rho_Z \leq 1$ is the intracluster correlation coefficient (ICC) calculated from the data (STAR Methods).

Equation 3 and several methods for calculating ρ_Z are well known in the theory of cluster-randomized experiments, e.g., cluster-randomized clinical trials.[28,29] To illustrate the concept, consider reporting of deaths in a drug trial. Death rates $n_{Z,i}$ may vary significantly between participant clusters at different trial sites i. These rates are more reliable in clusters with thousands of participants (e.g., large cities) than in those with dozens of participants (e.g., small towns). They may be affected not only by the cluster size (N_i) but also by other factors, e.g., differences in lifestyle and healthcare between the clusters. Correlations in the latter factors within each cluster necessitate weighing the counts based on Equations 2 and 3.[28,29] In the absence of differences in these factors, the same death rate is expected in all clusters and $\rho_Z = 0$. When $\rho_Z = 0$, $w_{Z,i}$ is proportional to N_i ($w_{Z,i} = N_i/\sum_i N_i$), thus aggregating counts $N_{Z,i}$ from all clusters and yielding $\bar{n}_Z = \sum_i N_{Z,i}/\sum_i N_i$ in Equation 2. At very strong correlations ($\rho_Z = 1$), the averaging of cluster death rates $n_{Z,i}$ becomes unweighted (equally weighted with all $w_{Z,i} = 1/N$).

Generally, ρ_Z balances the effects of true outcome variability across clusters and imprecisions in measuring/estimating the outcome (relative counts such as death rate) in each cluster. When $\rho_Z \neq 0$, measurement imprecisions become negligible in large clusters with $\rho_Z N_i \gg 1$, all of which then have approximately the same weight that is larger than the weights of smaller clusters. At larger ρ_Z, more clusters are sufficiently large to be weighed equally.

In STAR Methods, we show that $n_{Z,i}$ averaging in scRNA-seq is conceptually similar and describe how ρ_Z can be calculated from scRNA-seq data without model assumptions. In scRNA-seq, ρ_Z balances the effects of true biological variance between cells and imprecisions in measuring $n_{Z,i}$. The importance of weighing clinical trial results and accounting for the ICC has been recognized decades ago. We argue that the scRNA-seq data analysis should recognize the importance of this approach as well.

Instead, almost all common approaches to scRNA-seq data analysis use the unweighted averaging (AV), $w_{Z,i} = w_{AV} = 1/N$, implicitly neglecting variable precision of the transcript count measurement from cell to cell ($\rho_Z = 1$ assumption). PB approaches to data analysis from multiple samples aggregate the counts (AC), $w_{Z,i} = w_{AC} = N_i/\sum_i N_i$, implicitly neglecting the correlated transcript expression within each cell and, thereby, the expected biological variance in transcript counts between different cells ($\rho_Z = 0$ assumption). Statistical implications of such assumptions have been extensively studied in the context of cluster-randomized experiments. Both are known to cause false findings when the cluster size and composition are highly variable,[28–31] which is the case in scRNA-seq.

To examine implications of these assumptions in a real scRNA-seq experiment, we selected cells expressing *Gad1*, *Gad2*, and *Slc6a1* (markers of GABAergic neurons) and cells expressing *Gls*, *Grin2b*, and *Slc17a7* (markers of glutamatergic neurons) from E18 mouse brain scRNA-seq data from 10× Genomics (see STAR Methods). These genes are not the only markers for the corresponding neurons, yet they are highly expressed and separate two minimally overlapping (<1%), well-defined, and sufficiently large (~1,000 cells each) subsets of data (Figure 1A). Without necessarily requiring the biological precision or accuracy, this was all we needed to test different DGE analysis models.

Instead of assuming $w_{Z,i} = w_{AV}$ ($\rho_Z = 1$) or $w_{Z,i} = w_{AC}$ ($\rho_Z = 0$), we evaluated ρ_Z and the corresponding $w_{Z,i}$ from measured $n_{Z,i}$ (STAR Methods) and calculated the weight distribution coefficient (WDC) defined as normalized variance of $w_{Z,i}$ ($WDC_Z = Var(w_{Z,i})/Var(w_{AC})$). At $w_{Z,i} = w_{AV}$, $WDC_Z = 0$. At $w_{Z,i} = w_{AC}$, $WDC_Z = 1$. WDC values calculated from the data provide useful information about cell-to-cell variability in gene expression and reveal that neither of the two approximations is accurate for most genes (Figure 1B).

$WDC_Z \approx 0$ means that $w_{Z,i}$ is largely unaffected by the sequencing depth, indicating that the expression of a gene is variable enough for the transcript sampling error to be small compared with the gene expression variability. This is observed for the most highly expressed genes due to the low sampling error (Figure 1B) and may also be expected for highly variable genes with moderate expression (e.g., genes regulating cell response to stress and environment). $WDC_Z \approx 1$ means dominant contribution of the sequencing depth to $w_{Z,i}$, indicating low expression variability from cell to cell and/or low expression of the gene. This may be expected, e.g., for tightly regulated cell maintenance genes with low-to-moderate expression. At the same expression level, a low WDC indicates high expression variability.

The values of $w_{Z,i}$ significantly affect the fold change (FC) in the average gene expression, which is the reported measure of DGE. For instance, log_2(FC) between glutamatergic and GABAergic neurons is significantly different for many genes when calculated with measured $w_{Z,i}$ rather than assuming $w_{Z,i} = w_{AV}$ (Figure 2A). For some genes, this changes even the sign of log_2(FC) so that upregulated genes may appear to be downregulated, and vice versa. The difference between weighted and unweighted FC can be more than 2-fold and have dramatic implications for data interpretation when the genes of interest are affected, even though weighted and unweighted log_2(FC) values are close for many genes.

Weighted statistical testing describes DGE significance without model assumptions

Statistical weight values affect not only the log_2(FC) value but also the statistical significance of DGE (the probability p of the observed log_2(FC) when the expected log_2(FC) is zero). To account for this, we cannot use common approaches to calculate p values in scRNA-seq. Not only do these approaches use assumed weights but they also rely on additional assumptions, sometimes implicit.

For instance, p values are commonly obtained from the Wilcoxon test (default in Seurat), in which all data points are ranked (1,2,3 …) in the order of increasing $n_{Z,i}$. This test is appropriate and useful when the question is whether the distributions, rather than mean values, of random variables are different.[33] However, its p value is not the probability of the observed log_2(FC) when the expected one is zero.[33] The Wilcoxon test may return low p values more often than expected when the cell-to-cell distributions of transcript counts in two samples are different, even when the relative expression is the same (log_2(FC) = 0). For

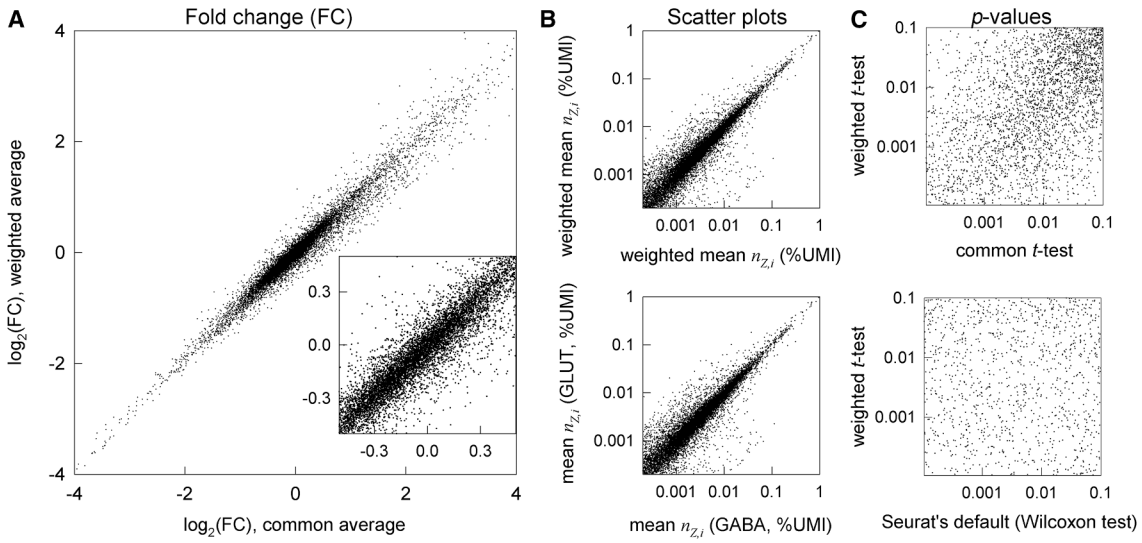


Figure 2. Statistical weights of cells significantly affect the calculated log_2(FC) and *p* values

(A) Effects of the statistical weights on log_2(FC) between glutamatergic vs. GABAergic neurons in E18 mouse brain (same as in Figure 1A) plotted for 12,087 genes with at least 50 total counts in glutamatergic neurons. 7% of these genes (847) have different signs of log_2(FC) for common and weighted averages. 17% of the latter (143) have the difference between the two log_2(FC) values > 0.25 and 3% (29) have the difference larger than 0.5.

(B) Scatterplots for the mean expression (bottom) and weighted mean expression (top) of these genes in GABAergic (*x* axis) and glutamatergic (*y* axis) neurons.

(C) Effects of the statistical weights on *p* values for genes plotted in (A). The common (unweighted) *t* test was performed based on relative count rather than logarithmic normalization to avoid artifacts of the latter. Among the genes with at least 50 total counts in glutamatergic neurons, the weighted *t* test found 3,993 genes with $p < 10^{-4}$ (of which 6 were not found by the common *t* test with $p < 0.05$ and 845 were not found by the Seurat's default Wilcoxon test with $p < 0.05$. The common *t* test found 3,747 genes with $p < 10^{-4}$ (of which 3 were not found by the weighted *t* test with $p < 0.05$). The Wilcoxon test found 8,045 genes with $p < 10^{-4}$ (of which 3,523 were not found by the weighted *t* test with $p < 0.05$), which is clearly inconsistent with the scatterplots for either mean expression or weighted mean expression.

instance, scatterplots indicated similar relative expression of most genes in glutamatergic and GABAergic neurons (Figure 2B). In contrast, the Wilcoxon test found over 60% of genes being differentially expressed with $p < 10^{-4}$ (Figure 2C), in part, because of very different total numbers of transcripts per cell in these neurons (Figure 1A). The variation in total (and consequently individual gene) transcript distribution is common in scRNA-seq because of different transcriptional activities of cells, different sequencing depths (detected fraction of transcripts), and effects of sample preparation. Similar details must be considered when applying common *t* test on log-normalized data in Seurat. The test compares the mean values of log-normalized counts (logarithms of scaled relative counts plus a pseudo count.) This differs from comparing the mean values of relative counts, and we show the consequences in the next section.

Rather than assuming any properties of transcript distribution explicitly or implicitly, we calculated *p* values directly from raw data, relying only on sufficient numbers of relevant cells (from ~100 to thousands of cells per group is becoming common and likely to increase). Mean values in such large datasets are expected to have normal distributions (central limit theorem[33]). Therefore, the corresponding *p* values can be found from a *t* test, a weighted version of which accounts for variable statistical weights (STAR Methods). This approach is similar to statistical testing of weighted averages in cluster-randomized experiments, which has been used for many years.[28,29,31] For instance, *p* values for DGE between glutamatergic and GABAergic neurons calculated from the weighted *t* test are very different from those calculated from the Wilcoxon test or the common (unweighted) *t* test (Figure 2C). Almost half of the genes with highly significant DGE ($p < 10^{-4}$) found by the Wilcoxon test were not found to be differentially expressed in the weighted or unweighted *t* test ($p \geq 0.05$, Figure 2C), consistent with the scatterplots for relative gene expression (Figure 2B).

It is important to keep in mind, however, that the weighted *t* test treats individual cells as independent experimental replicates (like all common approaches to scRNA-seq analysis). Depending on the experiment and cell type identification algorithm, this may not be a good approximation (STAR Methods). When multiple sets of experimental data are available, each independent set can be used as a biological replicate to resolve this problem. When such replicates are not available, answering a different statistical question with a supplemental χ^2 test may be useful for reducing the resulting false-positive findings.

The χ^2 test compares aggregate transcript counts just in the sequenced cells, rather than average counts per cell in the specimens from which the cells have been sampled (STAR Methods). Because it accounts only for variance in stochastic transcript sampling but not for variance in the sampling of cells or variance in gene expression from cell to cell, it may significantly underestimate the total variance and should not be used alone. However, it may mitigate cell sampling bias effects by excluding the weighted *t* test findings based on χ^2 test *p* values larger than a minimum significance threshold, e.g., 0.05.

To illustrate this and other approaches to combining the weighted t test with the χ^2 test, consider DGE between GABAergic and glutamatergic neurons selected for Figure 2. For these cells, the weighted t test detected 3,993 genes with $p < 10^{-4}$, among which 3 genes had χ^2 test $p \geq 0.05$ (no significant difference in the aggregate counts). The highly significant weighted t test for the latter genes was, therefore, caused by a difference in the transcript distribution among the cells, rather than a difference in the aggregate gene expression. This may be a real biological difference between the two types of neurons or just an effect of biased cell sampling from mouse brain. The best solution for this dilemma is replicate or independent experiments, eliminating the need for the χ^2 test. A moderately conservative proxy approach when such experimental data are not available is to use p value from the χ^2 test, instead of the weighted t test one, when the former but not the latter exceeds the minimum significance threshold. This is the approach utilized hereafter. One may also just select the larger of the two p values to minimize false-positive findings, which is the most conservative approach to combining the tests.

We found the DGE results based on the weighted t test-χ^2 test combination to be very different from those produced by standard DGE analysis approaches. In the same glutamatergic vs. GABAergic neuron example from Figure 2, the Wilcoxon test detected 3,629 genes with $p < 10^{-4}$ that were not significant ($p > 0.05$) in the weighted t test-χ^2 test combination (note that Figure 2 presents analysis based on the weighted t test alone). The latter approach detected 845 genes with $p < 10^{-4}$ that were not significant in the Wilcoxon test. Among these 845 genes, 147 had more than 2-fold difference in expression. For instance, the Wilcoxon test did not detect a change in the expression of cancer-associated genes *Tacc3* and *Birc5*, which had more than 0.2 counts/cell in both groups of neurons, more than 3-fold change in expression, and false discovery rate (FDR)-adjusted $p < 10^{-5}$ in the weighted t test-χ^2 test combination. Other tests built into Seurat's FindMarkers function had similarly large discrepancies with the weighted t test-χ^2 test combination when the logarithmic data normalization recommended by Seurat was used. The discrepancy with the unweighted t test was less dramatic when relative count normalization was used, yet it was significant (Figure 2B). For instance, the unweighted t test found 21 genes with $p < 10^{-4}$ that had $p \geq 0.05$ in the weighted t test-χ^2 test combination.

Weighted statistical testing improves robustness and sensitivity of DGE analysis

This analysis suggests that traditional approaches may produce misleading results, yet it does not prove their failure. Comparison of experimental data reveals major differences between model predictions but cannot establish which model is correct because the ground truth in these experiments is unknown. It is, therefore, also important to test predictions of different models by using simulated data that are based on real-world experiments yet model well-defined data patterns, in which the ground truth is known.

First, we tested the simplest pattern of UMI subsampling in glutamatergic neurons (Figures 3A–3G). To generate this pattern, we used glutamatergic neuron data, in which we randomly subsampled all UMIs for all genes with 40% probability of retaining each UMI. Such 40% subsampling mimics reduced sequencing depth (60% fewer counts detected by sequencing), which is well within the normal experimental variation. We performed 10 random subsampling simulations, using only genes with more than 30 total counts in the original data, and analyzed DGE effects of the subsampling. Because random subsampling does not change the relative gene expression, any DGE finding in this simulation is a false-positive one. This simulation revealed dramatic failure of all commonly used DGE analysis models (Figures 3C, 3D, 3F, and 3G). The false-positive DGE was still unacceptable at 60% subsampling (Figures S1A–S1G).

The weighted t test, χ^2 test, and their combination had no false-positive findings in any of the 10 simulations, even at $p < 0.05$ cutoff (Figures 3A and 3B). In contrast, Seurat's FindMarkers function with default Wilcoxon test consistently identified over 75% of all genes as differentially expressed, with $p < 0.001$, and over 60% as differentially expressed, with $p < 10^{-5}$ (Figures 3C and 3G). Seurat's FindMarkers function with the common t test option consistently identified over 20% of all genes with $p < 0.001$ and >1% with $p < 10^{-5}$. Over 1,000 genes with highly significant false DGE discovered by the default Seurat's pipeline for DGE analysis had $|log_2(FC)|>0.1$ (the default threshold in Seurat) and dozens had $|log_2(FC)|>0.3$ (as reported by Seurat). The results for other tests built into Seurat were qualitatively similar. When the commonly recommended logarithmic normalization (default in Seurat) was replaced with nonparametric RC (relative count) one (Equation 1), the unweighted t test did not identify DGE (Figures 3E and 3G), revealing unacceptable artifacts of logarithmic normalization (which affects the t test but not the Wilcoxon test).

The main cause of the dramatic failure of standard DGE analysis models was count dropout (loss of some nonzero counts due to subsampling). This is a well-known effect of sample-to-sample variations in sequencing depth and/or RNA recovery from cells, which is one of the reasons behind model-dependent data regression and/or rescaling.[34–37] RC normalization combined with the weighted t test-χ^2 test combination or the unweighted t test makes such model-dependent data processing unnecessary.

We found that SCT—one of widely used data rescaling procedures—was inefficient in reducing DGE caused by UMI subsampling and produced its own artifacts. SCT, which rescales counts based on negative binomial regression, was developed, in part, to reduce the count dropout effects.[7] We tested how SCT version 2 with default parameters affected DGE analysis in Seurat. Even the most conservative weighted t test revealed false DGE between original data before and after SCT, i.e., SCT introduced artificial expression patterns absent from the data (Figure 3F). When SCT was used to generate "corrected" counts for original and subsampled data independently, preserving the median sequencing depth in each of them, it increased, rather than reducing, the artifacts of count dropouts. It increased false-positive DGE in the Wilcoxon test and caused false DGE, even in the weighted t test-χ^2 test combination (Figure 3G). These SCT artifacts were particularly strong for low-expression genes with 30–50 counts per 1,000 cells and gradually disappeared at higher expression. When the original and subsampled

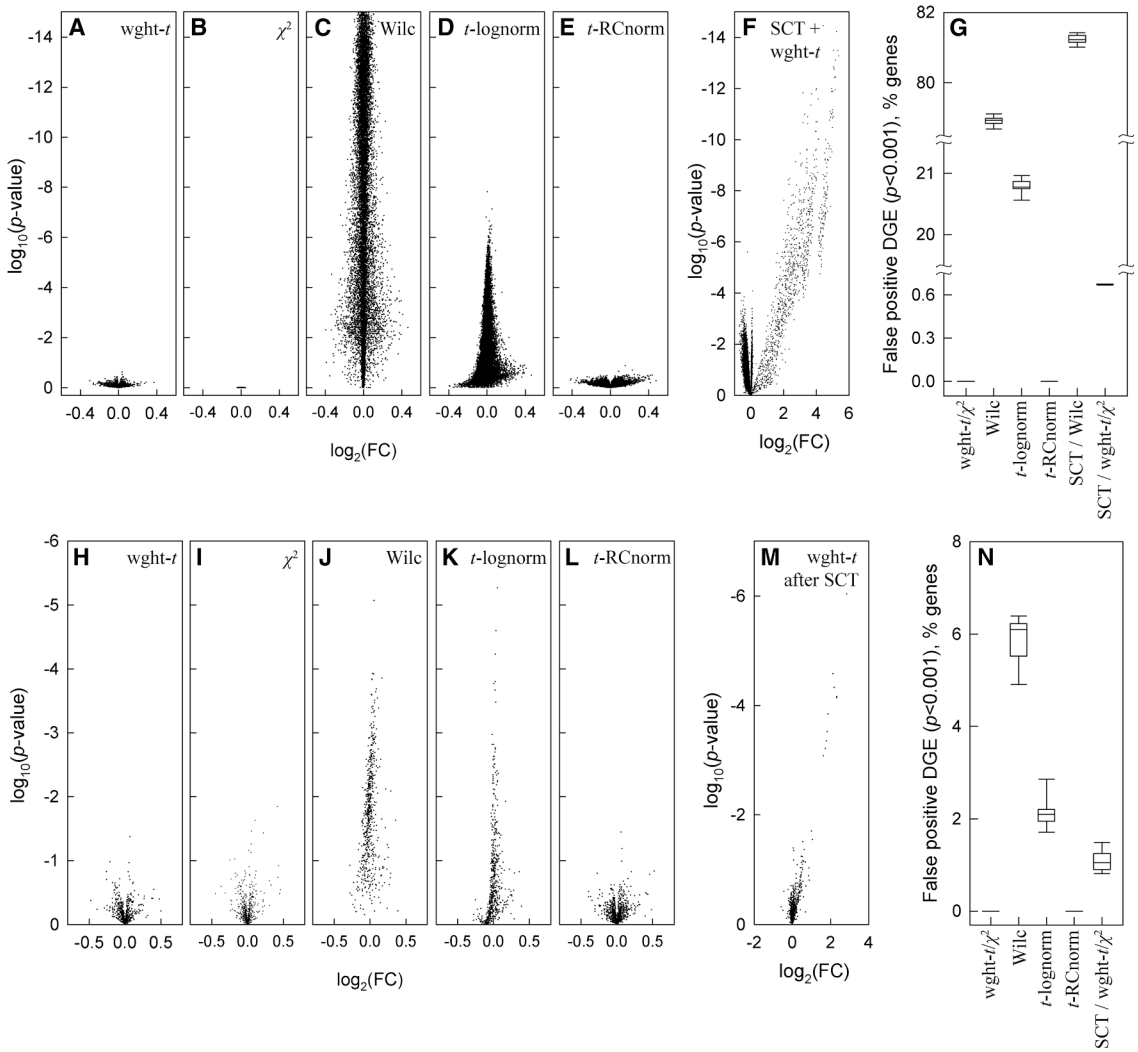


Figure 3. Weighted statistical testing eliminates false-positive DGE in the analysis of original vs. subsampled glutamatergic neurons data

(A–G) False DGE discovery after SCTransform (SCT) or after 40% count subsampling (retaining 40% of counts for each gene by randomly selecting counts to be removed). 12,911 genes are shown, each with more than 30 counts in all cells in the original data.

(A–E) Volcano plots for a single subsampling simulation after weighted *t* test (A), χ^2 test (B), Seurat's FindMarkers function default Wilcoxon test (C), Seurat's FindMarkers optional *t* test for logarithmically normalized data (D), and *t* test for RC normalized data (E, Equation 1).

(F) DGE artificially introduced into the data by count rescaling with SCT, as revealed by weighted *t* test comparison of the data before and after SCT.

(G) Fraction of the 12,911 genes (%) falsely discovered as differentially expressed with $p < 0.001$ after 10 subsampling simulations.

(H–N) DGE analysis for 1,000 randomly selected genes (each with more than 30 total counts in all cells) altered by random, uniformly distributed noise.

(H–L) Volcano plots for the same tests as in (A–E).

(M) Volcano plot for weighted *t* test comparison preceded by SCT of both original and noise-altered data. (N) Fraction the 1,000 noise-altered genes falsely discovered as differentially expressed with $p < 0.001$ after 10 simulations.

Boxes show the 25th–75th percentile range, whiskers show the 10th–90th percentile range, and lines show median values (in G and N).

data were merged before SCT to equalize the sequencing depth of all transformed data, the artifacts were much stronger and affected over 35% of all genes, including moderate-expression ones (Figure S1H).

Next, we altered a subset of 1,000 randomly selected genes in the glutamatergic neuron gene count matrix $N_{Z,i}$ (the dataset in Figure 2A) by random, uniformly distributed noise. Again, only genes with more than 30 total counts in all cells were included. The added noise had zero average amplitude, maximum amplitude between $0.8N_{Z,i}$ and $N_{Z,i}$ at $N_{Z,i} \leq 5$, and variance $5N_{Z,i}$ at large $N_{Z,i}$ (see STAR Methods). 10 random sets of 1,000 genes were tested. The noise was randomly regenerated for each simulation.

Again, the weighted *t* test-χ^2 test combination and common *t* test after RC normalization found no false-positive DGE (Figures 3H, 3I, 3L, and 3N). The other tests discovered DGE for many more genes than expected by chance at the corresponding *p* values (Figures 3J, 3K, 3M, and 3N). Like in the

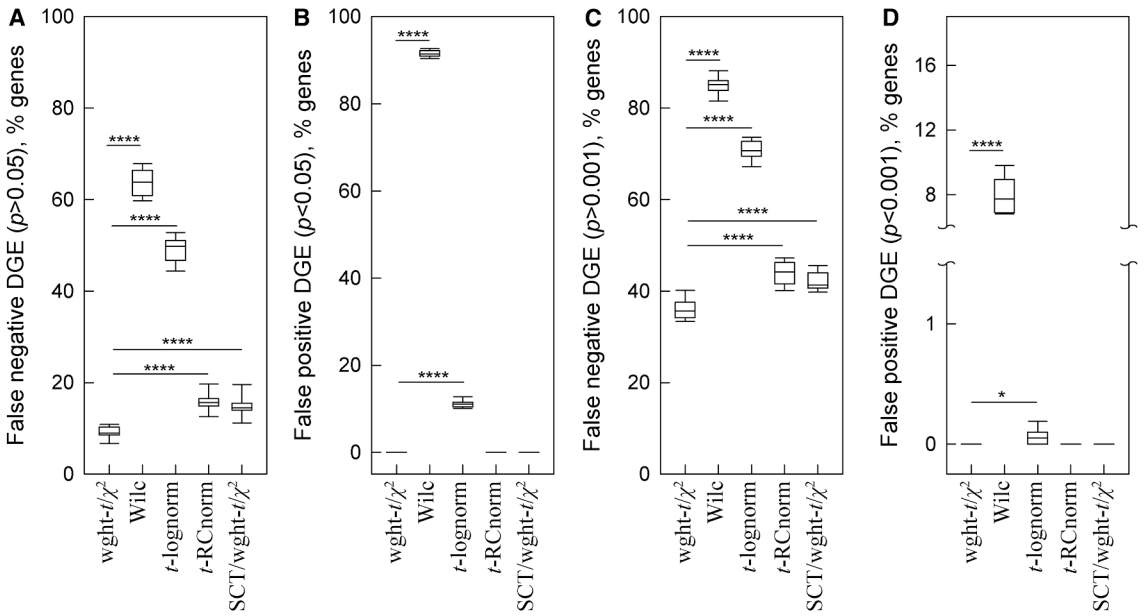


Figure 4. Weighted statistical testing reduces false-negative and false-positive DGE in simulations of gene upregulation superimposed with comparable noise

Random subsets of moderately-expressed genes (0.3–3 counts per cell) in glutamatergic neurons were selected for the simulations.

(A) False-negative DGE ($p > 0.05$ and/or log_2(FC) < 0.1) among 300 genes modified by 50% increase in expression and comparable random noise. Weighted *t* test-χ^2 test combination was used without SCT (wght-*t*/χ^2) and with SCT (SCT/wght-*t*/χ^2). Wilcoxon test (Wilc), *t* test for log-normalized data (*t*-lognorm), and *t* test for RC normalized data (*t*-RCnorm) were used without SCT.

(B) False-positive DGE ($p < 0.05$ and log_2(FC) > 0.1) among 1,000 genes modified only by random noise.

(C and D) Same as (A) and (B), respectively, but at $p < 0.001$ threshold. The box and whiskers parameters are the same as in Figure 3G; **** means $p < 0.0001$ and * means $p < 0.05$.

case of count subsampling, SCT produced rather than removed false DGE. Because the simulated noise produced fewer count dropouts than the count subsampling, its effects were less dramatic yet significant. Overall, the effects of random noise and count subsampling were very similar.

We then tested how well different models discover DGE masked by noise and tested whether SCT helped the best-performing model to suppress noise and reveal DGE. For these simulations, we combined the same random noise as above with a proportional increase (50%) in the counts of a 300-gene subset of the 1,000 genes modified by noise.

The weighted *t* test-χ^2 test combination was by far the most accurate and reliable (Figure 4). Its benefits were most apparent for moderate-expression genes (average 0.3 to 3 counts per cell), in which the added noise was comparable to or higher than added DGE signal. It missed significantly fewer upregulated genes compared with other tests (Figures 4A and 4C) and did not detect any false DGE (Figures 4B and 4D). The *t* test for RC normalized data did not detect false DGE either, yet it missed more upregulated genes. The Wilcoxon test and *t* test for log-normalized data reported significant false DGE.

In moderately expressed genes, SCT did not cause the detection of false DGE, but it reduced the fraction of upregulated genes detected by the weighted *t* test-χ^2 test combination (Figure 4A). In genes with average <0.3 counts/cell, the DGE signal was fully masked by noise and not detected by any tests. In genes with average >3 counts/cell, the DGE signal was stronger than the noise and reliably detected by all tests.

Finally, we confirmed the weighted *t* test performance by the analysis of simulated data with predefined data distributions (Figure S5). Because all *in silico* cells were independent replicates, in these simulations, we used only the weighted *t* test but not its combination with the χ^2 test. The weighted *t* test had the best combination of power for discovering true DGE (low false-negative findings) with type I error rate control (low false-positive findings). The Wilcoxon test had poor type I error rate control when total numbers of transcripts per cell in the two samples were different (an analogue of different sequencing depths). The unweighted *t* test was too conservative and had significantly higher false-negative findings.

Weighted statistical testing outperforms the PB approach

When scRNA-seq data are available from multiple samples per genotype or treatment, the biological variability and differences in measurement accuracy between the samples become important. The PB approach, which has been developed and is commonly used for analysis of such experiments,[1,38] however, involves several questionable and unnecessary simplifying approximations.

First, the PB approach uses count aggregation within each sample, instead of weighted averaging, implicitly assuming $w_{Z,i} = w_{AC} = N_i/\sum_i N_i$. As discussed above, this

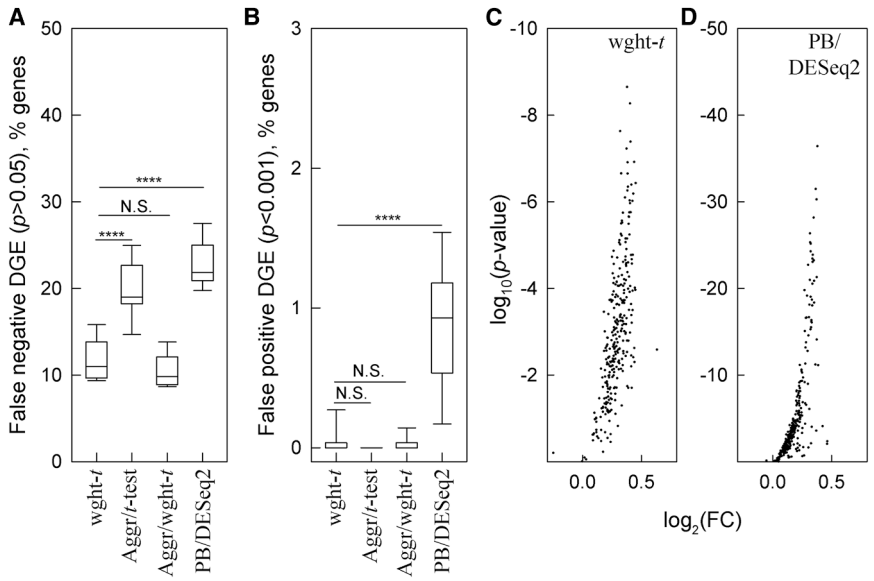


Figure 5. Weighted statistical testing outperforms the pseudo-bulk approach to DGE analysis

Two 5-sample sets were generated from glutamatergic neuron data by random cell subsetting. Like in Figure 4, we added random noise to a 1,000 gene set that had 0.3–3 counts/cell and 30% increase in expression to a 300-gene subset.

(A) False-negative DGE among 300 genes with increased expression (wght-*t* is the weighted *t* test for weighted average counts, Aggr/*t*-test and Aggr/wght-*t* are unweighted and weighted *t* tests for aggregate counts, respectively, and PB/DESeq2 is pseudo-bulk aggregate count analysis with DESeq2).

(B) False-positive DGE among 700 genes modified only by noise.

(C and D) Volcano plots for the genes modified by the expression increase and noise analyzed by weighted *t* test (C) and PB/DESeq2 (D).

****, $p < 0.0001$; N.S., not significant (A and B).

approximation may work well for some genes, yet it may significantly affect the results for many other genes (Figure 1B).

Second, the PB approach neglects sample-to-sample variations in measurement accuracy. This accuracy can be characterized by the variance of $\bar{n}_{Z,s} = \sum_{i=1}^{N_s} w_{Z,i,s} n_{Z,i,s}$ for each sample *s*. To be internally consistent, the aggregated count analysis must use weighted average of $\bar{n}_{Z,s}$ with statistical weights dependent on this variance,[32] which can be calculated from the experimental data as described in STAR Methods.

Third, the PB approach determines the variances for statistical testing from negative binomial regression (e.g., using DESeq2[39] for identifying DGE in Seurat). Instead, these variances can be calculated directly from the data without any approximations as described in STAR Methods.

To avoid these approximations, one can use $n_{Z,i}$ measured in each scRNA-seq experiment to determine the experimental values of $w_{Z,i}$ as described above followed by calculating the corresponding $\bar{n}_{Z,s}$ and $Var(\bar{n}_{Z,s})$. The log_2(FC) and *p* values can then be calculated by using the weighted *t* test, with weights determined from the experimental data as described in STAR Methods. Although $\bar{n}_{Z,s}$ distribution may not be normal, the *t* test is still the most robust approach when only a small number of replicates are available, and the statistical distribution cannot be reliably established.[40] This is almost always the case in high-cost scRNA-seq experiments.

Note that a PB approach incorporating weights was also described, but these weights were based on fitting logarithmically normalized data to a generalized linear model rather than determined directly from the data without model assumptions.[41,42]

To test our minimal assumptions approach, we again simulated datasets with the same well-defined noise and DGE signal based on the same glutamatergic neuron data. We generated 10 samples by randomly subsampling the cells from the original dataset. To mimic the effects of variable sample size, the samples contained 20%, 40%, 60%, 80%, or 100% of the neurons. In 5 of the samples, 1,000 moderate-expression genes (average 0.3–3 counts per cell) were modified by random noise. In the other 5 samples, the same 1,000 genes were modified by similar random noise and by an increase in the expression of a 300-gene subset. The noise-only and noise + DGE signal-modified samples were then compared using the weighted *t* test for weighted average counts (Figure 5, wght-*t*), unweighted *t* test for aggregate counts (Figure 5, Aggr/*t*-test), weighted *t* test for aggregate counts (Figure 5, Aggr/wght-*t*), and DESeq2 test for aggregate counts (Figure 5, PB/DESeq2, Seurat's default for traditional PB analysis).

Because we found multiple sample analysis of aggregate counts to be less sensitive to noise and more sensitive to changes in expression, we made the comparison more stringent than in Figure 4 and increased the expression by only 30%. We also selected only 500 from the 1,000 modified genes randomly. We selected the other 500 as the genes with the largest cell-to-cell expression variability (coefficient of variation), while we still selected a 300-gene subset of the 1,000 genes for expression upregulation randomly.

These simulations supported our theoretical deductions. Weighted *t* test again detected the largest fraction of upregulated genes, and weighted averaging resulted in fewer false-negative findings than count aggregation (Figure 5A). Weighted averaging and count aggregation produced similar results in this simulation, probably because count aggregation worked well for lower-expression genes used in it (Figure 1B). Both weighted and unweighted *t* tests for relative counts still had low false-positive findings (Figure 5B). DESeq2 analysis of aggregate counts[39,43]—the Seurat's default and widely used PB method—had the largest number of false-negative (Figure 5A) and false-positive (Figure 5B) findings. As expected,[1] this PB analysis performed better than the Wilcoxon test (cf., Figure 4), yet it still had too many false findings (e.g., 1% of ~20,000 genes detected by modern assays would result in ~200 false-positive findings). Moreover, its up to 50 orders of

magnitude lower p values compared with the weighted t test indicated severe variance underestimation (Figures 5C and 5D). When we simulated DGE ideally matching DESeq2 assumptions (equal number of 150 upregulated and 150 downregulated genes), the false-positive findings disappeared, yet the false-negative findings remained high (Figure S2).

Weighted statistical testing reduces false-positive findings in spatial transcriptomics

The approach to DGE analysis described above can also be used for spRNA-seq applications after several small modifications. The spRNA-seq data are a matrix $N_{Z,i}$ of gene Z counts in a spatial bin i. In some assays (e.g., Visium HD from 10× Genomics), the spatial bins are fixed areas like squares. In other assays (e.g., Xenium from 10× Genomics), the bins are segmented areas designed to represent individual cells. From the perspective of DGE analysis, the main differences between scRNA-seq and spRNA-seq are: (a) significantly fewer transcript counts per bin, (b) overlaps of different cells within the same bin (even after cell-based segmentation), and (c) assay noise correlations between adjacent bins or cells due to RNA probe diffusion.

Because of these differences, the aggregated count approximation with $w_{Z,i} = w_{AC} = N_i/\sum_i N_i$ ($\rho_Z = 0$) may be the best one for the currently available high-resolution assays (STAR Methods). In full transcriptome assays like Visium HD, log_2(FC) can then be calculated from the corresponding weighted averages, and p values can be determined from the weighted t test-χ^2 test combination. These log_2(FC) and p values must then be interpreted as the difference in aggregated, rather than individual cell, gene expression, which is the downside of the aggregated count approximation. An alternative solution—enabling one to account for biological variability between cells of the same type within each sample—is to define the sample as a group of multiple cell clusters with each cluster containing multiple bins (see STAR Methods).

For instance, in a Visium HD assay, mRNA within a tissue section is hybridized with gene-barcoded probes, which are then transferred onto a slide coated with UMI and spatially barcoded primers. The spatial barcode distinguishes 2 × 2 μm squares on the slide surface. After probe hybridization with the primers, a cDNA library is constructed, amplified by PCR, sequenced by next-generation sequencing (NGS) like in scRNA-seq, and converted into an $N_{Z,i}$ matrix of gene Z counts in spatial bin i. A user may select the number of 2 × 2 μm squares included in each spatial bin. Our testing suggests 10–15 μm smearing of the probes due to diffusion when they are transferred from tissue section onto a Visium HD slide. We, therefore, agree with the 10× Genomics recommendation of using 8 × 8 μm^2 binning.

The $N_{Z,i}$ data matrix generated by Visium HD can be analyzed in the same way as in scRNA-seq. The values of log_2(FC) and p comparing two groups of bins can be calculated by setting $\rho_Z = 0$ and combining the weighted t test and χ^2 test. In this case, the p values predicted by the weighted t test and χ^2 test may be similar. At lower counts per bin/cell, the variance associated with random transcript sampling for NGS may be dominant, in which case the weighted t test and χ^2 test p values are expected to be close to each other (STAR Methods). We recommend using the more conservative p value estimate from the two tests. Groups of bins preferentially or exclusively containing the cell type(s) of interest may be selected for such comparison, as illustrated in Figure 6A for proliferating chondrocytes in the growth plate. Alternatively, bins may be grouped or clustered based on gene expression patterns. Analysis of data with multiple samples per genotype or treatment again follows the protocol described above.

Note that accounting for statistical weight is even more important for spRNA-seq than for scRNA-seq. Large (∼100-fold) variations in the statical weight of bins cannot be eliminated, even in a hypothetical perfect spatial assay, because of large lateral variations in RNA density in tissue sections, which are biological rather than technical. Figure 6B illustrates their effects based on comparing proliferating (and possibly some pre-hypertrophic) chondrocytes in 3 sections from the same growth plate. Our approach identified differential expression only in several extracellular matrix genes, expression of which may vary locally within the same growth plate. In contrast, the analysis based on Seurat's FindMarkers function identified many additional genes not expected to be differentially expressed within the same growth plate. It was the attempt to understand this failure of Seurat's traditional analysis that initiated this whole study.

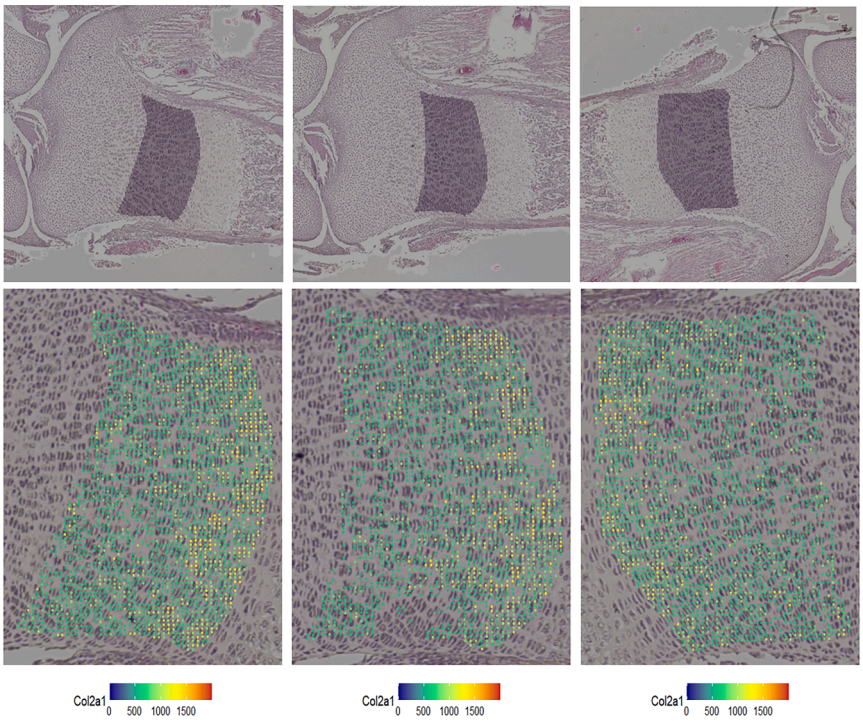
Importantly, analysis of imaging rather than sequencing-based spatial transcriptomics assays (e.g., Xenium) requires a different approach to data normalization. These assays image signals from probes for a limited, preselected subset of genes. They also cannot identify more that several total transcripts per μm^2 because of limited optical resolution. Moreover, ratios of counts between different gene transcripts appear to change from experiment to experiment (at least in our experience of using the Xenium assay). Together, these assay features preclude relative count normalization and utilization of statistical approaches to data analysis described above. We have found several alternative solutions, but their discussion is beyond the scope of this paper.

DISCUSSION

The key features distinguishing our approach to scRNA- and spRNA-seq analysis from commonly used ones are: (1) No assumptions about noise statistics besides randomness. (2) Model-free noise averaging, rather than model-dependent noise regression. (3) Unbiased accounting for technical noise variations from cell to cell and sample to sample based on statistical weights deduced directly from the data.

Fewer assumptions inherently produce less biased outcome because data may not meet the assumptions. Using statistical weights reduces uncertainty in whether a gene is downregulated or upregulated (Figure 2). Weighted statistical tests reduce the likelihood of interpreting random noise and variations in sequencing depth as DGE (Figure 3). Together, accurate weighted averaging and statistical tests remove noise artifacts and detect DGE signals more reliably than data parametrization, rescaling, or noise regression (Figures 3 and 4). They produce more consistent results when data from multiple samples are available as well (Figure 5). Besides, the same approach with only minor adjustments produces more meaningful results for high-resolution spatial transcriptomics (Figure 6).

A



B

Differentially expressed genes in all 3 pairwise comparisons ($p < 0.05$)						
Seurat's FindMarkers (default Wilcoxon test)						wght-t/χ^2-test
Cd63	*Ckap4*	*Col1a2*	*Col2a1*	*Comp*	*Fmod*	*Col1a1*
Hmgcs1	*Lrp6*	*Matn1*	*Matn3*	*mt-Atp8*	*mt-Nd2*	*Col2a1*
Mydgf	*Nrk*	*Pcolce2*	*Pofut2*	*Prelp*	*Rack1*	*Comp*
Rlim	*Sec22b*	*Slc26a2*	*Sox9*	*Stmn1*	*Tgfbi*	*Matn1*
Ybx1	*Ywhag*					*Prelp*

Figure 6. Weighted statistical testing reduces false-positive DGE in spRNA-seq (Visium HD assay)

(A) Top: Proliferating chondrocytes were selected in 3 sections from the same E18.5 tibia; shaded areas show the selected regions in high resolution images of H&E-stained sections. Bottom: Zoomed images of the selected areas. Colored dots mark 8 × 8 μm^2 bins used for the analysis of gene expression, which passed quality control based on the following criteria: >20 features per bin, >100 counts other than *Col1a1* or *Col1a2* per bin, mitochondrial genes between 0.1% and 5% UMI, and *Col2a1* > 5% UMI. The bin color represents the relative expression of *Col2a1* ($n_{Col2a1,i}$×10,000).

(B) DGE analysis performed with the weighted t test-χ^2 test combination and Seurat's FindMarkers function for all 3 possible pairwise comparisons between the selected areas. Only genes detected in at least 10% of the bins were analyzed.

To be beneficial, additional assumptions must accurately describe the assay noise. This does not seem to be the case for common scRNA-seq analysis tools, as indicated in Figure 1B and increased false findings by these tools (Figures 3, 4, 5, and 6). One explanation is the noise associated with the cell-to-cell variation in the sequencing depth (counts per cell, Figure 1A), which may be impossible to model (STAR Methods). Consider, e.g., the SCT rescaling of transcript counts designed to correct the data for this variation.[7] Figure 3F shows a strong signal introduced by the SCT into the data. This signal is, at least, partially artificial rather than unmasked from the noise, which is indicated by the highly significant DGE between the original and randomly subsampled data after the SCT (Figures 3G and S1H). The SCT artifacts are more pronounced for lower-expression genes because of stronger sequencing depth effects on them. Together, Figures 3F, 3G, and S1H show that model-dependent tools like the SCT are not universally appropriate for scRNA-seq. Therefore, we think that such tools should be avoided or at least tested for each dataset. Random subsampling testing is essential because of the expected sequencing depth variation between samples.

While developing our minimal assumptions approach, we have made several observations that are worth keeping in mind when interpreting the results produced by our or any other DGE analysis.

(1) Statistical weights of cells are not just mathematical constructions we used for data analysis; they are measures of gene count accuracy in each cell. Unweighted averaging may alter the apparent change in gene expression by more than a factor of 2 and affect even the sign of the change (Figure 2A). While providing reasonable control of false-positive findings, unweighted t test for relative counts is too conservative and has less power in detecting DGE than the weighted t test (Figure S5). Statistical weights are, therefore, important for data interpretation.[28,29,31,32]

(2) Violin plots of normalized transcript counts are visually appealing yet may be misleading, even when one looks at the data points rather than the violins themselves. More graphically accurate representation of the same data by bona fide histograms may be equally misleading because histograms of normalized data do not account for the statistical weights of cells. A peak due to cells with low confidence count values can be easily misinterpreted. For instance, violin plots or bona fide histograms may look very different for samples with identical gene expression when the sequencing depth is different (Figure S6). Similarly, it is possible to imagine the situation when the average gene expression is different in two samples, while the corresponding violin plots look the same. In contrast, histograms for unnormalized total transcript counts per cell faithfully represent the reality (Figure 1A)

because the statistical weights of cells are not important for their interpretation.

(3) Analysis of DGE between cells selected based on cell clustering algorithms is questionable, unlike DGE analysis between cells selected based on the known marker gene expression. This concern is recognized in Seurat documentation (FindMarkers function). However, being often ignored in practical applications, it may be an important contributing factor to false-positive and false-negative DGE results. All statistical models for *p* value calculation presume random, independent sampling. All common cell-clustering algorithms group cells based on differences in the expression of thousands of highly variable genes (e.g., 2,000–3,000 genes in Seurat). The *p* values calculated for these genes (and genes dependent on them) are questionable because cell clustering based on the expression of these genes is inconsistent with independent sampling. Clustering based on several marker genes like in Figure 1A and excluding these marker genes from DGE analysis resolve this concern. Strategies for identifying such marker genes when they are not known *a priori* are beyond the scope of the present work.

(4) Analysis of the FDR due to multiple comparisons in scRNA-seq is important[44] yet may be misused. Proper FDR analysis is question and context dependent. FDR is the fraction of genes with *p* values below a predetermined threshold that is expected to be false positive by chance upon multiple comparisons.[44,45] It is different for a specific group of genes, a pathway, or a list of candidate genes for follow up. However, the FDR is often calculated for the full transcriptome, which is not always appropriate. Validating genes selected based on uncorrected *p* values by additional or independent experiments reduces the risk of missing important genes by chance due to an FDR threshold, although it increases the study cost. With all this in mind, our scripts do not calculate corrected *p* values, which can be easily determined, e.g., by using the p.adjust() function in R, as described in our GitHub tutorial.

In conclusion, more accurate scRNA- and spRNA-seq data analysis with fewer assumptions will hopefully simplify identification of upregulated/downregulated genes and eventually reduce or eliminate the need for follow up. In the meantime, our findings suggest caution in interpreting unvalidated output of common DGE analysis approaches (Figures 2, 3, 4, 5, and 6). Because we have used these approaches before, we reanalyzed our past experiments and learned useful lessons. Revisiting such data may be worth the invested time.

Limitations of the study

The goals of this study are to show that no assumptions beyond random and independent sampling are necessary for scRNA- and spRNA-seq analysis and to provide the corresponding tools. To demonstrate that additional, unnecessary assumptions lead to analysis bias and resulting artifacts, we compared our tools with those most widely used in Seurat. However, many other tools are available in numerous packages like edgeR, Scanpy, scVI, etc., which we have not formally tested here. Most are based on the same assumptions and similar algorithms to those implemented in Seurat and are expected to perform like Seurat tools. Some, however, also employ more complex computational data transformation models (including generative AI), making the underlying model assumptions less transparent. Assumption-related analysis bias in the latter models may depend on the data being analyzed and is difficult to examine at the conceptual level of the present study. Therefore, we invite interested readers to examine potential biases of their favorite tools by applying the tests we described to the specific data.

To be accurate, our minimal-assumption statistics requires large datasets with many cells per cell group of interest and thousands of total UMIs per cell (needed for ICC estimates, STAR Methods). Such datasets are routine in modern scRNA-seq and are the subject of this study, yet there are smaller ones. Therefore, our DGE analysis functions incorporate and report multiple quality control parameters. A tutorial and function documentation provided at https://github.com/sergeyleikin/sc-sp-RNASeq describe how to use these parameters to ensure accurate ICC estimates and what can be done when the data do not meet the requirements (e.g., samples with too few cells and/or transcripts of interest). The tutorial and function documentation also discuss the quality control and what can be done when it fails in spRNA-seq.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Sergey Leikin (leikins@mail.nih.gov).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Spatial data have been deposited at NCBI GEO as https://www.ncbi.nlm.nih.gov/geo: GSE299816 and are publicly available.
- Additional data and simulations not included in this publication have been deposited at Synapse.org and are publicly available as of the date of publication at https://doi.org/10.7303/syn66520062.
- This paper also analyzes existing, publicly available data, accessible at 10× Genomics: https://www.10xgenomics.com/datasets/10-k-mouse-e-18-combined-cortex-hippocampus-and-subventricular-zone-cells-single-indexed-3-1-standard-4-0-0 and https://www.10xgenomics.com/datasets/10-k-mouse-e-18-combined-cortex-hippocampus-and-subventricular-zone-cells-dual-indexed-3-1-standard-4-0-0.
- All original code has been deposited at GitHub (https://github.com/sergeyleikin/sc-sp-RNASeq) and is publicly available at https://doi.org/10.5281/zenodo.19041480.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This work was supported by the Intramural Research Program of NICHD. The Visium HD assay was performed with the assistance of Drs. James Iben and Fabio Rueda Faucz at the Molecular Genomics Core of NICHD. The authors thank Drs. Ryan Dale, Patrick Fletcher, Edward Mertz, and Shyamal Peddada for many useful discussions. A blog post by Jan Vanhove (https://janhove.github.io/posts/2016-02-16-cluster-randomisation-correction/) helped G.M. to realize the connection between our work and theory of cluster-randomized experiments.

AUTHOR CONTRIBUTIONS

Conceptualization, G.M. and S.L.; methodology, G.M. and S.L.; investigation, G.M., A.T., and S.L.; writing – original draft, G.M. and S.L.; writing – review & editing, G.M. and S.L.; funding acquisition, S.L.; resources, S.L.; supervision, S.L.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Animals
  - Visium HD experiment
- METHOD DETAILS
  - Mouse brain scRNASeq datasets
  - Simulated random noise and gene upregulation
  - Computational benchmarking
- METHOD DETAILS: THEORY
  - Sampling and technical noise in scRNASeq and spRNASeq
  - Average and variance calculation
  - Weighted *t* test
  - Violin plot interpretation
  - Average and variance calculation in multiple sample experiments
  - Chi-squared test and its combination with weighted *t* test
  - Sampling and technical noise in spRNASeq
- METHOD DETAILS: SUPPLEMENTAL DERIVATIONS
  - Derivation of Equation 4
  - Derivation of Equation 8
  - Model of colored balls in a basket
  - Binomial trials with variable outcome probability
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2026.101383.

Received: July 28, 2025
Revised: February 9, 2026
Accepted: March 17, 2026
Published: April 13, 2026

REFERENCES

1. Squair, J.W., Gautier, M., Kathe, C., Anderson, M.A., James, N.D., Hutson, T.H., Hudelle, R., Qaiser, T., Matson, K.J.E., Barraud, Q., et al. (2021). Confronting false discoveries in single-cell differential expression. Nat. Commun. *12*, 5692. https://doi.org/10.1038/s41467-021-25960-2.
2. Murphy, A.E., Fancy, N., and Skene, N. (2023). Avoiding false discoveries in single-cell RNA-seq by revisiting the first Alzheimer's disease dataset. eLife *12*. https://doi.org/10.7554/eLife.90214.
3. Hao, Y., Stuart, T., Kowalski, M.H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., et al. (2024). Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat. Biotechnol. *42*, 293–304. https://doi.org/10.1038/s41587-023-01767-y.
4. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888–1902.e21. https://doi.org/10.1016/j.cell.2019.05.031.
5. Chen, Y., Chen, L., Lun, A.T.L., Baldoni, P.L., and Smyth, G.K. (2025). edgeR v4: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. Nucleic Acids Res. *53*, gkaf018. https://doi.org/10.1093/nar/gkaf018.
6. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15. https://doi.org/10.1186/s13059-017-1382-0.
7. Hafemeister, C., and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. *20*, 296. https://doi.org/10.1186/s13059-019-1874-1.
8. Wu, C.H., Zhou, X., and Chen, M. (2025). Exploring and mitigating shortcomings in single-cell differential expression analysis with a new statistical paradigm. Genome Biol. *26*, 58. https://doi.org/10.1186/s13059-025-03525-6.
9. Tzec-Interián, J.A., González-Padilla, D., and Góngora-Castillo, E.B. (2025). Bioinformatics perspectives on transcriptomics: A comprehensive review of bulk and single-cell RNA sequencing analyses. Quant Biol *13*, ARTN.e78. https://doi.org/10.1002/qub2.78.
10. Gatlin, V., Gupta, S., Romero, S., Chapkin, R.S., and Cai, J.J. (2025). Exploring cell-to-cell variability and functional insights through differentially variable gene analysis. NPJ Syst. Biol. Appl. *11*, 29. https://doi.org/10.1038/s41540-025-00507-z.
11. Kim, M.C., Gate, R., Lee, D.S., Tolopko, A., Lu, A., Gordon, E., Shifrut, E., Garcia-Nieto, P.E., Marson, A., Ntranos, V., et al. (2024). Method of moments framework for differential expression analysis of single-cell RNA sequencing data. Cell *187*, 6393–6410.e16. https://doi.org/10.1016/j.cell.2024.09.044.
12. Silkwood, K., Dollinger, E., Gervin, J., Atwood, S., Nie, Q., and Lander, A.D. (2024). Leveraging gene correlations in single cell transcriptomic data. BMC Bioinf. *25*, 305. https://doi.org/10.1186/s12859-024-05926-z.
13. Lee, H., and Han, B. (2024). Pseudobulk with proper offsets has the same statistical properties as generalized linear mixed models in single-cell case-control studies. Bioinformatics *40*, btae498. https://doi.org/10.1093/bioinformatics/btae498.
14. Yan, J., Zeng, Q., and Wang, X. (2024). RankCompV3: a differential expression analysis algorithm based on relative expression orderings and applications in single-cell RNA transcriptomics. BMC Bioinf. *25*, 259. https://doi.org/10.1186/s12859-024-05889-1.
15. Missarova, A., Dann, E., Rosen, L., Satija, R., and Marioni, J. (2024). Leveraging neighborhood representations of single-cell data to achieve sensitive DE testing with miloDE. Genome Biol. *25*, 189. https://doi.org/10.1186/s13059-024-03334-3.
16. Han, G., Yan, D., Sun, Z., Fang, J., Chang, X., Wilson, L., and Liu, Y. (2024). Bayesian-frequentist hybrid inference framework for single cell RNA-seq analyses. Hum. Genomics *18*, 69. https://doi.org/10.1186/s40246-024-00638-0.
17. Ospina, O.E., Soupir, A.C., Manjarres-Betancur, R., Gonzalez-Calderon, G., Yu, X., and Fridley, B.L. (2024). Differential gene expression analysis of spatial transcriptomic experiments using spatial mixed models. Sci. Rep. *14*, 10967. https://doi.org/10.1038/s41598-024-61758-0.
18. Mason, K., Sathe, A., Hess, P.R., Rong, J., Wu, C.Y., Furth, E., Susztak, K., Levinsohn, J., Ji, H.P., and Zhang, N. (2024). Niche-DE: niche-differential gene expression analysis in spatial transcriptomics data identifies context-dependent cell-cell interactions. Genome Biol. *25*, 14. https://doi.org/10.1186/s13059-023-03159-6.
19. Cuevas-Diaz Duran, R., Wei, H., and Wu, J. (2024). Data normalization for addressing the challenges in the analysis of single-cell transcriptomic datasets. BMC Genom. *25*, 444. https://doi.org/10.1186/s12864-024-10364-5.

20. Pullin, J.M., and McCarthy, D.J. (2024). A comparison of marker gene selection methods for single-cell RNA sequencing data. Genome Biol. *25*, 56. https://doi.org/10.1186/s13059-024-03183-0.

21. Liu, Y., Zhao, J., Adams, T.S., Wang, N., Schupp, J.C., Wu, W., McDonough, J.E., Chupp, G.L., Kaminski, N., Wang, Z., et al. (2023). iDESC: identifying differential expression in single-cell RNA sequencing data with multiple subjects. BMC Bioinf. *24*, 318. https://doi.org/10.1186/s12859-023-05432-8.

22. Cui, T., and Wang, T. (2023). A comprehensive assessment of hurdle and zero-inflated models for single cell RNA-sequencing analysis. Brief. Bioinform. *24*, bbad272. https://doi.org/10.1093/bib/bbad272.

23. Liu, Y., Huang, J., Pandey, R., Liu, P., Therani, B., Qiu, Q., Rao, S., Geurts, A.M., Cowley, A.W., Jr., Greene, A.S., et al. (2023). Robustness of single-cell RNA-seq for identifying differentially expressed genes. BMC Genom. *24*, 371. https://doi.org/10.1186/s12864-023-09487-y.

24. Gagnon, J., Pi, L., Ryals, M., Wan, Q., Hu, W., Ouyang, Z., Zhang, B., and Li, K. (2022). Recommendations of scRNA-seq Differential Gene Expression Analysis Based on Comprehensive Benchmarking. Life *12*, 850. https://doi.org/10.3390/life12060850.

25. Shi, Y., Lee, J.H., Kang, H., and Jiang, H. (2022). A Two-Part Mixed Model for Differential Expression Analysis in Single-Cell High-Throughput Gene Expression Data. Genes *13*, 377. https://doi.org/10.3390/genes13020377.

26. Bouland, G.A., Mahfouz, A., and Reinders, M.J.T. (2021). Differential analysis of binarized single-cell RNA sequencing data captures biological variation. NAR Genom. Bioinform. *3*, lqab118. https://doi.org/10.1093/nargab/lqab118.

27. Li, H., Zhu, B., Xu, Z., Adams, T., Kaminski, N., and Zhao, H. (2021). A Markov random field model for network-based differential expression analysis of single-cell RNA-seq data. BMC Bioinf. *22*, 524. https://doi.org/10.1186/s12859-021-04412-0.

28. Fleiss, J.L. (1999). The Design and Analysis of Clinical Experiments (John Wiley & Sons).

29. Hayes, R.J., and Moulton, L.H. (2022). Cluster Randomized Trials (CRC Press).

30. Ridout, M.S., Demétrio, C.G., and Firth, D. (1999). Estimating intraclass correlation for binary data. Biometrics *55*, 137–148. https://doi.org/10.1111/j.0006-341x.1999.00137.x.

31. Hartung, J., Knapp, G., and Sinha, B. (2008). Statistical Meta-Analysis with Applications (Wiley-Interscience).

32. Bevington, P.R., and Robinson, D.K. (2002). Data Reduction and Error Analysis for the Physical Sciences, Third Edition Edition (McGraw-Hill).

33. Ramachandran, K.M., and Tsokos, C.P. (2021). Mathematical Statistics with Applications in R, Third Edition Edition (Academic Press).

34. Jiang, R., Sun, T., Song, D., and Li, J.J. (2022). Statistics or biology: the zero-inflation controversy about scRNA-seq data. Genome Biol. *23*, 31. https://doi.org/10.1186/s13059-022-02601-5.

35. Silverman, J.D., Roche, K., Mukherjee, S., and David, L.A. (2020). Naught all zeros in sequence count data are the same. Comput. Struct. Biotechnol. J. *18*, 2789–2798. https://doi.org/10.1016/j.csbj.2020.09.014.

36. Zhao, P., Zhu, J., Ma, Y., and Zhou, X. (2022). Modeling zero inflation is not necessary for spatial transcriptomics. Genome Biol. *23*, 118. https://doi.org/10.1186/s13059-022-02684-0.

37. Svensson, V. (2020). Droplet scRNA-seq is not zero-inflated. Nat. Biotechnol. *38*, 147–150. https://doi.org/10.1038/s41587-019-0379-5.

38. Lun, A.T.L., and Marioni, J.C. (2017). Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data. Biostatistics *18*, 451–464. https://doi.org/10.1093/biostatistics/kxw055.

39. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.

40. Rasch, D., Teuscher, F., and Guiard, V. (2007). How robust are tests for two independent samples? J Stat Plan Infer *137*, 2706–2720. https://doi.org/10.1016/j.jspi.2006.04.011.

41. Hoffman, G.E., Lee, D., Bendl, J., Prashant, N.M., Hong, A., Casey, C., Alvia, M., Shao, Z., Argyriou, S., Therrien, K., et al. (2023). Efficient differential expression analysis of large-scale single cell transcriptomics data using dreamlet. Preprint at bioRxiv. https://doi.org/10.1101/2023.03.17.533005.

42. Hoffman, G.E., and Roussos, P. (2021). Dream: powerful differential expression analysis for repeated measures designs. Bioinformatics *37*, 192–201. https://doi.org/10.1093/bioinformatics/btaa687.

43. Amezquita, R.A., Lun, A.T.L., Becht, E., Carey, V.J., Carpp, L.N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C., et al. (2020). Orchestrating single-cell analysis with Bioconductor. Nat. Methods *17*, 137–145. https://doi.org/10.1038/s41592-019-0654-x.

44. Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. J R Stat Soc B *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

45. Benjamini, Y., and Cohen, R. (2017). Weighted false discovery rate controlling procedures for clinical trials. Biostatistics *18*, 91–104. https://doi.org/10.1093/biostatistics/kxw030.

46. Lause, J., Berens, P., and Kobak, D. (2021). Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. Genome Biol. *22*, 258. https://doi.org/10.1186/s13059-021-02451-7.

47. Townes, F.W., Hicks, S.C., Aryee, M.J., and Irizarry, R.A. (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. Genome Biol. *20*, 295. https://doi.org/10.1186/s13059-019-1861-6.

48. Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. *17*, 333–351. https://doi.org/10.1038/nrg.2016.49.

49. Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. Nat. Methods *11*, 163–166. https://doi.org/10.1038/nmeth.2772.

50. Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat. Rev. Genet. *14*, 618–630. https://doi.org/10.1038/nrg3542.

51. Feynman, R.P. (1998). Statistical Mechanics: A Set of Lectures (CRC Press)).

52. Blitzstein, J.K., and Hwang, J. (2015). Introduction to Probability (CRC Press).

53. Cochran, W.G. (1939). The use of the analysis of variance in enumeration by sampling. J Am Stat Ass *34*, 492–510.

54. Elston, R.C., Hill, W.G., and Smith, C. (1977). Query: estimating "heritability" of a dichotomous trait. Biometrics *33*, 231–236.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Mouse E18.5 tibia	this paper	NCBI GEO: GSE299816
Mouse E18 cortex cells (hippocampus and subventricular zone)	10X Genomics	https://www.10xgenomics.com/datasets/10-k-mouse-e-18-combined-cortex-hippocampus-and-subventricular-zone-cells-single-indexed-3-1-standard-4-0-0
Mouse E18 cortex cells (hippocampus and subventricular zone)	10X Genomics	https://www.10xgenomics.com/datasets/10-k-mouse-e-18-combined-cortex-hippocampus-and-subventricular-zone-cells-dual-indexed-3-1-standard-4-0-0
Experimental models: Organisms/strains		
C57BL/6J mice maintained on C57BL/6J background	Jackson Laboratory	stock #000664; RRID:IMSR_JAX:000664
Critical commercial assays		
Visium HD	10X Genomics	cat #1000676
Software and algorithms		
R functions for DGE analysis in scRNASeq and spRNASeq	this paper	https://github.com/sergeyleikin/sc-sp-RNASeq https://doi.org/10.5281/zenodo.19041480
R-Studio	Posit Software	v. 2024.12.0 Build 467; RRID:SCR_000432
Seurat	Satja Lab (https://satijalab.org/seurat/)	v. 5.2.0; RRID:SCR_007322
SigmaPlot	Systat Software	v. 13.0; RRID:SCR_003210

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Animals

C57BL/6J (stock #000664) mice were purchased from the Jackson Laboratory (Bar Harbor, ME, USA) and maintained on the C57BL/6J background. Animal care and experiments were performed in accordance with a protocol approved by the Eunice Kennedy Shriver National Institute of Child Health and Human Development Animal Care and Use Committee.

Visium HD experiment

Tibias from mouse embryo at embryonic day E18.5 were dissected, fixed in freshly prepared 4% methanol-free formaldehyde solution in phosphate-buffered saline for 24 h at room temperature, and embedded into paraffin. Three sections from the same tibia were placed on the same slide within 6.5 × 6.5 mm area, stained with Hematoxylin and Eosin, imaged on a Nikon Ti2-E inverted microscope with a 20X/0.8NA objective and analyzed with the Visium HD assay (10X Genomics) as described by the manufacturer. The sequencing statistics was: 356,837,522 reads, 1080 mean reads per 8 × 8 μm bin, 89.3% valid barcodes, 99.9% valid UMIs, 28.5% sequencing saturation. A low-resolution image of the sequenced area captured by the CytAssist instrument (10X Genomics) was replaced with the high-resolution Nikon Ti2-E image as described by the manufacturer using Loupe Browser software (10X Genomics). The data were processed with SpaceRanger v.3.0.1 software (10X Genomics). Proliferating chondrocyte areas in the proximal growth plates of the 3 tibias were selected in the Loupe Browser software. The barcodes of 8 × 8 μm bins within these areas were exported from Loupe Browser for further data processing in R (see sampling and technical noise in spRNASeq).

METHOD DETAILS

Mouse brain scRNASeq datasets

scRNASeq datasets from E18 mouse brain were downloaded from the 10X Genomics (see key resources table). Filtered count matrices from the two datasets (filtered features.h5 files) were imported into Seurat objects and merged without renormalization to increase the number of neurons in subsequent DGE analysis. GABAergic neurons were selected based on nonzero counts of *Gad1*, *Gad2*, and *Slc6a1*. Glutamatergic neurons were selected based on nonzero counts of *Gls*, *Grin2b*, and *Slc17a7*.

Simulated random noise and gene upregulation

To test effects of well-defined noise and changes in gene expression, the $N_{Z,i}$ matrix of glutamatergic neuron gene counts was modified by either random noise [$N_{Z,i}$+$f_n(\alpha,N_{Z,i})$] or noise and expression up(down)regulation [$N_{Z,i}$+$f_n(\alpha,N_{Z,i})$+$f_{ex}(Z,i)$]. The noise function $f_n(\alpha,N_{Z,i})$,

$$f_n(\alpha, N_{Z,i}) = \text{round}\left(\text{runif}(1, min = -1, max = 1) \cdot sf \cdot \sqrt{N_{Z,i}}\right)$$

produces integer values uniformly distributed between $-sf\cdot\sqrt{N_{Z,i}}$ and $sf\cdot\sqrt{N_{Z,i}}$. The scaling factor

$$sf = \left(\sqrt{3\alpha} - \frac{\sqrt{3\alpha} - 1}{1 + 0.1(N_{Z,i} - 1)}\right)$$

is the simplest expression ensuring $Var[f_n(\alpha,N_{Z,i}\rightarrow\infty)] = \alpha N_{Z,i}$ (reason behind the $\sqrt{3\alpha}$ factor), $f_n(\alpha,0) = 0$, $f_n(\alpha,N_{Z,i})\geq 0$ at $\alpha < 10$, and maximum noise amplitude ($sf\sqrt{N_{Z,i}}$) between $0.5N_{Z,i}$ and $N_{Z,i}$ for moderate expression genes with average 0.3 to 3 counts per cell. For the latter genes, simulation results were qualitatively similar at all $\alpha < 10$. The effects of noise weakly increased with α, and we selected mid-range $\alpha = 5$ for presenting the results. The results were affected only by the relative noise amplitude but not the specific form of its dependence on $N_{Z,i}$.

The expression up(down)regulation function

$$f_{ex}(Z, i) = \text{round}(\beta N_{Z,i})$$

produces integers proportional to the original transcript count $N_{Z,i}$, e.g., $\beta = 1.5$ is used to model 50% upregulation of gene expression.

Computational benchmarking

The goal of this paper is to demonstrate major artifacts caused by unneeded assumptions in commonly used scRNASeq and spRNA-Seq data analysis methods and show that a minimal-assumptions approach akin to that in cluster-randomized clinical trials largely eliminates such artifacts. To facilitate its understanding and testing, we provide the source code for functions implementing this minimal-assumptions approach in R programming language for Seurat data objects. The code is designed to ease understanding of the functions rather than to increase their computation speed. For instance, it performs all calculations sequentially in a loop for individual genes instead of vectorizing them. Much faster vectorized calculations could be easily accomplished for ANOVA-ICC-based analysis, yet this would be more challenging for iterative-ICC-based analysis. Thus, it makes more sense to first test multiple datasets for whether both ICC estimators produce similar results (as we suspect) or whether one of them performs better (which is possible). Such testing is beyond the scope of the present paper.

Nonetheless, we found the functions we provide to be fast enough for practical analysis of any datasets on a standard Dell Precision 7680 laptop workstation, despite the heavy burden of the IT security software running on the background (Figure S3). In our hands, the DGE calculations by these functions were never even close to being the bottleneck of the overall data analysis. Much more time was consumed by standard preprocessing of raw sequencing data, data quality analysis, and assembly and annotation of Seurat objects. Typical analysis of DGE between two groups of several hundred cells was completed within a few seconds. More challenging DGE analysis of ~12,000 genes between ~1,100 glutamatergic neurons and ~1,100 GABAergic neurons was completed in less than 1 min (Figure S3A). DGE analysis of ~20,000 genes for such neurons in 3 different datasets (~2,500 total neurons) took ~30 s (Figure S3B). Even the most challenging analysis of DGE between two sets of data containing 3 samples each, ~20,000 genes, and ~400,000 cells took ~1 h (Figure S3C). For comparison, the time to prepare for the latter analysis after the tissue samples were collected was ~3 weeks for sample processing and sequencing assay plus ~1 week for data processing to assemble the quality-assured, annotated Seurat object.

For completeness, we nevertheless compared the performance of our functions to the common Seurat workflow. When comparing just two groups of cells, our functions were faster than most tests built into the Seurat's FindMarkers function (Figure S3A), including the common limma implementation of the Wilcoxon test (https://rdrr.io/bioc/limma/man/rankSumTestwithCorrelation.html). The only exception was the fast presto implementation of the Wilcoxon test (https://github.com/immunogenomics/presto), although the latter provided different results from the limma implementation. When comparing ~2,500 cells in multiple samples, our functions were noticeably slower than the DESeq2-based pseudo-bulk analysis in Seurat. The difference in performance became much more dramatic when multiple sample analysis was scaled to 400,000 cells (Figure S3C), because the Seurat pseudo-bulk analysis was fully vectorized while our algorithm was not (as discussed above). In addition, the pseudo-bulk analysis simply aggregates counts while we fully analyze each sample, calculating ICC, mean, and variance for all genes.

Overall, we found that the nonoptimized implementation of our minimal-assumptions approach should be fast enough for practical analysis of any existing datasets. Much faster implementation based on vectorizing some of the calculations within our functions is possible, yet the ever-increasing computing power is likely to be sufficient for the ever-increasing data sizes even without it.

METHOD DETAILS: THEORY

Sampling and technical noise in scRNASeq and spRNASeq

The scRNASeq data matrix $N_{Z,i}$ of gene Z counts in cell i is widely interpreted assuming transcript sampling probability being consistent from cell to cell in the original biological sample, e.g., for data parametrization within generalized linear models (GLMs) of count statistics.[7,37,43,46,47] However, these are desired rather than actual data properties.

Consider common droplet-based scRNASeq assays. All of them involve the following basic steps: Step 1. Cell extraction from tissue and dissociation. Step 2. Encapsulation of cells inside droplets containing primers and reagents, one cell per droplet. Step 3. Cells lysis, RNA hybridization with primers, and synthesis of barcoded cDNAs by reverse transcription. A primer-based barcode includes a sequence identifying the droplet and cell inside as well as a sequence unique for each primer (unique molecular identifier, UMI). Step 4. Purification and PCR amplification of cDNAs extracted from droplets. Step 5. Counting of transcripts in an aliquot of the resulting cDNA solution (cDNA library) by next generation sequencing (NGS). Each cDNA is mapped to a gene based on the transcript sequence and to an individual cell based on the droplet barcode. Each unique transcript is counted only once based on its UMI.

The NGS counting can indeed be described by traditional counting statistics, since the probability of each cDNA being counted is approximately the same. However, being remarkably efficient and more than 90% accurate,[48] it is not the main source of technical noise in scRNASeq. The only logical explanation of the variation from $\sim 10^3$ up to $\sim 10^5$ RNA counts per cell (Figure 1A) is that the preceding assay steps produce much larger technical noise. Among these technical noise sources is initial cell isolation and encapsulation into droplets, which alters the composition of transcripts inside cells because cells react to enzymes, disruption of cell-cell contacts, etc. These transcript composition changes are expected to result in significant cell-to-cell variations in the sampling probability for the same gene. These variations are further increased by fluctuations in primer and reagent composition and concentration from droplet to droplet, variable RNA degradation, variable likelihood of hybridization with primers, and variable reverse transcription efficiency. All these noise sources and biological cell variability add up to the observed $\sim$100-fold variation in the transcript sampling probability from cell to cell, which is not likely to be amenable to parametrization within GLMs or other simplified models.

As we show below, this complex physics is better described by calculating the average of relative counts $n_{Z,i} = N_{Z,i}/\sum_Z N_{Z,i}$ and its variance, without any oversimplification and parametrization. We argue that the only assumption needed for determining log_2(FC) and the corresponding p-values in DGE analysis is that all transcript counts are random and independent (as also assumed in other statistical analyses of scRNASeq). Such random, independent sampling of transcripts relies on low fraction of all transcripts in the cell being detected by the current technology. Indeed, a large fraction of live cell transcripts is expected to be degraded/damaged during cell encapsulation into droplets and lysis. Only some of the surviving transcripts are then converted into high quality (identifiable) cDNA, and usually no more than $\sim$50% of the latter are sequenced by NGS at the recommended sequencing depth. When all these effects are combined, the upper estimate for the sampling fraction in scRNASeq is $\sim$1–10%, justifying the assumption of random and independent sampling. Consistently, from 10^4 to $3{\cdot}10^4$ unique transcripts per cell are sequenced in most cells in Figure 1A. A mammalian cell is estimated to have $\sim 10^5$–10^6 transcripts depending on its transcriptional activity.[49,50] GABAergic and glutamatergic neurons are large, highly transcriptionally active cells likely to contain no less than $\sim 10^6$ transcripts.

All scRNASeq analysis methods assume random, independent transcript sampling. Unlike other methods, however, our approach does not utilize additional assumptions. Instead, we rely on the analysis framework developed for studies of cluster-randomized experiments,[28,29,31] since each cell is analogous to a cluster of independent transcript sequencing experiments. We adapt and modify this approach to account for the specific physical properties of scRNASeq assays.

Average and variance calculation

To calculate the average and variance, we use that Equation 1 for $n_{Z,i}$ can be rewritten as

$$n_{Z,i} = \frac{1}{N_i}\sum_{j=1}^{N_i} X_{Z,i,j} \qquad \text{(Equation 4)}$$

where $X_{Z,i,j} = 0,1$ is gene Z count from cell/droplet i and UMI j, $n_{Z,i}$ is the total normalized count of gene Z in cell i and $N_i = \sum_Z N_{Z,i}$ is the total number of unique transcripts (UMIs) detected in cell i (for discussion of failed transcript identifications see Derivation of Equation 4 in Supplemental Derivations).

One may be tempted to interpret $X_{Z,i,j}$ in terms of common counting models, e.g., as a Bernoulli trial, yet this may not be consistent with scRNASeq realities, and this is not what we do. Such models do not account for continuous and rapid change in the number and composition of transcripts, which happens both in a live cell and when the transcripts are being sampled after the cell is lysed. Moreover, even for given total number of transcripts in the cell (N_i^{tot}), total number of gene Z transcripts ($N_{Z,i}^{tot}$), and sampling failure rate (FR_i, fraction of transcripts destroyed during sampling but not identified), the probability density for $N_{Z,i}$ may not be just a function of N_i, N_i^{tot}, $N_{Z,i}^{tot}$, and FR_i (as assumed in all counting models). Instead, it is also likely to depend on transcript distribution in different locations (e.g., nucleus, cytosol, ribosomes), their physical state (e.g., folding, splicing, and interactions), concentration and composition of sequencing reagents, etc. In contrast, a textbook model of sampling of colored balls from a basket either has a constant outcome probability (canonical Bernoulli trials of sampling with replacement)

or a probability determined only by the outcome ($N_{Z,i}$ and N_i), and total numbers of balls ($N_{Z,i}^{tot}$, N_i^{tot}), which is unaffected by the spatial distribution of balls in the basket. We therefore use the approach we think is better justified, which is to assume that the sampling events are random and independent without making any other assumptions. Then, our results are not limited just to the common counting models.

Very large variation in N_i from cell to cell (Figure 1A) indicates large variation in the precision of measuring $n_{Z,i}$ due to the technical noise, in which case the average value of $n_{Z,i}$ in a group of N cells must be calculated as a weighted average[28,29,31,32] as defined by Equation 2, where

$$w_{Z,i} = \frac{1}{Var(n_{Z,i})} \Big/ \sum_{i=1}^{N} \frac{1}{Var(n_{Z,i})} \quad \text{(Equation 5)}$$

is the statistical weight of gene Z transcripts in cell i with N_i transcripts ($\sum_i w_{Z,i} = 1$). $Var(n_{Z,i})$ is the total variance of $n_{Z,i}$,

$$Var(n_{Z,i}) = \langle (n_{Z,i} - E(n_Z|N_i))^2 \rangle \quad \text{(Equation 6)}$$

which must be calculated by averaging over all possible statistical trajectories producing a cell with N_i transcripts. Here $E(n_Z|N_i)$ is the expected value of n_Z in any cell of the selected type that has N_i transcripts. The symbol $\langle\ \rangle$ indicates statistical trajectory averaging (akin to Feynman's path integrals in statistical physics[51]).

Using the law of total variance,[52] we can calculate the variance of $n_{Z,i}$ as

$$Var(n_{Z,i}) = \langle (n_{Z,i} - \langle n_{Z,i} \rangle_i)^2 \rangle + V_Z \quad \text{(Equation 7)}$$

where $\langle n_{Z,i} \rangle_i$ is the expected value of $n_{Z,i}$ (average over statistical trajectories in the cell i) and $V_Z = \langle (\langle n_{Z,i} \rangle_i - E(n_Z|N_i))^2 \rangle$. After substituting Equation 4 into Equation 7 and calculating the sums, we find

$$Var(n_{Z,i}) = \left\langle \frac{\langle n_{Z,i} \rangle_i (1 - \langle n_{Z,i} \rangle_i)}{N_i} \right\rangle + V_Z = \frac{\bar{n}_Z(1 - \bar{n}_Z) + V_Z(N_i - 1)}{N_i} \quad \text{(Equation 8)}$$

where we used that $Var(x) = \langle x^2 \rangle - \langle x \rangle^2$ and $\bar{n}_Z \cong \langle n_{Z,i} \rangle$. Substituting Equation 8 into Equation 5, we find

$$w_{Z,i} = \frac{N_i}{\bar{n}_Z(1 - \bar{n}_Z) + (N_i - 1)V_Z} \Big/ \sum_{i=1}^{N} \frac{N_i}{\bar{n}_Z(1 - \bar{n}_Z) + (N_i - 1)V_Z} \quad \text{(Equation 9)}$$

Note that we derived Equation 8 by directly calculating the sums, assuming only that all sequencing counts $X_{Z,i,j}$ are random and independent (see Derivation of Equation 8, Supplemental Derivations). We did not treat the sequencing reads as Bernoulli trials or utilize any specific statistical distribution. Another instructive way to arrive at Equation 8 would be to treat $X_{Z,i,j}$ as a count of gene Z in a random draw j from the total population of N_i^{tot} transcripts in the cell i. Then $\sum_j X_{Z,i,j}$ is described by the hypergeometric distribution, in which N_i is the total number of random draws (conversion to cDNA, sequencing, and identification of a unique transcript). Even though the value of N_i^{tot} is not known, $Var(n_{Z,i})$ is then described by Equation 8 as long as $N_i \ll N_i^{tot}$, which is the case in typical scRNASeq experiments (see model of colored balls in a basket and binomial trials with variable outcome probability, Supplemental Derivations).

Equation 9 is identical to Equation 3, in which $\rho_Z = V_Z / \bar{n}_Z(1 - \bar{n}_Z)$ is the intracluster correlation coefficient ICC ($0 \leq \rho_Z \leq 1$). Equation 3 is well-known in statistics of cluster-randomized experiments.[28–30,53,54] Previous models of scRNASeq data analysis correspond to $\rho_Z = 1$ ($w_{Z,i} = 1/N$) and $\rho_Z = 0$ ($w_{Z,i} = N_i / \sum_i N_i$). Statistical studies of cluster-randomized experiments showed that neither of these two limits is a good approximation. The importance of accurate ICC calculation has been recognized, and multiple different ICC estimator models have been developed.[30] The most widely used is the ANOVA ICC estimator.[30,53,54] Here we suggest a different approach for estimating ρ_Z directly from experimental data without relying on any specific type of data distribution or sampling. The benefit of this approach is that it makes no assumptions and can also be used for spRNASeq and multiple sample analysis in scRNASeq.

To calculate ρ_Z we use that for inverse variance weights[31,32]

$$Var(\bar{n}_Z) = 1 \Big/ \sum_{i=1}^{N} \left[1 / Var(n_{Z,i}) \right] = \bar{n}_Z(1 - \bar{n}_Z) \Big/ \sum_{i=1}^{N} \left(\frac{N_i}{1 + \rho_Z(N_i - 1)} \right) \quad \text{(Equation 10)}$$

At the same time, $Var(\bar{n}_Z)$ can be estimated directly from the data for given inverse-variance weights $w_{Z,i}$ using the unbiased estimator

$$Var(\bar{n}_Z) = \frac{1}{1 - \sum_{i=1}^{N} w_{Z,i}^2} \sum_{i=1}^{N} w_{Z,i}^2 (n_{Z,i} - \bar{n}_Z)^2 \quad \text{(Equation 11)}$$

Taken together, Equations 3, 10, and 11 allow calculation of $w_{Z,i}$ directly from the data without any assumptions as we show below.

While Equations 3 and 10 are well known in the theory of cluster-randomized experiments, we have not seen Equation 11 being used before. To derive this variance estimator, we used that

$$\left\langle \sum_{i=1}^{N} w_{Z,i}^2 \left(n_{Z,i} - \bar{n}_Z\right)^2 \right\rangle = \sum_{i=1}^{N} w_{Z,i}^2 \langle n_{Z,i}^2 \rangle + \sum_{i=1}^{N} w_{Z,i}^2 \langle \bar{n}_Z^2 \rangle - 2 \sum_{i=1}^{N} w_{Z,i}^2 \langle n_{Z,i} \bar{n}_Z \rangle \quad \text{(Equation 12)}$$

It follows from Equations 9 and 10 that

$$w_{Z,i} Var\left(n_{Z,i}\right) = Var(\bar{n}_Z) \quad \text{(Equation 13)}$$

Therefore,

$$w_{Z,i}^2 \langle n_{Z,i}^2 \rangle = w_{Z,i}^2 Var\left(n_{Z,i}\right) + w_{Z,i}^2 \langle \bar{n}_Z \rangle^2 = w_{Z,i} Var(\bar{n}_Z) + w_{Z,i}^2 \langle \bar{n}_Z \rangle^2 \quad \text{(Equation 14)}$$

$$w_{Z,i}^2 \langle \bar{n}_Z^2 \rangle = w_{Z,i}^2 Var(\bar{n}_Z) + w_{Z,i}^2 \langle \bar{n}_Z \rangle^2 \quad \text{(Equation 15)}$$

and

$$w_{Z,i}^2 \langle n_{Z,i} \bar{n}_Z \rangle = w_{Z,i}^3 Var\left(n_{Z,i}\right) + w_{Z,i}^2 \langle \bar{n}_Z \rangle^2 = w_{Z,i}^2 Var(\bar{n}_Z) + w_{Z,i}^2 \langle \bar{n}_Z \rangle^2 \quad \text{(Equation 16)}$$

After substitution of Equations 13, 14, 15, and 16 into Equation 12 we arrive at

$$Var(\bar{n}_Z) = \frac{1}{\left(1 - \sum_{i=1}^{N} w_{Z,i}^2\right)} \left\langle \sum_{i=1}^{N} w_{Z,i}^2 \left(n_{Z,i} - \bar{n}_Z\right)^2 \right\rangle \quad \text{(Equation 17)}$$

which is approximated by Equation 11 in the limit of large N.

It is worth noting that Equations 11 and 17 are approximations when the weights $w_{Z,i}$ are fitted from the data rather than fixed. However, this is justified when uncertainties of $w_{Z,i}$ are much smaller than variations of $n_{Z,i}$, which should be the case for $w_{Z,i}$ calculated based on averaging over multiple cells as described below.

Combining Equations 10 and 11 we arrive at

$$\bar{n}_Z(1 - \bar{n}_Z) \Big/ \sum_{i=1}^{N} \left(\frac{N_i}{1 + \rho_Z(N_i - 1)}\right) = \frac{1}{1 - \sum_{i=1}^{N} w_{Z,i}^2} \sum_{i=1}^{N} w_{Z,i}^2 \left(n_{Z,i} - \bar{n}_Z\right)^2 \quad \text{(Equation 18)}$$

Using Equation 3 to express $w_{Z,i}$ as a function of ρ_Z and N_i and Equation 5 to express $\bar{n}_Z$ as a function of $w_{Z,i}$ and $n_{Z,i}$, Equation 18 can be solved numerically to find ρ_Z. The values of $\bar{n}_Z$ and $Var(\bar{n}_Z)$ calculated from Equation 18 are in good agreement with those calculated by using the ANOVA ICC estimator (Figures S4A–S4C). Because the latter calculation is computationally faster, we provide it as an option in our scRNASeq data analysis functions (see https://github.com/sergeyleikin/sc-sp-RNASeq). Note that Equation 18 may yield negative values of ρ_Z instead of $\rho_Z \approx 0$ (like ANOVA ICC estimator), which is a known effect of sampling errors that should be corrected by changing negative ρ_Z to $\rho_Z = 0$.[28] Similarly, estimated $\rho_Z > 1$ resulting from sampling errors should be replaced with $\rho_Z = 1$ (although we have not observed estimated $\rho_Z > 1$ in our analysis of experimental data or simulations).

Finding $\bar{n}_Z$ and $Var(\bar{n}_Z)$ through data-driven estimates of $w_{Z,i}$ enables calculation of log_2(FC) and the corresponding p-values with increased accuracy and power. The data analysis reported in the main text of the paper and built into our R scripts is based on this approach. Its only underlying assumptions are that the transcript sampling in scRNASeq is random and independent, and that fluctuations in $\langle n_{Z,i} \rangle$ and N_i across cells are independent of each other, e.g., because of large technical noise. While technical noise might not completely remove correlation between $\langle n_{Z,i} \rangle$ and N_i across cells of different types, here we are mostly concerned with DGE between groups of similar cells. These two assumptions are essential for all unbiased models of scRNASeq analysis anyway. The benefit of our method is that unlike the commonly used approaches it does not use any additional assumptions, simplifications, or data parametrization, while taking full advantage of the data at hand.

Weighted *t* test

A variety of tests for comparing weighted average values in cluster-randomized experiments have been described, partly because of suboptimal weighted *t* test performance when statistical weights are imprecise.[28,29,31] We have found, however, that weighted *t* test performance issues are caused by common usage of the following variance estimator

$$Var(\bar{x}) = \frac{1}{N - 1} \sum_{i=1}^{N} w_{x,i} (x_i - \bar{x})^2 \quad \text{(Equation 19)}$$

for random variables x_i with weights $w_{x,i}$ in experiment clusters i. This expression provides unbiased estimate of $Var(\overline{x})$ at $w_{x,i} = [Var(x_i)]^{-1}/\sum_i [Var(x_i)]^{-1}$.[32] However, it may become biased and inaccurately estimate $Var(\overline{n}_Z)$ when $w_{Z,i}$ do not match the inverse variance of $n_{Z,i}$ (Figure S4D), e.g., resulting in increased false positive findings at $\rho_Z = 0$ (Figures S5A–S5D). Moreover, the test with variance estimator defined by Equation 19 can be anti-conservative even with inverse variance weights (Figures S5A and S5D). Apparently, Equation 19 can have larger uncertainty than Equation 11 and in some cases underestimate the variance, causing small p-values. We are not aware of any rigorous justification or derivation of Equation 19 as the variance estimator.

In contrast, the variance estimator given by Equation 11 remains unbiased and accurate at any $w_{Z,i}$ provided that $\sum_i w_{Z,i}^2 \ll 1$, which is almost always true for scRNASeq and spRNASeq. After replacing Equation 19 with Equation 11 as the variance estimator, we found that the weighted t test exhibited optimal or close to optimal type I error (false positive findings) rate control (Figures S5E and S5J–S5M). Since we are aware only of weighted t test implementations based on Equation 19 we have included an R code for weighted t test based on Equation 11 variance estimator into our package of R functions for scRNASeq and spRNASeq analysis. In addition, this code has two options for calculating degrees of freedom, which may affect p-values in some cases (for more details, see https://github.com/sergeyleikin/sc-sp-RNASeq). It is, however, important to keep in mind that the weighted t test is exact only in the asymptotic limit of large N, i.e., at large number of degrees of freedom. Otherwise, it is just a reasonable approximation.

Analysis of experimental data (Figures 3 and 4) showed that the weighted t test with the Equation 11 variance estimator might provide the best combination of type I error rate control and type II error (false negative) rate control. However, the ground truth in experimental data is not known. Therefore, we performed extensive comparison of all widely used statistical tests for simulated data, in which the ground truth is known (Figures S5F–S5M). We simulated single gene $N_{Z,i}$ counts in scRNASeq by using beta distribution for the count probability, varying the first shape parameter α to mimic changes in gene expression. We used a scaling factor n_{max} in the distribution function for the total counts per cell to mimic changes in the sequencing depth. We tested different versions of this distribution, all of which produced similar results (see additional data at https://doi.org/10.7303/syn66520062).

These simulations confirmed our conclusions based on experimental data analysis. The only two tests with good type I error rate control under all conditions were weighted t test with the Equation 11 variance estimator and unweighted (common) t test, although the unweighted t test was too conservative under some conditions (Figures S5J–S5M). The weighted t test with Equation 19 variance estimator had too many false positive findings (Figures S5D and S5E). Consistent with our analysis of real experimental data in scRNASeq, the Wilcoxon test exhibited poor type I error control for samples with different distributions of N_i, e.g., mimicking effects of sequencing depth variation (Figures S5C, S5K and S5M).

The weighted t test with the Equation 11 variance estimator was also among the best performers in detecting true changes in expression, outperforming the Wilcoxon test (Figures S5F, S5G and S5I) despite the type I error rate control issues in the latter (Figures S5K–S5M), depending on the underlying distributions. The unweighted t test was too conservative and had the worst type II error rate control among all tests.

Overall, the weighted t test with the Equation 11 variance estimator was the clear winner, like in the experimental data analysis (Figures S5F–S5M). It is worth noting that computationally faster weighted t test with ANOVA ICC performed nearly identical to the one based on numerical calculation of ρ_Z from Equation 18. The ANOVA ICC estimator is therefore available as an option in our scRNASeq data analysis functions.

Violin plot interpretation

Using equal statistical weights of different cells ($\rho_Z = 1$) may not only distort the calculated p-values but also cause misinterpretation of the popular violin plots for normalized gene expression (Figure S6). A violin plot is a vertical histogram of cell density distribution vs. normalized gene expression, which is meaningful only when all cells can be assumed to have the same statistical weight. A histogram peak showing only low confidence cells (with low statistical weights) is not an accurate representation of actual gene expression and may be misleading.

To demonstrate misleading violin plots, Figure S6 shows an effect of 40% subsampling of all genes (every count retained with 40% probability or discarded with 60% probability), which corresponds to reduced depth of sequencing in experiments and does not alter gene expression. However, this subsampling significantly alters the violin plots.

Average and variance calculation in multiple sample experiments

In multiple sample experiments, additional averaging over multiple samples must be performed

$$\overline{n}_Z = \sum_{s=1}^{K} \overline{n}_{Z,s} W_{Z,s}, \qquad \text{(Equation 20)}$$

where

$$\overline{n}_{Z,s} = \sum_{i=1}^{N_s} n_{Z,i,s} w_{Z,i,s}, \qquad \text{(Equation 21)}$$

$n_{Z,i,s}$ and $w_{Z,i,s}$ are $n_{Z,i}$ and $w_{Z,i}$ values for samples s = 1,2 … K, N_s is the number of cells in sample s and

$$W_{Z,s} = \frac{1}{Var\left(\bar{n}_{Z,s}\right)} \Big/ \sum_{s=1}^{K} \frac{1}{Var\left(\bar{n}_{Z,s}\right)} \quad \text{(Equation 22)}$$

is the statistical weight of sample s ($\sum_s W_{Z,s} = 1$). Ignoring differences in statistical weights of different samples may be just as problematic as neglecting variations in statistical weights of individual cells.

Equations 20 and 22 are identical to Equations 2 and 5, yet here $Var\left(\bar{n}_{Z,s}\right)$ is given by

$$Var\left(\bar{n}_{Z,s}\right) = \left\langle \left(\bar{n}_{Z,s} - \left\langle \bar{n}_{Z,s} \right\rangle\right)^2 \right\rangle + \mathcal{V}_Z = \frac{1}{1 - \sum_{i=1}^{N_s} w_{Z,i,s}^2} \sum_{i=1}^{N_s} w_{Z,i,s}^2 \left(n_{Z,i,s} - \bar{n}_{Z,s}\right)^2 + \mathcal{V}_Z \quad \text{(Equation 23)}$$

where $\mathcal{V}_Z$ is an analogue of V_Z in Equation 7 and we used Equation 11 for $\left\langle \left(\bar{n}_{Z,s} - \left\langle \bar{n}_{Z,s} \right\rangle\right)^2 \right\rangle$. Using the same logic as for deriving Equation 18, we arrive at

$$\sum_{s=1}^{K} \frac{1 - \sum_{i=1}^{N_s} w_{Z,i,s}^2}{\sum_{i=1}^{N_s} w_{Z,i,s}^2 \left(n_{Z,i,s} - \bar{n}_{Z,s}\right)^2 + \left(1 - \sum_{i=1}^{N_s} w_{Z,i,s}^2\right) \mathcal{V}_Z} = \frac{1 - \sum_{s=1}^{K} W_{Z,s}^2}{\sum_{s=1}^{K} W_{Z,s}^2 \left(\bar{n}_{Z,s} - \bar{n}_Z\right)^2} \quad \text{(Equation 24)}$$

where $n_{Z,i,s}$ are measured, $w_{Z,i,s}$ are calculated by numerically solving Equation 18 or determined from ANOVA ICC for each individual sample s, $\bar{n}_{Z,s}$ are calculated from Equation 21, and $\mathcal{V}_Z$ and $W_{Z,s}$ are calculated by numerically solving Equations 22, 23, and 24. The values of $W_{Z,s}$ can then be used for determining log_2(FC) and the corresponding p-values from the weighted t test. R functions implementing all these calculations for Seurat objects are available for download from https://github.com/sergeyleikin/sc-sp-RNASeq. Testing of the robustness of this approach and its comparison with previously described pseudo-bulk calculations are described in the main text of the paper.

Chi-squared test and its combination with weighted *t* test

The weighted t test with statistical weights calculated from Equation 18 is the minimal assumptions approach to evaluating differential expression between two groups of cells. It is more robust and general than other approaches, yet it is not perfect either. Like previous scRNASeq data analysis models and models of cluster-randomized experiments, it still relies on individual cells being random technical replicates and individual samples being random biological replicates. The weighted t test and all these models *conceptually* assume that the sequenced cells are randomly selected representatives of larger cell populations. In scRNASeq experiments, this is only partially true. Even in most careful experiments, at least a small fraction of cells is expected to be dramatically altered from their original state by the cell isolation process, e.g., resulting in apoptosis initiation. Expression of some genes in these cells may be orders of magnitude different from all other cells in the population and not random. As a result, even the most robust weighted t test may detect these genes as differentially expressed, which is not desirable.

To mitigate this problem when comparing two groups of cells A and B, we suggest combining the weighted t test with a χ^2 test for 2 x 2 contingency table $O_{m,n}$ for aggregate counts of gene Z vs. other genes. This contingency table is $O_{1,1} = N_{Z,A}$, $O_{1,2} = N_{Z,B}$, $O_{2,1} = N_A$-$N_{Z,A}$, $O_{2,2} = N_B$-$N_{Z,B}$, where $N_{Z,A} = \sum_{i \in A} N_{Z,i}$, $N_{Z,B} = \sum_{i \in B} N_{Z,i}$, $N_A = \sum_Z N_{Z,A}$, and $N_B = \sum_Z N_{Z,B}$.

The benefit of this approach is that the χ^2 test does not assume individual cells to be properly randomized replicates. Instead, it calculates the p-values only for the aggregated counts in specific cells sequenced in the scRNASeq experiment without generalizing it to all cells of the same type in the sample (see the derivation below). However, $n_{Z,i}$ variance within the sequenced cells may be smaller than the variance in the overall population of cells of the same type (even within the same sample). Therefore, positive findings of the χ^2 test may not mean differential expression in the latter population. At the same time, negative findings of the χ^2 test mean low likelihood of differential expression in the sequenced cells, helping to eliminate weighted t test positives related to sampling bias. Therefore, performing the χ^2 test and weighted t test for all genes may be beneficial for reducing type I errors caused by experimental cell selection and processing bias.

Briefly, the validity of the χ^2 test for this specific application can be proven as follows. This test for the contingency table $O_{m,n}$ is based on a χ^2 distribution for

$$X^2 = \sum_{m=1}^{2} \sum_{n=1}^{2} \frac{\left(O_{m,n} - E_{m,n}\right)^2}{E_{m,n}} \quad \text{(Equation 25)}$$

Here $E_{m,n}$ are the expected values of $O_{m,n}$ at $N_{Z,A}/N_A = N_{Z,B}/N_B$ (log_2(FC) = 0). X^2 has the χ^2 distribution in the asymptotic limit of large $O_{m,n}$ when $Var(O_{m,n}) = E_{m,n}$.[33] This simply requires $N_{Z,A}$ or $N_{Z,B}$ to be sufficiently large (our preference is > 30–50). Provided that this condition is satisfied, the accuracy of the χ^2 test hinges only on how well $Var(N_Z)$ is approximated by $\langle N_Z \rangle$. At $Var(N_Z) < \langle N_Z \rangle$, the test becomes conservative. At $Var(N_Z) > \langle N_Z \rangle$, it becomes anti-conservative.

Taking into account that $N_{Z,i} = N_i n_{Z,i}$ and therefore $Var(N_{Z,i}) = N_i^2\langle(n_{Z,i} - \langle n_{Z,i}\rangle)^2\rangle$, similar to the derivation of Equation 8 we find that

$$Var(N_{Z,i}) = N_i\langle\langle n_{Z,i}\rangle_i(1 - \langle n_{Z,i}\rangle_i)\rangle \cong \langle N_{Z,i}\rangle \quad \text{(Equation 26)}$$

as long as sequencing of each transcript j in the cell i can be considered an independent transcript sampling event. Note that here the averaging indicated by the symbol $\langle\ \rangle$ is performed only over statistical trajectories that begin after encapsulation of each cell i into a droplet for single cell sequencing and does not include averaging over possible cell sampling trajectories. Therefore, $Var(N_{Z,i})$ does not include the variance associated with cell sampling, which is described by V_Z in Equation 8.

Because each cell i is encapsulated in an independent droplet, the sequencing of its transcripts is independent from sequencing of transcripts in other cells. Therefore,

$$Var(N_Z) = \sum_i Var(N_{Z,i}) \cong \sum_i \langle N_{Z,i}\rangle = \langle N_Z\rangle \quad \text{(Equation 27)}$$

This proves that the χ^2 test provides an accurate estimate for the probability of the observed $N_{Z,A}$ and $N_{Z,B}$ at $N_{Z,A}/N_A = N_{Z,B}/N_B$. However, this is the probability only for the cells sequenced by NGS. It bears reminding that $Var(N_Z)$ for statistical trajectories that begin at the sample stage before cell isolation and encapsulation into droplets may be larger, even significantly larger. Thus, the χ^2 test should not be used for inferring the p-value for a larger population of cells (even from the same sample).

Sampling and technical noise in spRNASeq

Aside from spatial information (Figure 6), the output of sequencing based spRNASeq assays is like scRNASeq, i.e., $N_{Z,i}$ matrix of gene Z counts in a spatial bin i. In contrast to scRNASeq, however, a single bin i is likely to include transcripts from multiple cells, and adjacent bins may contain transcripts from the same cell. As a result, physics of transcript sampling and technical noise in spRNASeq depends on the spatial resolution of the assay.

At low resolution (bin size ~50 μm or larger), adjacent bins can be considered independent. DGE analysis in such assays is similar to scRNASeq. Provided that enough bins are included in the analysis (preferably ~100 or more), the approach described in the main text can be utilized without modifications. The only difference is that the bins must be interpreted as variable composition mixtures of cells. Note that relative count normalization may not be the best choice for some applications of these assays. For instance, normalization based on cell-type-specific endogenous control genes may enable analysis of gene expression in specific cell types despite the lack of single cell resolution. Approaches to such normalization are, however, context specific and are beyond the scope of the present paper.

At high spatial resolution (bin size ~10 μm or smaller), transcript counts in adjacent bins are likely to be correlated, precluding treatment of individual bins as independent replicates. One solution is correlation-based models, which are justifiable for isotropic tissues with similar correlations between bins in all spatial directions. Since we are more interested in anisotropic tissues like those shown in Figure 6, we are not pursuing this approach.

Another solution is clustering of bins based on spatial localization (Figure 6) or marker gene expression followed by analysis of aggregated transcript counts $\overline{n}_{Z,c} = \sum_{i\in c}N_{Z,i}/\sum_{i\in c}N_i$. Here $N_i = \sum_Z N_{Z,i}$ is the total number of counts in bin i. Transcript counts in sufficiently large clusters can be considered independent. Within such a cluster, the statistical weight of each bin is $w_{Z,i} = N_i/\sum_i N_i$, which is a good approximation at $\rho_Z(N_i\text{-}1)\ll 1$ (see Equation 3). Figure 1B shows that this approximation works reasonably well for many genes even at N_i between 10^4 and 10^5 in scRNASeq. We can then expect it to work even better in high resolution spRNASeq, where N_i ~1000 or lower. At the current state of the technology, transcript counts in spRNASeq are simply not sufficient for analyzing biological variability within individual cell clusters.

DGE between two cell clusters can then be analyzed based on combining the weighted t test and χ^2 test as described in the main text. A more accurate DGE comparison can be made between two groups of multiple clusters, using weighted t test like in multiple sample scRNASeq experiments (replacing $\overline{n}_{Z,s}$ and $W_{Z,s}$ in Equations 22, 23, and 24 with $\overline{n}_{Z,c}$ and the corresponding statistical weight $W_{Z,c}$ for each cluster). The same procedure can be further repeated for multiple samples.

METHOD DETAILS: SUPPLEMENTAL DERIVATIONS

Derivation of Equation 4

To address possible transcript sampling and identification failures in scRNASeq experiments, consider a more general form of Equation 4,

$$n_{Z,i} = \frac{1}{N_i}\sum_{j=1}^{N_i^{tot}} X_{Z,i,j} \quad \text{(Equation 28)}$$

where N_i^{tot} is the unknown total number of transcripts in the cell i, including those identified and not identified by scRNASeq. $X_{Z,i,j} = 1$ for transcripts j in the cell i that are identified by scRNASeq as gene Z mRNAs. $X_{Z,i,j} = 0$ for all other transcripts (successfully identified

by scRNASeq as transcripts of other genes, sampled but not identified by failed sequencing, and not sampled at all). Since $X_{Z,i,j} \neq 0$ only for identified transcripts, we can limit the summation in Equation 28 to the latter, i.e.,

$$\frac{1}{N_i}\sum_{j=1}^{N_i^{tot}} X_{Z,i,j} = \frac{1}{N_i}\sum_{j=1}^{N_i} X_{Z,i,j}, \quad \text{(Equation 29)}$$

recovering the original Equation 4.

Derivation of Equation 8

Substitution of Equation 4 into Equation 7 yields

$$Var(n_{Z,i}) = \left\langle \left(\frac{1}{N_i}\sum_{j=1}^{N_i} X_{Z,i,j} - \left\langle \frac{1}{N_i}\sum_{j=1}^{N_i} X_{Z,i,j} \right\rangle_i \right)^2 \right\rangle + V_Z = \left\langle \left(\frac{1}{N_i}\sum_{j=1}^{N_i} \left(X_{Z,i,j} - \langle X_{Z,i,j}\rangle_i \right) \right)^2 \right\rangle + V_Z. \quad \text{(Equation 30)}$$

Mathematically, our assumption of random and independent transcript identification trials means that $\langle (X_{Z,i,j} - \langle X_{Z,i,j}\rangle_i)(X_{Z,i,k} - \langle X_{Z,i,k}\rangle_i)\rangle_i = 0$ at $j \neq k$. Therefore,

$$Var(n_{Z,i}) = \left\langle \frac{1}{N_i^2}\sum_{j=1}^{N_i} \left(X_{Z,i,j} - \langle X_{Z,i,j}\rangle_i \right)^2 \right\rangle + V_Z. \quad \text{(Equation 31)}$$

After rewriting Equation 31 as

$$Var(n_{Z,i}) = \left\langle \frac{1}{N_i^2}\sum_{j \in \{X_{Z,i,j}=0\}}^{N_i} \left(X_{Z,i,j} - \langle X_{Z,i,j}\rangle_i \right)^2 + \frac{1}{N_i^2}\sum_{j \in \{X_{Z,i,j}=1\}}^{N_i} \left(X_{Z,i,j} - \langle X_{Z,i,j}\rangle_i \right)^2 \right\rangle + V_Z, \quad \text{(Equation 32)}$$

where $\{X_{Z,i,j} = 0\}$ and $\{X_{Z,i,j} = 1\}$ are the sets of j values with the corresponding $X_{Z,i,j}$, and using that

$$\langle n_{Z,i}\rangle_i = \frac{1}{N_i}\sum_{j=1}^{N_i} \langle X_{Z,i,j}\rangle_i = \langle X_{Z,i,j}\rangle_i, \quad \text{(Equation 33)}$$

we find

$$Var(n_{Z,i}) = \left\langle \frac{1}{N_i^2}\sum_{j \in \{X_{Z,i,j}=0\}}^{N_i} \langle n_{Z,i}\rangle_i^2 + \frac{1}{N_i^2}\sum_{j \in \{X_{Z,i,j}=1\}}^{N_i} \left(1 - \langle n_{Z,i}\rangle_i \right)^2 \right\rangle + V_Z. \quad \text{(Equation 34)}$$

Then, after substitution of

$$\sum_{j \in \{X_{Z,i,j}=1\}}^{N_i} 1 = N_i n_{Z,i}, \quad \sum_{j \in \{X_{Z,i,j}=0\}}^{N_i} 1 = N_i\left(1 - n_{Z,i}\right) \quad \text{(Equation 35)}$$

into Equation 34, we arrive at

$$Var(n_{Z,i}) = \left\langle \frac{1}{N_i^2}\langle n_{Z,i}\rangle_i^2 N_i\left(1 - n_{Z,i}\right) + \frac{1}{N_i^2}\left(1 - \langle n_{Z,i}\rangle_i\right)^2 N_i n_{Z,i} \right\rangle + V_Z. \quad \text{(Equation 36)}$$

Since the overall averaging can be performed in two steps, first averaging within each cell i and then across different cells, we can replace $n_{Z,i}$ in Equation 36 with $\langle n_{Z,i}\rangle_i$ and simplify this equation to

$$Var(n_{Z,i}) = \left\langle \frac{\langle n_{Z,i}\rangle_i\left(1 - \langle n_{Z,i}\rangle_i\right)}{N_i} \right\rangle + V_Z. \quad \text{(Equation 37)}$$

Next, using that $\langle \langle n_{Z,i}\rangle_i \rangle = \langle n_{Z,i}\rangle$ and

$$V_Z = Var\left(\langle n_{Z,i}\rangle_i\right) = \langle \langle n_{Z,i}\rangle_i^2 \rangle - \langle \langle n_{Z,i}\rangle_i \rangle^2 \quad \text{(Equation 38)}$$

we arrive at

$$Var(n_{Z,i}) = \frac{\langle n_{Z,i}\rangle}{N_i} - \frac{Var\left(\langle n_{Z,i}\rangle\right) + \langle n_{Z,i}\rangle^2}{N_i} + V_Z = \frac{\langle n_{Z,i}\rangle}{N_i} - \frac{V_Z + \langle n_{Z,i}\rangle^2}{N_i} + V_Z\,. \quad \text{(Equation 39)}$$

Finally, after taking into account that

$$\overline{n}_Z = \sum_{i=1}^{N} n_{Z,i} w_{Z,i} = \langle n_{Z,i} \rangle, \quad \text{(Equation 40)}$$

we arrive at Equation 8.

In other words, Equations 8 and 9 are valid for any distributions of $X_{Z,i,j}$ in the asymptotic limit of large N, as long as the sampling of transcripts is random and independent. These distributions include but are not limited to the popular binomial, negative binomial, and zero-inflated negative binomial distributions. Because these equations are based on unknown $\overline{n}_Z$ and V_Z rather than a defined probability distribution for $X_{Z,i,j} = 1$, they do not require assuming anything about cell transcripts that have not been identified by scRNASeq sampling. The unknown $\overline{n}_Z$ and V_Z are then calculated directly from the experimental data without any additional assumptions by combining Equations 2, 3, and 18, as described in the Theory section.

Model of colored balls in a basket

To illustrate how Equations 8 and 39 can be derived from common counting models, consider a textbook model of colored balls in baskets. Specifically, each cell i corresponds to a basket i, $N_{Z,i}$ corresponds to the number of balls with color Z drawn from that basket, N_i corresponds to the total number of random draws from the basket i. The total numbers of color Z balls ($N_{Z,i}^{tot}$) and all balls (N_i^{tot}) in the basket i are not known, just like the total numbers of gene Z and all mRNA transcripts. The goal is to estimate

$$\overline{n}_Z = \left\langle N_{Z,i}^{tot} \middle/ N_i^{tot} \right\rangle \quad \text{(Equation 41)}$$

and $Var\left(N_{Z,i}^{tot} / N_i^{tot}\right)$ based on the known $N_{Z,i}$ and N_i, so that the difference in the prevalence of color Z balls between two different sets of baskets can be evaluated based on the weighted t test.

This model is described by the hypergeometric probability distribution for $N_{Z,i}$ that has the within-basket mean $\langle N_{Z,i} \rangle_i = N_i \left(N_{Z,i}^{tot} / N_i^{tot}\right)$ and the variance

$$Var_i\left(N_{Z,i}\right) = N_i \left(\frac{N_{Z,i}^{tot}}{N_i^{tot}}\right)\left(\frac{N_i^{tot} - N_{Z,i}^{tot}}{N_i^{tot}}\right)\left(\frac{N_i^{tot} - N_i}{N_i^{tot} - 1}\right), \quad \text{(Equation 42)}$$

where $Var_i(N_{Z,i})$ is the variance within a single basket i. Then, using the law of total variance, we find that the total valiance of $n_{Z,i} = N_{Z,i}/N_i$ for all baskets with N_i draws is

$$Var\left(n_{Z,i}\right) = \frac{\langle Var_i\left(N_{Z,i}\right)\rangle}{N_i^2} + V_Z, \quad \text{(Equation 43)}$$

where $V_Z = Var\left(N_{Z,i}^{tot} / N_i^{tot}\right)$ is unknown. After substituting $n_{Z,i}^{tot} = N_{Z,i}^{tot}/N_i^{tot}$ and Equation 42 into Equation 43, we find

$$Var\left(n_{Z,i}\right) = \frac{1}{N_i}\left\langle n_{Z,i}^{tot}\left(1 - n_{Z,i}^{tot}\right)(1 - SR_i)\right\rangle + V_z, \quad \text{(Equation 44)}$$

where $SR_i = (N_i - 1)/\left(N_i^{tot} - 1\right)$ is the sampling ratio in the basket i. Taking into account that

$$\left\langle n_{Z,i}^{tot}\left(1 - n_{Z,i}^{tot}\right)\right\rangle = \left\langle n_{Z,i}^{tot}\right\rangle - \left\langle \left(n_{Z,i}^{tot}\right)^2\right\rangle = \left\langle n_{Z,i}^{tot}\right\rangle - \left\langle n_{Z,i}^{tot}\right\rangle^2 - V_z, \quad \text{(Equation 45)}$$

we find

$$Var\left(n_{Z,i}\right) = \frac{\overline{n}_Z(1 - \overline{n}_Z) + V_Z(N_i - 1)}{N_i} - \frac{\left\langle SR_i n_{Z,i}^{tot}\left(1 - n_{Z,i}^{tot}\right)\right\rangle}{N_i}. \quad \text{(Equation 46)}$$

At $SR_i \ll 1$, i.e., when the sampling is random and independent, the last term in Equation 46 is negligible and we recover Equations 8 and 39. Then, at sufficiently large number of baskets, Equations 2, 3, and 18 can be used to find the unknown values of $\overline{n}_Z$ and V_Z without any additional assumptions based on matching the predicted and observed values of weighted average and variance.

The instructive value of this model is that it illustrates when and why one can analyze relative prevalence $n_{Z,i}$ and accurately estimate its average $\overline{n}_Z$ and variance V_Z from the known $N_{Z,i}$ and N_i, even though $N_{Z,i}^{tot}$ and N_i^{tot} are not known. In the context of this specific model, the key condition is the low sampling ratio ($N_i \ll N_i^{tot}$), which ensures random and independent sampling. At $N_i \sim N_i^{tot}$, the problem becomes more complex and may require additional assumptions. In the context of scRNASeq, $N_i \ll N_i^{tot}$ in all existing and foreseeable assays, so that such assumptions (which may not be justified) are not necessary. More generally, random and independent sampling is all that is needed to accurately estimate relative prevalence of a trait in a large group of independent population clusters (baskets in this example or cells in scRNASeq).

Binomial trials with variable outcome probability

In a canonical sequence of Bernoulli trials, the positive outcome probability (rate) p is the same in all trials, yielding binomial distribution. In the limit of infinite N_i^{tot}, Equation 8 can be derived by a straightforward application of the law of total variance to an ensemble of cells with fixed N_i binomial trials each, where each cell has a different, but fixed p. However, Equation 8 is valid also when p is variable rather than constant even within each cell (as our derivation above proves.) It is therefore instructive to illustrate this by using the binomial/Bernoulli distribution logic yet relaxing the requirement of constant within-cell p to ensure more realistic description of scRNASeq.

Within a cell i, an individual trial j that has the success probability p_j is described by the Bernoulli distribution with $\langle X_{Z,i,j}\rangle_j = p_j$ and $Var_j(X_{Z,i,j}) = p_j(1\text{-}p_j)$. When p_j varies for individual transcripts of the same gene Z even within the same cell, the within-cell variance of $X_{Z,i,j}$ is

$$Var_i\left(X_{Z,i,j}\right) = \left\langle \left(X_{Z,i,j} - p_j + p_j - p\right)^2 \right\rangle_i = \left\langle \left(X_{Z,i,j} - p_j\right)^2 \right\rangle_i + \left\langle \left(p_j - p\right)^2 \right\rangle_i = \left\langle Var_j\left(X_{Z,i,j}\right)\right\rangle_i + Var_i\left(p_j\right) = \left\langle p_j\left(1 - p_j\right)\right\rangle_i + Var_i\left(p_j\right) = p(1 - p), \quad \text{(Equation 47)}$$

where $p = \langle p_j\rangle_i$ is the within-cell average rate, and we used that $\langle x\rangle_i = \langle\langle x\rangle_j\rangle_i$ and $\langle p_j^2\rangle_i = \langle p_j\rangle_i^2 + Var_i\left(p_j\right)$.

For N_i random and independent trials we can then calculate the across-cell variance of $n_{Z,i}$ as

$$Var\left(n_{Z,i}\right) = \frac{\langle Var_i\left(N_{Z,i}\right)\rangle}{N_i^2} + V_Z = \frac{\left\langle \sum_{j=1}^{N_i} Var_i\left(X_{Z,i,j}\right)\right\rangle}{N_i^2} + V_Z = \frac{\langle p(1 - p)\rangle}{N_i} + V_Z\,. \quad \text{(Equation 48)}$$

Here $\langle\ \rangle$ indicates averaging across all cells (as before), $V_Z = Var(p)$, we used the law of total variance (Equation 7), and we used that the variance of a sum of independent random variables is equal to the sum of variances of these variables. After substitution of $\langle p^2\rangle = \langle p\rangle^2 + Var(p)$ into Equation 48, we find

$$Var\left(n_{Z,i}\right) = \frac{\langle p\rangle(1 - \langle p\rangle) + V_Z(N_i - 1)}{N_i} \quad \text{(Equation 49)}$$

Equation 49 has the same level of generality as our original Equation 8. Moreover, since $\langle p\rangle = \bar{n}_Z$, the two equations are actually identical. The key observations here are: (a) The only difference between Equation 49 and the corresponding expression for binomially distributed $N_{Z,i}$ is that $\langle p\rangle$ averages p variations both within and across the cells rather than just p variations across the cells. (b) The variance for a single independent trial with a binary outcome and random probability p is $\langle p\rangle(1\text{-}\langle p\rangle)$, i.e., it depends only on $\langle p\rangle$ but not on the variability of p (Equations 47 and 49 with $N_i = 1$).

These observations suggest that Equation 46 may work beyond the hypergeometric distribution not only when $SR_i \to 0$ but also at larger SR_i. However, this is a nontrivial question because at $SR_i \sim 1$ the expected variance is no longer a sum of variances for individual trials. Analysis of this question is beyond the scope of our study, since the case of $N_i \sim N_i^{tot}$ is not practically relevant for scRNASeq.

QUANTIFICATION AND STATISTICAL ANALYSIS

All data analysis was performed in R-Studio v. 2024.12.0 Build 467 (Posit Software) using R v. 4.4.2, Seurat v. 5.2.0 3 and custom R code shared at https://github.com/sergeyleikin/sc-sp-RNASeq. All statistical tests were performed as described in the main text and Method Details: Theory section above. All t tests and weighted t tests were two-tailed with unequal variances. SigmaPlot v.13.0 (Systat Software) was utilized for generating all plots.