Cell Systems

CellPress

Commentary

Beyond differential expression: Differential variability as an underappreciated axis in transcriptomics

Chenxin Flora Jiang,[1] Changhu Wang,[2] and Jingyi Jessica Li[1,2,3,*]
[1]Department of Statistics and Data Science, University of California, Los Angeles, Los Angeles, CA 90095, USA
[2]Biostatistics Program, Public Health Science Division, Fred Hutchinson Cancer Center, Seattle, WA 98109, USA
[3]Department of Biostatistics, University of Washington, Seattle, WA 98195, USA
*Correspondence: lijy03@fredhutch.org
https://doi.org/10.1016/j.cels.2026.101706

Differential expression has long been the default lens for comparing transcriptomes between conditions, but mean shifts alone can miss important biological signals. Using a recent debate on memory-associated transcriptome changes, this commentary highlights differential variability as an underappreciated axis for interpreting transcriptome variation.

The rise of single-cell and spatial transcriptomics is transforming biomedical research, and while analytical approaches are diverse, practice still centers on a few standard perspectives. The prevailing approach for comparing biological contexts (e.g., cell types or conditions) is differential expression (DE), which tests for shifts in mean gene expression. Yet this focus alone is often insufficient because it overlooks an equally informative axis—differential variability (DV)—which captures changes in expression heterogeneity between contexts. DV can arise from mechanisms such as transcription-factor-driven expression pulses, chromatin-state switching, and shifts in cellular states.[1,2] Considering DV alongside DE can reveal biology that mean-based tests miss.

A recent debate surrounding a study by Sun et al.,[3] critiqued by Mukamel and Yu[4] and defended by Sun et al.,[5] provides a timely case study. The controversy centers on the existence of DE genes associated with long-term memory: Sun et al.[3,5] reported DE genes in single-cell transcriptomic data analysis, whereas Mukamel and Yu[4] argued that the findings did not meet rigorous statistical criteria. Here, we present a comprehensive reanalysis of this dataset to highlight the limits of a DE-centric view and show that incorporating DV offers a more complete and rigorous interpretation of single-cell transcriptomic data.

Following analyses by the two groups in the debate (Sun et al.[3,5] and Mukamel and Yu[4]), we examined transcriptomic differences between the "fear training and recall" (FR) and "no fear" (NF) conditions in Targeted Recombination in Active Populations (TRAP)-labeled BLA.Int.Gpr88 neurons. The two groups' divergent results stemmed from several factors: (1) using pseudobulk versus single-cell analyses and (2) performing different statistical procedures, particularly in whether multiple-testing correction is applied and whether gene preselection is performed on the same dataset prior to DE testing, as such preselection can inflate false positives (Figure 1A). Motivated by these differences, we conducted an independent, expanded DE analysis on the same dataset in which we incorporated rigorous stability assessments at both the pseudobulk and single-cell levels.

Mukamel and Yu[4] aggregated single-cell data into pseudobulk profiles and, with a small cohort (4 FR vs. 7 NF mice), applied the nonparametric Wilcoxon rank-sum test, reporting zero DE genes at a false discovery rate (FDR) of 5%. We first asked whether limited statistical power at this sample size could explain the null result. Accordingly, we reanalyzed the pseudobulk data using two popular parametric DE methods, DESeq2[6] and edgeR,[7] to check whether model-based approaches, which are generally more powerful statistically, could reveal any DE genes. Although DESeq2 reported 43 DE genes at an FDR of 5%, a label-permutation analysis, which randomly shuffled the 4 FR and 7 NF labels across the 11 mice, revealed that the findings were not reliable: DESeq2 frequently identified as many or more "DE" genes in the permuted datasets than in the original, and this indicates inflated false positives (Figure 1B). Furthermore, the lack of a reliable DE signal was corroborated by additional results: edgeR detected no DE genes, and after calibration by NullstrapDE,[8] an add-on we developed to mitigate false positives arising from model misspecification (e.g., violations of the negative binomial assumptions in DESeq2 and edgeR), neither DESeq2 nor edgeR yielded any DE genes (Figure 1B).

In contrast to the pseudobulk approach, Sun et al.[3,5] carried out DE analysis at the single-cell level. In our reanalysis, when we tested 3,574 genes, no genes met significance after multiple-testing correction at an FDR of 5% (Figure 1C), consistent with Mukamel and Yu.[4] Sun et al.[3,5] reported DE genes within a preselected panel of 56 genes; however, because that panel was defined from the same dataset using a log-fold-change filter, this introduces a risk of selection bias (also known as "double-dipping"[9]) (Figure 1A). To assess whether a 5% FDR threshold might simply be too conservative, we directly examined the functional annotations of the top 5% ranked DE genes without imposing FDR control. We also applied a stability selection[10] procedure to identify the consistently top-ranked 5% of genes (i.e., genes most frequently appearing in the top 5% DE set across subsamples) among all

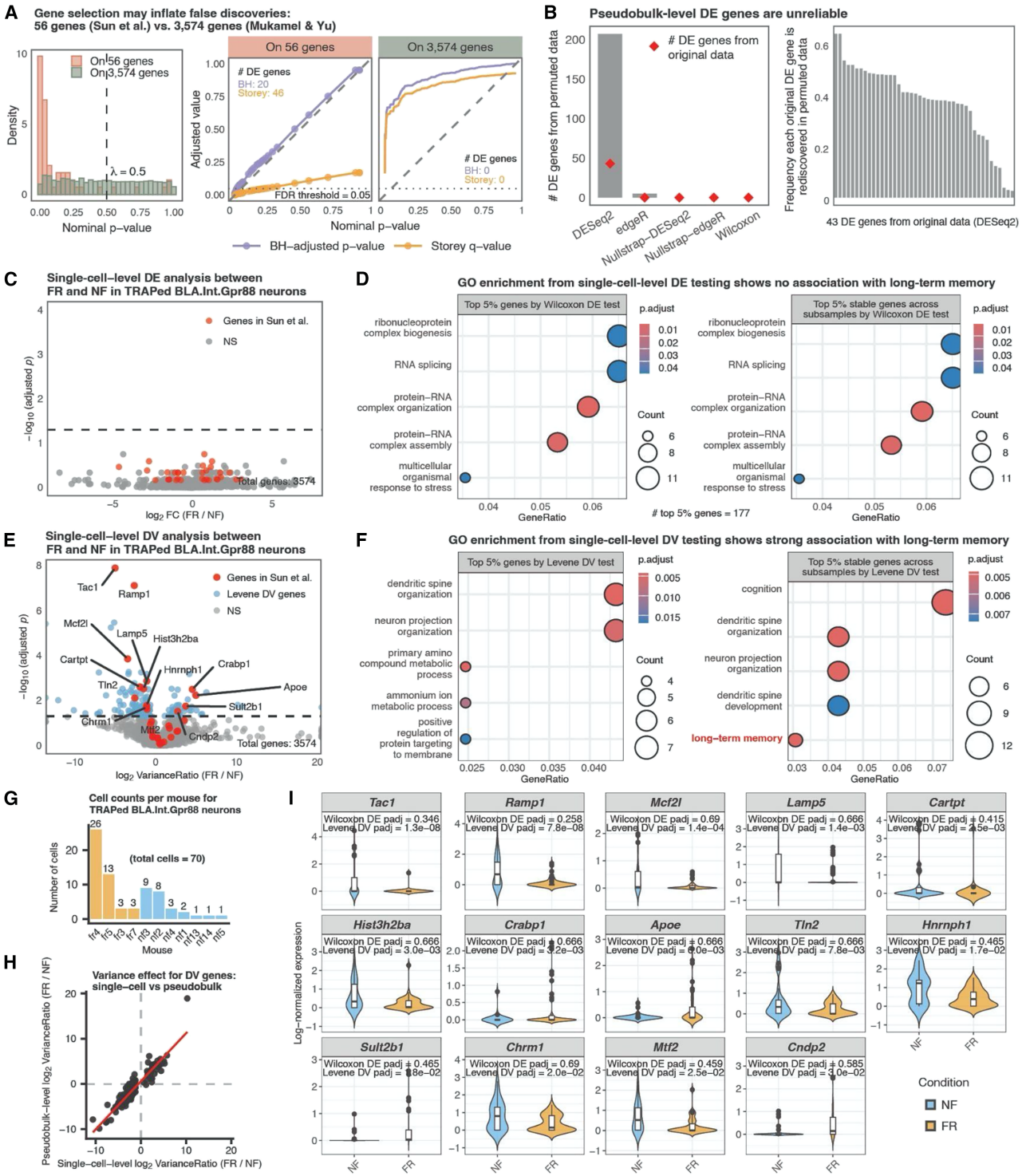

Figure 1. A comparative analysis revealing the limitations of differential expression (DE) and the importance of differential variability (DV) in single-cell transcriptomic data analysis

(A) Gene selection may inflate false discoveries in DE analysis. Left: distributions of nominal *p* values from Sun et al.[3,5] (56 genes), which are highly concentrated near zero, and from Mukamel and Yu[4] (3,574 genes), which are close to uniform. Middle: for the preselected 56 genes, multiple-testing correction using the Benjamini-Hochberg (BH) or Storey procedure is ineffective due to violation of correction assumptions that makes the adjusted *p* values as large as or even smaller than the unadjusted ones. Right: for all 3,574 genes, multiple-testing correction by either procedure behaves appropriately, with adjusted *p* values larger than the corresponding nominal values.

(B) Pseudobulk-level DE analyses using DESeq2, edgeR, and Nullstrap-DE add-on—coupled with permutation-based sanity checks—suggest that no DE genes can be reliably detected; in particular, permutation reveals that DESeq2 detects spurious DE genes, which are removed by Nullstrap-DE calibration.

(legend continued on next page)

3,574 genes. Gene Ontology (GO) enrichment analyses of both gene sets revealed no enrichment for memory-related processes, instead highlighting general cellular functions (e.g., "ribonucleoprotein complex biogenesis"; Figure 1D). Moreover, inspection of the single-cell profiles indicated that many genes previously reported as DE by Sun et al.[3] show clearer differences in variability than in mean expression between FR and NF conditions for TRAP-labeled BLA.Int.Gpr88 neurons (Figure 1I).

Taken together, these analyses across scales and methods indicate that this dataset lacks statistically and biologically compelling DE signals and suggest that a DE-centric perspective is insufficient to characterize the transcriptomic differences between FR and NF in TRAP-labeled BLA.Int.Gpr88 neurons.

Building on these results, we hypothesized that biologically meaningful changes in this setting may lie not in the mean but in the variability of gene expression. DV analysis captures changes in variability that need not coincide with shifts in average expression[1] and is particularly pertinent in complex neural systems, where transcriptional heterogeneity can be functionally consequential.[11] Accordingly, we performed DV analysis at the single-cell level rather than using pseudobulk, because (1) the extremely limited and uneven number of cells contributed by each mouse makes mouse-level differential variance testing underpowered (Figure 1G), and (2) variance ratios estimated from pooled single-cell data are highly concordant with those from pseudobulk mouse-level aggregation, which indicates that there is no evident systematic bias between the two approaches at the point estimate level (Figure 1H). We then applied the Levene test[12] at the single-cell level to assess whether gene-expression variance differs between FR and NF in TRAP-labeled BLA.Int.Gpr88 neurons.

The DV results provide a complementary analytical perspective and a clear path to reconcile the prior debate. Our analysis identified 169 DV genes (FDR = 5%; Figure 1E), including several genes previously highlighted by Sun et al.,[3] such as *Tac1* and *Ramp1*. Violin plots for genes overlapping between our DV set and the "DE" genes reported by Sun et al.[3] show pronounced differences in expression variance between FR and NF, with minimal shifts in mean levels (Figure 1I). This overlap is presented for illustration and suggests that these previously reported signals can be more reliably captured by variability rather than mean expression and should not be interpreted as a validation of the original DE claims. These patterns remain after Winsorization at the 95th percentile, indicating that the observed variability differences are not driven by outlier cells (Figure S1A). To assess biological relevance, mirroring our DE workflow, we analyzed two sets of genes: (1) the top 5% ranked DV genes among the 3,574 tested and (2) the top 5% genes identified by stability selection. We then performed GO enrichment analysis on both sets. In contrast to the functionally uninformative DE results, both DV gene sets, particularly the stability-selected set, were enriched for memory-related processes (e.g., "dendritic spine organization" and "long-term memory"; Figure 1F). We further assessed the robustness of DV signals using leave-one-mouse-out and cell permutation analyses. DV genes were consistently rediscovered across leave-one-mouse-out subsamples but rarely under permutation; this indicates that the signal is not driven by individual mice or random, condition-irrelevant variability (Figures S1B–E). Together, these findings indicate that DV analysis captures the primary signal in this dataset and suggest that variability, rather than mean expression, is the key transcriptomic difference between FR and NF.

In summary, our reanalysis of the dataset at the center of the Sun et al.[3,5] and Mukamel and Yu[4] debate underscores three lessons for transcriptomic data analysis. First, statistical analysis should be guided by the hypothesis being tested, rather than by the *a priori* assumption of a particular type of signal; for example, DE tests for shifts in mean expression, whereas DV tests for changes in expression variation across cells. Evaluating both can provide a more comprehensive view of the data. Second, our findings offer a partial reconciliation of the debate: DV analysis confirms the biological relevance of several genes reported by Sun et al.,[3] while simultaneously affirming Mukamel and Yu's[4] emphasis on statistical rigor—specifically the need to control the FDR, avoid gene preselection (double-dipping), and conduct robust analyses. Third, our results demonstrate that DV is not merely a technical alternative but an underappreciated and essential axis of transcriptomic variation, one that can reveal biologically meaningful insights missed by DE-focused approaches. This case study illustrates the power of aligning the analytic perspective with the underlying biology and emphasizes that new insights often emerge when we look beyond the conventional axis of mean differences.

Methods

The processed single-cell RNA-sequencing (RNA-seq) gene expression data were downloaded from GEO (GEO: GSE246147). Cell annotations for neuron subtypes, including BLA.Int.Gpr88 and others, were kindly provided by Dr. Wenfei Sun. Following Sun et al.,[3] TRAPed (tdT+) "engram" neurons were identified by applying a threshold of log-normalized *tdT* expression greater than 1. We focused on the TRAPed BLA.Int.Gpr88 neurons, which yielded a total of 70 cells

(C and D) (C) Single-cell-level DE analysis using the Wilcoxon rank-sum test identifies no DE genes at an FDR threshold of 0.05. "NS" stands for non-significant genes. (D) Single-cell-level DE analyses show that the top 5% of genes, whether ranked directly by the Wilcoxon rank-sum test or via subsampling-based stability analysis, are not enriched for long-term memory-related functions, even without FDR correction.
(E and F) (E) Single-cell-level DV analysis using the Levene test identifies 169 DV genes at an FDR threshold of 0.05, many of which overlap with those originally highlighted by Sun et al.[3,5] "NS" stands for non-significant genes. (F) The top 5% of genes, selected either by the Levene test or by subsampling-based stability analysis, are both enriched for memory-related GO terms, with the latter showing stronger and more specific enrichment for long-term memory.
(G) Number of TRAP-labeled BLA.Int.Gpr88 neurons contributed by each mouse in the FR and NF conditions, highlighting the limited per-mouse cell counts.
(H) Concordance between variance ratios (FR vs. NF) estimated from pooled single-cell data and from pseudobulk mouse-level aggregation, with the relationship approximated by a diagonal line passing through the origin, indicating no evident systematic bias between the two approaches.
(I) Violin plots of DV genes identified by the Levene test that overlap with the significant genes reported by Sun et al.[3,5] For each gene, the distributions of log-normalized expression levels in cells from the NF and FR conditions are shown. The BH-adjusted *p* values for DE (Wilcoxon rank-sum test) and DV (Levene test) are indicated above each plot.

(45 FR and 25 NF). Pseudobulk-level data were generated by aggregating single-cell-level data at the individual (mouse) level, which resulted in 11 samples (4 FR and 7 NF).

For direct comparison with Sun et al.[3,5] and Mukamel and Yu,[4] we applied the same gene prefiltering criteria as Mukamel and Yu,[4] retaining the 3,574 genes with a maximum normalized expression (counts per million) greater than 600 in at least one cell. This gene set also approximately matches the genes analyzed in Figure 2E of Sun et al.[3]

For pseudobulk-level DE testing, we used DESeq2 (version 1.48.1) and edgeR (version 4.6.2) implemented in R. For single-cell-level DE testing, we applied the Wilcoxon rank-sum test using the wilcox.test function from the stats package (version 4.5.0). Single-cell-level DV testing was performed using the Levene test, implemented via the leveneTest function in the car package (version 3.1-3). GO enrichment analysis was conducted with the clusterProfiler package (version 4.16.0), using the enrichGO function with ontology set to "BP" (biological process). Gene sets with a Benjamini-Hochberg-adjusted p value < 0.05 were considered significant.

Data and code availability

The processed single-cell RNA-seq gene expression data were downloaded from GEO (GEO: GSE246147). Cell annotations for neuron subtypes, including BLA.Int.Gpr88 and others, were kindly provided by Dr. Wenfei Sun. The source code and data for reproducing the analyses presented here are available at https://doi.org/10.5281/zenodo.19991665.

ACKNOWLEDGMENTS

We thank Dr. Wenfei Sun for generously sharing the code and assisting with access to the data, which enabled our reanalysis. This work was supported by the National Institute of General Medical Sciences (R35GM140888), the National Human Genome Research Institute (R01HG014687 and Opportunity Fund subaward from the Technology Development Coordinating Center U24HG011735), and the National Science Foundation (DBI-1846216 and DMS-2113754) to J.J.L.

DECLARATION OF INTERESTS

J.J.L. serves on the *Cell Systems* advisory board.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2026.101706.

REFERENCES

1. Osorio, D., Yu, X., Zhong, Y., Li, G., Serpedin, E., Huang, J.Z., and Cai, J.J. (2019). Single-Cell Expression Variability Implies Cell Function. Cells *9*, 14. https://doi.org/10.3390/cells9010014.

2. Andrews, C.J., Costa, E.J., Janer Carattini, G.L., DelRosso, N.V., Fujimori, T., Shimasawa, M., and Bintu, L. (2025). Bifunctional transcriptional effector domains control gene expression pulses in an occupancy-dependent manner. Preprint at bioRxiv. https://doi.org/10.64898/2025.12.10.693551.

3. Sun, W., Liu, Z., Jiang, X., Chen, M.B., Dong, H., Liu, J., Südhof, T.C., and Quake, S.R. (2024). Spatial transcriptomics reveal neuron–astrocyte synergy in long-term memory. Nature *627*, 374–381. https://doi.org/10.1038/s41586-023-07011-6.

4. Mukamel, E.A., and Yu, Z. (2025). False positives in study of memory-related gene expression. Nature *642*, E1–E3. https://doi.org/10.1038/s41586-025-08988-y.

5. Sun, W., Liu, Z., Jiang, X., Chen, M.B., Dong, H., Liu, J., Südhof, T.C., and Quake, S.R. (2025). Reply to: False positives in study of memory-related gene expression. Nature *642*, E4–E6. https://doi.org/10.1038/s41586-025-08989-x.

6. 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.

7. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140. https://doi.org/10.1093/bioinformatics/btp616.

8. Jiang, C., Wang, C., and Li, J.J. (2025). Nullstrap-DE: A General Framework for Calibrating FDR and Preserving Power in DE Methods, with Applications to DESeq2 and edgeR. Preprint at arXiv. https://doi.org/10.48550/arXiv.2507.20598.

9. Taylor, J., and Tibshirani, R.J. (2015). Statistical learning and selective inference. Proc. Natl. Acad. Sci. USA *112*, 7629–7634. https://doi.org/10.1073/pnas.1507583112.

10. Yu, B. (2013). Stability. Bernoulli *19*, 1484–1500. https://doi.org/10.3150/13-BEJSP14.

11. Upadhya, S., Klein, J.A., Nathanson, A., Holton, K.M., and Barrett, L.E. (2025). Single-cell analyses reveal increased gene expression variability in human neurodevelopmental conditions. Am. J. Hum. Genet. *112*, 876–891. https://doi.org/10.1016/j.ajhg.2025.02.011.

12. Levene, H. (1960). Robust tests for equality of variances. In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, I. Olkin, ed. (Stanford University Press), pp. 278–292.

Cell Systems, Volume *17*

Supplemental information

Beyond differential expression: Differential variability as an underappreciated axis in transcriptomics

Chenxin Flora Jiang, Changhu Wang, and Jingyi Jessica Li

Supplementary Material for “Beyond differential expression: differential variability as an underappreciated axis in transcriptomics”

Supplementary Figures

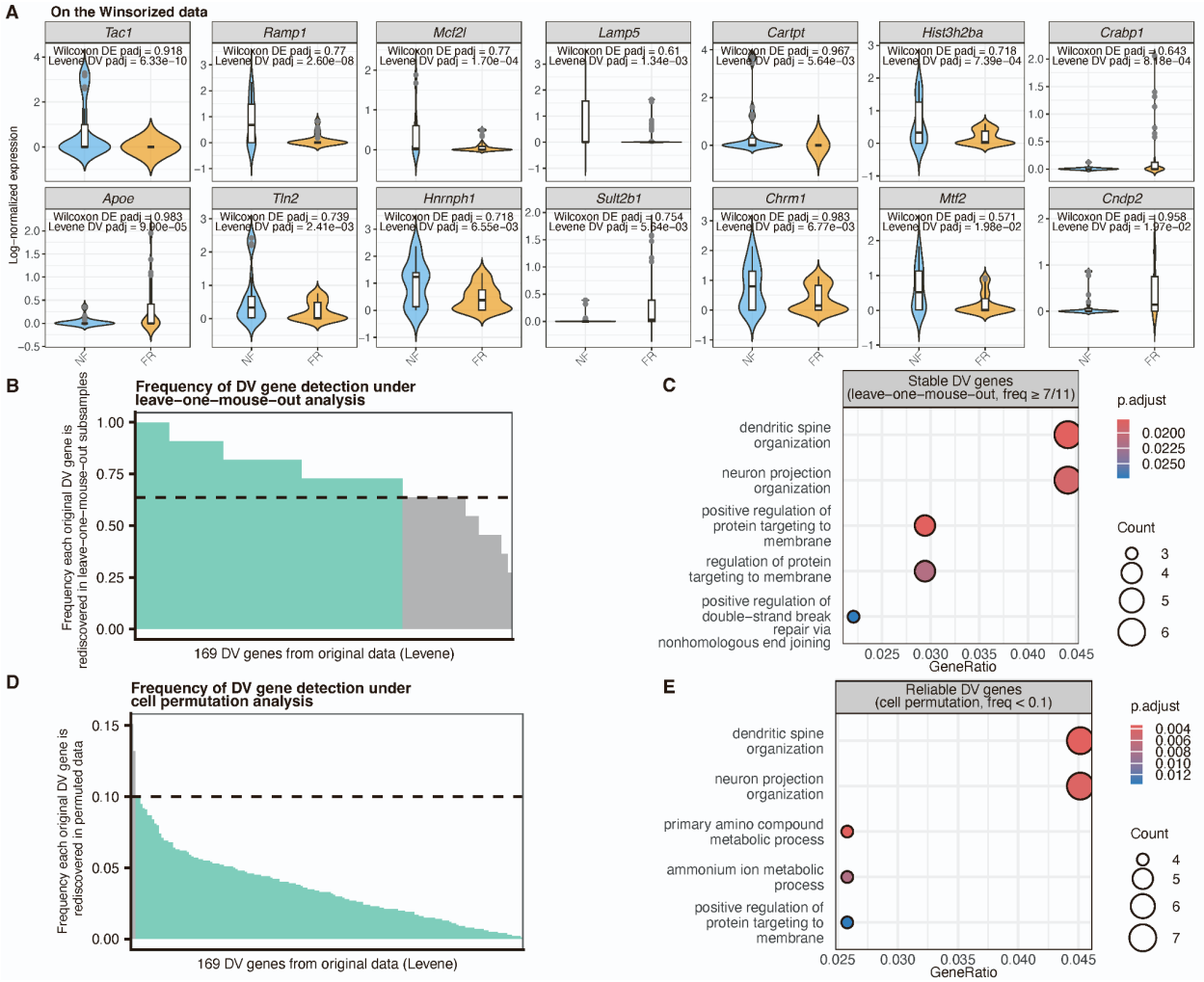

Figure S1 | A comparative analysis revealing the limitations of differential expression (DE) and the importance of differential variability (DV) in single-cell transcriptomic data analysis.
(A) Violin plots of DV genes identified by the Levene test that overlap with the significant genes reported by Sun et al.[1,2], after Winsorization at the 95th percentile within each condition. For each gene, the distributions of log-normalized expression levels in cells from the NF and FR conditions are shown. The BH-adjusted p-values for DE (Wilcoxon rank-sum test) and DV (Levene test) are indicated above each plot. Compared to **Figure 1I**, these patterns are preserved after Winsorization, indicating that the observed variability differences are not driven by outlier cells. **(B)** Frequency of DV gene detection under leave-one-mouse-out analysis. Each bar represents the proportion of subsamples in which a DV gene identified in the original analysis is rediscovered. The dashed line marks the stability threshold (detection frequency ≥ 7/11), with stable genes shown in cyan and others in gray. Most DV genes exceed this threshold, indicating robustness to the removal of individual mice. The threshold of 7/11 was chosen based on the number of mice with sufficient cell counts, excluding mice with only one cell for variance estimation. **(C)** GO enrichment analysis of DV genes stably detected across subsamples (frequency ≥

7/11). Enriched terms are primarily related to neuronal processes. **(D)** Frequency of DV gene detection under permutation analysis, where condition labels are permuted across cells. Each bar represents the proportion of permuted datasets in which a DV gene identified in the original analysis is rediscovered. The dashed line marks the stability threshold (detection frequency < 0.1), with stable genes shown in cyan and others in gray. The DV genes are rarely detected under permutation, indicating that the observed signals are unlikely to arise from random, condition-irrelevant variability. **(E)** GO enrichment analysis of reliable DV genes with low detection frequency under permutation (frequency < 0.1). Enriched terms are primarily related to neuronal processes.

References

1. Sun, W. *et al.* Spatial transcriptomics reveal neuron–astrocyte synergy in long-term memory. *Nature* **627**, 374–381 (2024).

2. Sun, W. *et al.* Reply to: False positives in study of memory-related gene expression. *Nature* **642**, E4–E6 (2025).