Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Differences in molecular sampling and data processing explain variation among single-cell and single-nucleus RNA-seq experiments

John T. Chamberlin, Younghee Lee, Gabor T. Marth, et al.

Genome Res. 2024 34: 179-188 originally published online February 14, 2024
Access the most recent version at doi:10.1101/gr.278253.123

References This article cites 46 articles, 10 of which can be accessed free at:
http://genome.cshlp.org/content/34/2/179.full.html#ref-list-1

Open Access Freely available online through the *Genome Research* Open Access option.

Creative Commons License This article, published in *Genome Research*, is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**

To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

© 2024 Chamberlin et al.; Published by Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Differences in molecular sampling and data processing explain variation among single-cell and single-nucleus RNA-seq experiments

John T. Chamberlin,[1] Younghee Lee,[1,2] Gabor T. Marth,[3] and Aaron R. Quinlan[1,3]

[1]*Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah 84108, USA;* [2]*Seoul National University, College of Veterinary Medicine, Seoul, 08826, South Korea;* [3]*Department of Human Genetics, Utah Center for Genetic Discovery, University of Utah, Salt Lake City, Utah 84112, USA*

A mechanistic understanding of the biological and technical factors that impact transcript measurements is essential to designing and analyzing single-cell and single-nucleus RNA sequencing experiments. Nuclei contain the same pre-mRNA population as cells, but they contain a small subset of the mRNAs. Nonetheless, early studies argued that single-nucleus analysis yielded results comparable to cellular samples if pre-mRNA measurements were included. However, typical workflows do not distinguish between pre-mRNA and mRNA when estimating gene expression, and variation in their relative abundances across cell types has received limited attention. These gaps are especially important given that incorporating pre-mRNA has become commonplace for both assays, despite known gene length bias in pre-mRNA capture. Here, we reanalyze public data sets from mouse and human to describe the mechanisms and contrasting effects of mRNA and pre-mRNA sampling on gene expression and marker gene selection in single-cell and single-nucleus RNA-seq. We show that pre-mRNA levels vary considerably among cell types, which mediates the degree of gene length bias and limits the generalizability of a recently published normalization method intended to correct for this bias. As an alternative, we repurpose an existing post hoc gene length–based correction method from conventional RNA-seq gene set enrichment analysis. Finally, we show that inclusion of pre-mRNA in bioinformatic processing can impart a larger effect than assay choice itself, which is pivotal to the effective reuse of existing data. These analyses advance our understanding of the sources of variation in single-cell and single-nucleus RNA-seq experiments and provide useful guidance for future studies.

[Supplemental material is available for this article.]

Single-nucleus RNA-seq (snucRNA-seq) is typically used as a substitute for single-cell RNA-seq (scRNA-seq) when dissociation of intact single cells is difficult. Although there are some gene-specific biological differences in mRNA localization and stability between the nucleus and cytoplasm (Fazal et al. 2019), the primary obstacle identified by initial studies was excessive data sparsity in snucRNA-seq. Nuclei contain a minority of total cellular mRNA but are inherently enriched for pre-mRNAs (Fig. 1A), yielding fewer total exonic read alignments and a higher fraction of intronic read alignments (Habib et al. 2017; Hu et al. 2017). Early publications found that the lower transcript yield from nuclei could be mitigated by incorporating these intronic reads during gene expression quantification (Lake et al. 2017, 2018; Bakken et al. 2018; Wu et al. 2019; Selewa et al. 2020). Other studies extended this strategy to scRNA-seq to level comparisons with single-nucleus experiments (Mereu et al. 2020; Selewa et al. 2020). For the widely used CellRanger preprocessing software, this initially required use of a specific option, but the tool has since been updated to incorporate intronic reads by default for both cell and nuclear data sets (https://www.10xgenomics.com/support/software/cell-ranger/latest/release-notes/cr-release-notes#v7-0-0). Although there are some minor changes in the assignment of alignments owing to overlaps between exons and introns of different genes (Soneson et al. 2021), the overt impact is an increase in total detected transcripts in the range of ∼10% to >100% (https://support.10xgenomics.com/single-cell-gene-expression/sequencing/doc/technical-note-interpreting-intronic-and-antisense-reads-in-10x-genomics-single-cell-gene-expression-data). The paradigm shift to universal incorporation of intronic reads is of central importance to downstream analysis of new or existing data because mRNA and pre-mRNA tend to be captured by distinct sampling mechanisms with accompanying biases. RNAs are primarily reverse transcribed from the poly(A) tail in a transcript length–independent manner (Phipson et al. 2017), whereas pre-mRNAs are subject to a gene length–associated bias via *internal priming* (La Manno et al. 2018; Thrupp et al. 2020; Eraslan et al. 2022; Svoboda et al. 2022; Gorin and Pachter 2023b). This term refers to the hybridization of oligo(dT) primers to internal adenosine homopolymers, which frequently occur in introns and covary strongly with gene length (Fig. 1B). Because each priming event introduces a new molecular barcode (unique molecular identifier [UMI]), the chance of capturing and counting a pre-mRNA transcript, as well as the potential for multiple counting of a single transcript, increases with gene length. Overall gene length bias, or the degree to which measured abundance covaries with gene length, is greater in nuclei than in cells because they contain proportionally more pre-mRNA (Fig. 1C). We note that the change in preprocessing does not always coincide with an mRNA versus pre-mRNA dichotomy, as some exonic reads are produced by pre-mRNAs. This may explain why a mild gene length effect has also

Corresponding author: aaronquinlan@gmail.com
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.278253.123. Freely available online through the *Genome Research* Open Access option.

© 2024 Chamberlin et al. This article, published in *Genome Research*, is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

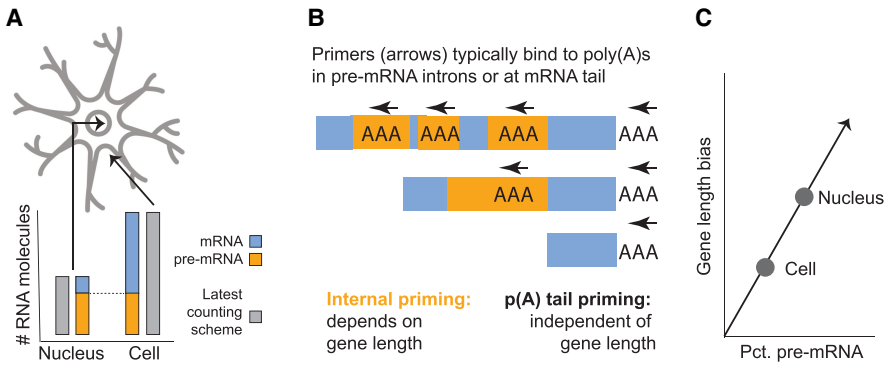

Figure 1. Conceptual depictions of key technical aspects of sc/snucRNA-seq. (*A*) Nuclei contain less total RNA but are enriched for pre-mRNA compared with cells. All pre-mRNAs in the nucleus are also in the cell, but most mRNAs are outside the nucleus. (*B*) Pre-mRNA tends to generate intronic reads via internal priming, which is biased in a gene length–associated matter. mRNA generates exonic reads from priming at the poly(A) tail, irrespective of length. Gray arrows indicate the position and direction of reverse transcription initiation by poly(dT) primers. (*C*) Gene length bias, or the degree to which long genes are overestimated by the assay, is a function of pre-mRNA content. Nuclei contain relatively more pre-mRNA than equivalent cells.

been reported from nuclei using the "exon-only" counting scheme (Kuo et al. 2023).

Whereas a number of studies have simply noted the excess of gene length bias in snucRNA-seq (Lake et al. 2018; Selewa et al. 2020; Eraslan et al. 2022), others have advanced incompatible or inconsistent recommendations to account for it in different types of analyses. First, Thrupp et al. (2020) argued that snucRNA-seq is fundamentally "not suitable" as a substitute for scRNA-seq in part owing to gene length bias. However, they did not comment on the role of internal priming in pre-mRNA capture, nor did they control for differences in the inclusion of intronic reads between experiments. More recently, Svoboda et al. (2022) advocated for excluding internally primed alignments to improve concordance with bulk RNA-seq, whereas Gupta et al. (2022) proposed a normalization scheme that scales down the contribution of pre-mRNA transcript counts based on gene length to improve the similarity between matched single-cell and single-nucleus samples. Critically, neither group showed the generalizability of recommendations across cell types. This is important because pre-mRNA recovery, or implicitly, gene length bias, is known to be highly tissue specific (Hu et al. 2017, 2018). As such, the potential impact of the inclusion, exclusion, and/or normalization of pre-mRNA data is also highly variable. Moreover, the conventional bioinformatic workflow does not separately report exonic and intronic gene abundances in the output gene expression matrix, so additional steps are needed to infer the level and impact of pre-mRNA across individual cell types.

In summary, the presence and effect of variation in pre-mRNA sampling bias across cell types both within and between assays is obscured by standard analysis practices. Previous efforts have generally failed to disentangle the effect of assay choice (cell or nucleus) and bioinformatic processing (quantification with or without intronic reads), such that best practices in data analysis have not been established. The goal of this study is to address these gaps by reanalyzing existing experiments including a large data set from the mouse cortex (Yao et al. 2021) and human microglia data generated by Thrupp et al. (2020). We quantify gene expression with and without intronic alignments and reimplement the Gupta et al. gene length–correction method to evaluate the effect of pre-mRNA content and bioinformatic choices across cell types. We focus on nervous tissues because these are known to show the highest levels of pre-mRNA content (Cao et al. 2020) and are frequently assayed as single nuclei, particularly in human studies. We also revisit data and metadata from an atlas of human fetal organs (Cao et al. 2020) to position our results in a broader context. Finally, we explore potential post hoc solutions for improving biological interpretation of differences among cell types and experiments.

Results

Incorporating pre-mRNA data improves similarity of cells and nuclei despite gene length bias

We focused our analysis on a large data set from the mouse motor cortex (Yao et al. 2021). To minimize potential batch effects, we limited our analyses to only the single-cell and single-nucleus data generated by the Allen Institute for Brain Science using 10x Genomics V3 (Zheng et al. 2017) assay chemistry. We also reanalyzed human microglial data (10x Genomics V3) generated by Thrupp et al. (2020) in order to reevaluate their conclusions with regard to bioinformatic processing choices. We aligned and quantified the sequence data, with and without intronic reads, from both studies with STARsolo (Methods) (Kaminow et al. 2021). Finally, we analyzed preprocessed data from a single-cell atlas of human fetal development to understand the range of pre-mRNA recovery across tissues and its contribution to gene length bias. Because the nomenclature for quantification schemes is not standardized in the literature, herein we use the terms *Exon* (UMI counts from exonic alignments alone), *Intron* (UMI counts not from exonic alignments), and *Intron&Exon* (UMI counts from alignments in exons or introns). We inferred *Intron* counts by subtracting *Exon* counts from *Intron&Exon*. Concretely, the "intron fraction" is the fraction of total counts that were contributed only by intronic alignments, which we assume to be approximately proportional to the true pre-mRNA content. We also reimplemented the Gupta et al. method, which we refer to as *length-corrected* counts. This entails dividing *Intron* counts for each gene by the expected number of internal priming motifs in the gene (15 or more consecutive A's) given its length (Methods).

To explore the impact of bioinformatic choices on estimated RNA abundances (UMI counts per million [CPM]), we began with L5 IT neurons, the most common cell type in the data set. Gene length bias (the correlation between average gene abundance and gene length) was stronger in nuclei than in cells when using *Intron&Exon* quantification, consistent with the difference in intron fraction ($R = 0.44$ vs. $R = 0.29$, 66% vs. 40% intron fraction) (Fig. 2A). The gene length effect was also evident in a between-assay differential expression test, in which the genes significantly more abundant in nuclei were statistically longer than those in cells (fold change >1.5, adjusted P-value < 0.05) (Fig. 2B). *Exon* counts showed a milder gene length bias ($R = 0.21$ and $R = 0.13$), whereas the bias was strongest but most similar with *Intron* counts ($R = 0.47$ in both) (Supplemental Fig. S2A). Nevertheless, the net similarity

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

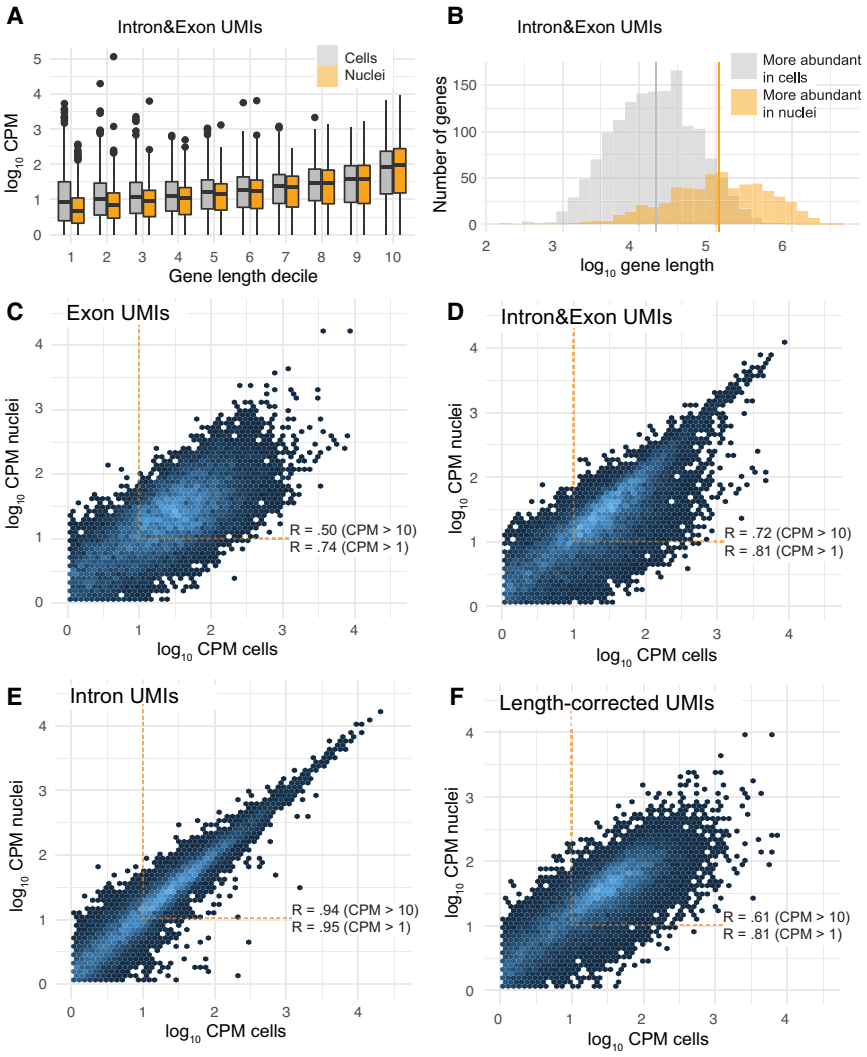

Figure 2. RNA sampling properties explain the effect of bioinformatic processing on assay similarity. (*A*) Gene length bias is evident in averaged gene abundances in both L5 IT cells and nuclei under the *Intron&Exon* quantification but is stronger in nuclei. Pearson's correlations between $\log_{10}$ (mean CPM) and gene length are R = 0.51 in nuclei and R = 0.29 in cells. (*B*) The gene length distribution of genes that are significantly more abundant (fold change >1.5, adjusted *P*-value < 0.05) either in L5 IT cells (gray) or in nuclei (orange). Mean $\log_{10}$ gene lengths are 5.0 versus 4.2, $P < 2.2 \times 10^{-16}$ (*t*-test on $\log_{10}$ lengths). (*C*) Hexbin plot showing the correlation of *Exon* abundances between L5 IT cells and nuclei. Pearson's correlations are computed on $\log_{10}$ (mean CPM across all cells or nuclei) for genes above one or 10 mean CPM in both assays. (*D*) *Intron&Exon* abundances are more strongly correlated and show fewer total differences. (*E*) The correlation of *Intron* abundances is very high, consistent with pre-mRNA localization within the nucleus, which is within the cell. (*F*) *Length-corrected* abundances are no better correlated than the baseline result. Total differences increase, consistent with the worsened correlation among more highly expressed genes. The length-correction method depresses *Intron* counts, which indirectly amplifies the prominence of *Exon*.

between cells and nuclei, in terms of the correlation between average gene abundance ($\log_{10}$ of mean CPM), was better with *Intron&Exon* (R = 0.81) than with *Exon* (R = 0.74) (Fig. 2C,D), and the number of differentially abundant genes decreased by ~40% (Supplemental Fig. S2B). Data from the human fetal atlas reinforced the role of pre-mRNAs in gene length bias, as the tissues with higher intronic read fraction showed a stronger correlation between average gene abundance and gene length (Supplemental Fig. S1).

To understand the basis of the improvement, we compared *Intron* counts between cells and nuclei, which revealed a very strong correlation of 0.94 and ~8% as many differentially abundant genes (Fig. 2E). This indicates that pre-mRNA sampling is much more similar than mRNA between nuclei and cells, consistent with the subcellular localization patterns of the two RNA species: all pre-mRNAs are inside the cell, but not all mRNAs are inside the nucleus. This contrasts with previous studies that suggested that increased library size from *Intron* counts helped to mitigate differences between cells and nuclei (Wu et al. 2019). However, upon down-sampling the *Intron&Exon* counts to the same depth as the *Exon* counts, the Pearson's correlation between average gene abundance in cells and nuclei remained R = 0.81, indicating that the improvement over *Exon* is not also owing to a decrease in sparsity. We also note that the library sizes differ between the two assays: 40,000 UMIs/cell and 10,000/nucleus using *Intron&Exon* and 24,000 UMIs/cell and 3500 UMIs/nucleus using *Exon*. We down-sampled the cells to 10,000 UMIs/cell and repeated the comparison and, again, found that the correlation remained roughly 0.81.

The Gupta et al. length-correction method is intended to improve upon *Intron&Exon* similarity by reducing gene length bias. Gupta et al. (2022) reported a ~40% decrease in differentially abundant genes and an increased correlation in average gene abundance from 0.5 to 0.6 in a comparison of human preadipocyte cells to nuclei. However, we found the opposite effect in L5 IT neurons (Fig. 2F): Differential genes increased from 492 to 654, and correlation did not improve. We attribute this to large differences in pre-mRNA enrichment between preadipocytes and L5 IT neurons. In preadipocytes, the intron read fraction was both lower overall and more different between nuclei and cells (40% vs. 9%, 4.4-fold). In L5 IT neurons, the levels were 66% versus 40%, an enrichment ratio of just 1.5-fold. Given that length correction reduces *Intron* but not *Exon* counts, we conclude that reducing gene length bias is counteracted by an increased emphasis on mRNA differences in a sample-specific manner.

Variation in pre-mRNA recovery explains bias within and between assays

We next extended the analysis to the remaining mouse cortex cell types and to the human microglia data from Thrupp et al. (2020). We were surprised to find that a higher intron fraction in nuclei did not necessarily equate to a higher fraction in cells of the

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

same annotated type ($P = 0.45$) (Fig. 3A). In other words, the relative enrichment of pre-mRNA between nuclei and cells varies across cell types, such that the magnitude of gene length bias between assays is unpredictable a priori. Using *Intron&Exon* counts from nuclei and cells for each cell type, we found that the count of differentially abundant genes between assays increased with the ratio of intron content between cortex nuclei and cells (Spearman's $\rho = 0.92$) (Fig. 3B). Human microglia, despite representing a different species and study, clustered among the other nonneuronal cell types from the mouse data set (Fig. 3B, yellow diamond), suggesting that the effect of differences in pre-mRNA enrichment between cells and nuclei is not unique to the mouse cortex data set. The value reported by Gupta et al. (2022) (shown as a white diamond) appears as an outlier but is also consistent with the general trend.

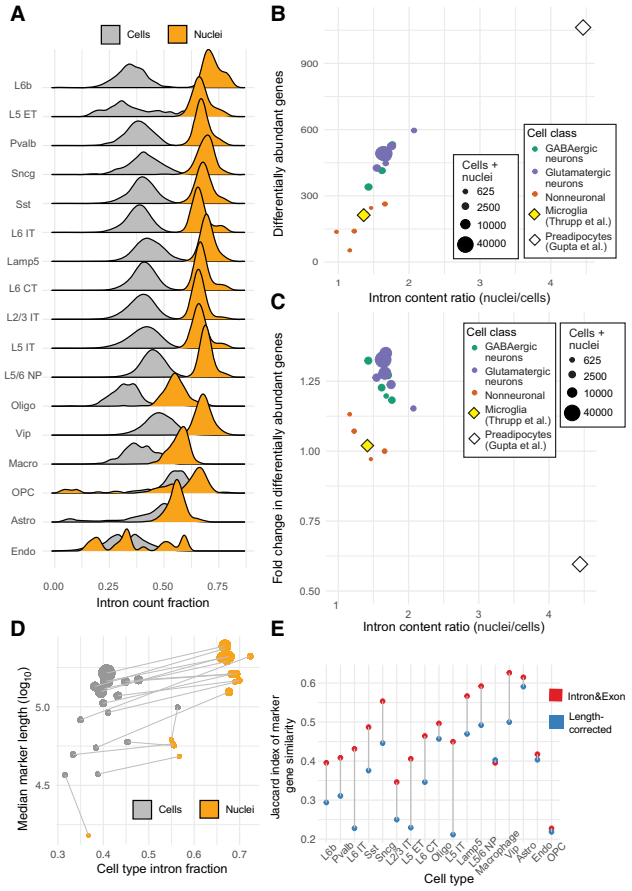

Figure 3. Variation in pre-mRNA enrichment moderates assay similarity with and without gene length normalization. (*A*) Intron content distribution (total *Intron* counts divided by total *Intron&Exon* counts per cell or nucleus) for each cell type in the mouse cortex. Mean intron content in cells does not increase with mean intron content in nuclei (Spearman's $\rho = 0.19$, $P = 0.46$). Nonneuronal cells are abbreviated: (astro) astrocytes, (endo) endothelial cells, (oligo) oligodendrocytes, (OPC) oligodendrocyte precursor cells, and (macro) macrophages. The remaining labels specify types of neuronal cells. (*B*) Differential expression testing of cells versus nuclei for each cell type. The number of differentially abundant genes (FC > 2) increases with the ratio of mean intron content in the cell type (Spearman's $\rho = 0.89$, $P = 9.9 \times 10^{-7}$). Cells are colored by cell type class. (*C*) Fold change in the number of differentially abundant genes for cells versus nuclei after applying the Gupta et al. length-correction procedure. The white diamond shows the result reported by Gupta et al. (2022) for white preadipocytes. Values greater than one indicate an adverse performance of the length-correction method. Linear modeling (of cortex cells only) identifies cell class and intron content ratio as significant predictors of method performance ($R^2 = 0.78$, $P = 0.0001$). (*D*) Marker genes discovered from nuclei tend to be longer than those from cells of the same type, except for in very rare cell types. (*E*) Marker gene similarity (Jaccard index) for the top 50 markers (by fold change) is variable but not in relation to the number of cells or the intron content ratio. Columns are ordered by total (cells + nuclei).

The gene length–correction method from Gupta et al. (2022) was generally ineffective, as the number of differentially abundant genes between cells and nuclei increased instead of decreased for human microglia and 15 of the 17 mouse cortex cell types. (Fig. 3C; for correlations, see Supplemental Fig. S3). We emphasize that these results are not incompatible with the values reported by Gupta et al. (2022), in which the pre-mRNA enrichment (i.e., intron content ratio) between preadipocyte nuclei and cells was much stronger than any cortex type. Instead, it suggests that the utility of their method is likely to be limited to cell types in which gene length bias contributes more strongly than mRNA-level bias to inter-assay differences, that is, when pre-mRNA recovery is highly different between nuclei and cells.

It is also important to address how these mechanisms translate to downstream analyses that would be performed in a typical single-assay experiment. We performed marker gene analysis to identify genes with significantly higher expression in each cell type within a sample. Systematic gene length bias was again evident, as marker genes were consistently longer in nuclei than in cells except in the rarest cell types (Fig. 3D). We compared the similarity of the top 50 marker genes between the two assays for each cell type using the Jaccard similarity index, or the fraction of total markers discovered by both. Unlike the previous pairwise analysis, marker similarity was not correlated with pre-mRNA enrichment; this is likely because marker gene testing is a function of all cell types in aggregate. Gene length correction slightly worsened the marker gene Jaccard similarity for 15 of 17 cell types, consistent with the observed decrease in direct similarity (Fig. 3E). Finally, we confirmed that the magnitude of gene length bias was stronger in cell types with a higher intron fraction (Supplemental Fig. S4A) and that the elevated intron fraction was not a result of gene-specific effects (Supplemental Fig. S4B).

A strategy for post hoc correction of gene length bias

Increased sampling of pre-mRNA from longer genes results in increased statistical power to detect differential expression in these genes and vice versa. This phenomenon is similar to the fragmentation bias in conventional RNA-seq, in which longer transcripts yield more fragments that can be sequenced and, in turn, more stable expression estimates, irrespective of scaling. In the context of gene set enrichment analysis, this was addressed by the GOseq R package published in 2010 (Young et al. 2010). GOseq models the relationship between the bias term (in our case, gene length) and the chance of appearing differentially expressed (Fig. 4A) in order to downweigh the prominence of GO categories that happen to contain unusually long genes and vice versa. As a proof of principle, we applied the algorithm to the set of genes that were significantly more abundant in L5 IT nuclei compared with cells (those shown in Fig. 2B). We expect these genes to reflect true biological differences in RNA localization between nuclei and cytoplasm. Without bias correction, the most overrepresented categories were terms indicative of neuronal function, such as "synapse organization," which is not biologically informative given that these cells and nuclei are ostensibly the same cell type. After correcting

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

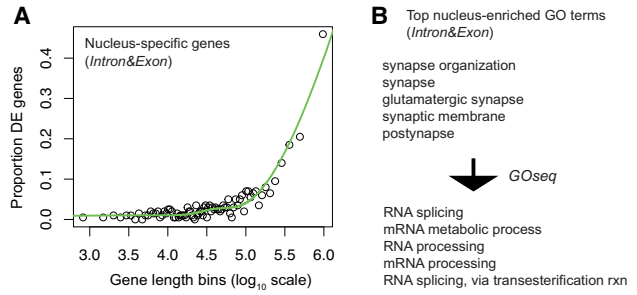

Figure 4. Approaches for measuring and mitigating sampling bias. (*A*) GOseq depicts and models the relationship between a bias term (gene length) and the chance of appearing differentially expressed. Points represent bins of 300 genes each on $\log_{10}$ scale. (*B*) After correcting for gene length, GO terms enriched in nuclei are consistent with known patterns of RNA localization.

for gene length, the top terms were consistent with nuclear patterns observed from RNA localization assays, such as "RNA splicing" (Fig. 4B; Supplemental Table S1; Fazal et al. 2019): Fazal et al. (2019) reported that "mRNAs enriched in nuclear locations tend to code for proteins enriched in nuclear speckles and nucleoplasm." A similar pattern was seen by Bakken et al. (2018) using a technically distinct scRNA-seq assay. Many of the same terms were also produced from *Exon* differential expression and from the length-corrected counts, without applying GOseq bias correction (Supplemental Table S1). This shows that gene length bias can be modeled and accounted for post hoc to achieve a biologically meaningful interpretation without manipulation of raw gene counts and without discarding intronic reads. Next, we tested GOseq on marker genes from the cortex samples (from Fig. 3D). Here, we also measured the change in rank of GO categories according to the average gene length, similar to the original GOseq publication. We began with oligodendrocytes, as these showed a large number of marker genes and have a well-defined physiological role. As expected, the corrected rank of GO categories shifted according to average gene length (Supplemental Fig. S5). Of the top five categories, GOseq caused "myelin sheath" to increase from rank 3 to rank 1 and "myelination" to enter at rank 3, which coincides with the role of oligodendrocytes in myelin formation. However, the change is less stark than in Figure 4B, likely because there are stronger biological differences between cell types than between cells and nuclei of the same type. We also found qualitative improvement harder to interpret in most cell types. In any case, we argue that these examples support the use of GOseq in snucRNA-seq analysis for measuring and potentially accounting for gene length bias, especially in combination with domain-specific expertise. Complete marker gene GOseq results are included in Supplemental Table S2.

Bioinformatic processing can have a larger role than assay choice

The prior analysis performed comparisons within quantification schemes, which does not address scenarios in which data are preprocessed discordantly. This includes the historical convention in which cells were quantified as *Exon* and nuclei as *Intron&Exon* but could also occur if new single-cell data are compared to older, already-processed data (Fig. 5A). Thrupp et al. (2020) is a case of the former. Their study is a response to the apparent failure of human snucRNA-seq studies (Del-Aguila et al. 2019; Grubman et al. 2019; Mathys et al. 2019) to reproduce results from two earlier scRNA-seq studies in mouse models (Keren-Shaul et al. 2017; Sala Frigerio et al. 2019). Specifically, they show that two gene sets purported to correspond to microglial activation in mouse models of Alzheimer's disease are significantly depleted in normal human microglial nuclei. However, the motivating studies were analyzed under their contemporary paradigms: scRNA-seq as *Exon* and snucRNA-seq as *Intron&Exon*. In effect, this maximizes potential differences because the results reflect both assay and bioinformatic effects. Thrupp et al. (2020) analyzed their newly generated data under the unified *Intron&Exon* method, which means they were only able to address the role of the assay.

We performed additional differential expression tests using Thrupp et al. data, but with different combinations of preprocessing. In the discordant test (cells as *Exon*, nuclei as *Intron&Exon*), we observed 1119 differentially abundant genes (fold change >2), whereas *Intron&Exon* in both assays yielded just 213 (Fig. 5B). To understand the role of the quantification method in isolation, we also compared *Exon* to *Intron&Exon* within cells and nuclei separately. This yielded 566 and 579 genes that significantly changed in response to preprocessing in cells and nuclei, respectively. In other words, preprocessing had a larger effect on apparent differential gene expression than the assay itself in the Thrupp et al. microglia data set. This suggests that the original failure to reproduce scRNA-seq results in nuclei could be explained by the change in preprocessing rather than assay. Next, we used GOseq to assess the role of gene length on the reported depletion of the "disease-associated microglia" (DAM) gene set (Keren-Shaul et al. 2017) in nuclei. We first replicated the report of Thrupp et al. (2020) that the DAM genes tend to be significantly more abundant in cells when using *Intron&Exon* quantification at a more than twofold change threshold: 40 DAM genes were more abundant in cells versus none in nuclei. However, GOseq correction did not remove the strong statistical enrichment of DAM genes ($P = 1.6 \times 10^{-37}$ without correction and $P = 1.9 \times 10^{-31}$ with correction). This suggests that although DAM genes do tend to be shorter than average, enrichment in cells is not substantially explainable by gene length bias. Next, we compared cells quantified as *Exon* to the same cells quantified as *Intron&Exon*. At a fold change threshold of two or more, only seven DAM genes were represented in either direction. However, we also tested a relaxed threshold of more than 1.5-fold change in order to evaluate possible differences in effect size between preprocessing and assay. In this case, we found that the DAM genes were strongly enriched in "*Exon*"-specific genes. We then repeated the analysis of cell-specific genes (vs. nuclei) at the same 1.5-fold threshold (see Fig. 5C,D) and found that DAM genes appear to be significantly depleted in response to both a change in assay (from cells to nuclei) or a change in preprocessing (from *Exon* to *Intron&Exon*). This indicates that preprocessing also has a strong directional effect on the measured abundance of DAM genes, albeit at less than a twofold effect size. Full GOseq results for this analysis are found in Supplemental Table S3.

These results have two important implications. First, the DAM gene signature likely would have contained different genes if the original study had used *Intron&Exon* preprocessing. Repeating the original analysis that defined these genes but with *Intron&Exon* quantification would be illuminating. Second, although the DAM genes do appear to be less abundant in nuclei independent of length, the poor replicability of the DAM gene signature in the initial follow-up studies can be partially explained by the change in data preprocessing. In other words, follow-up single-cell studies (were they to exist) would likely have also shown reduced microglial activation simply owing to the change in

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

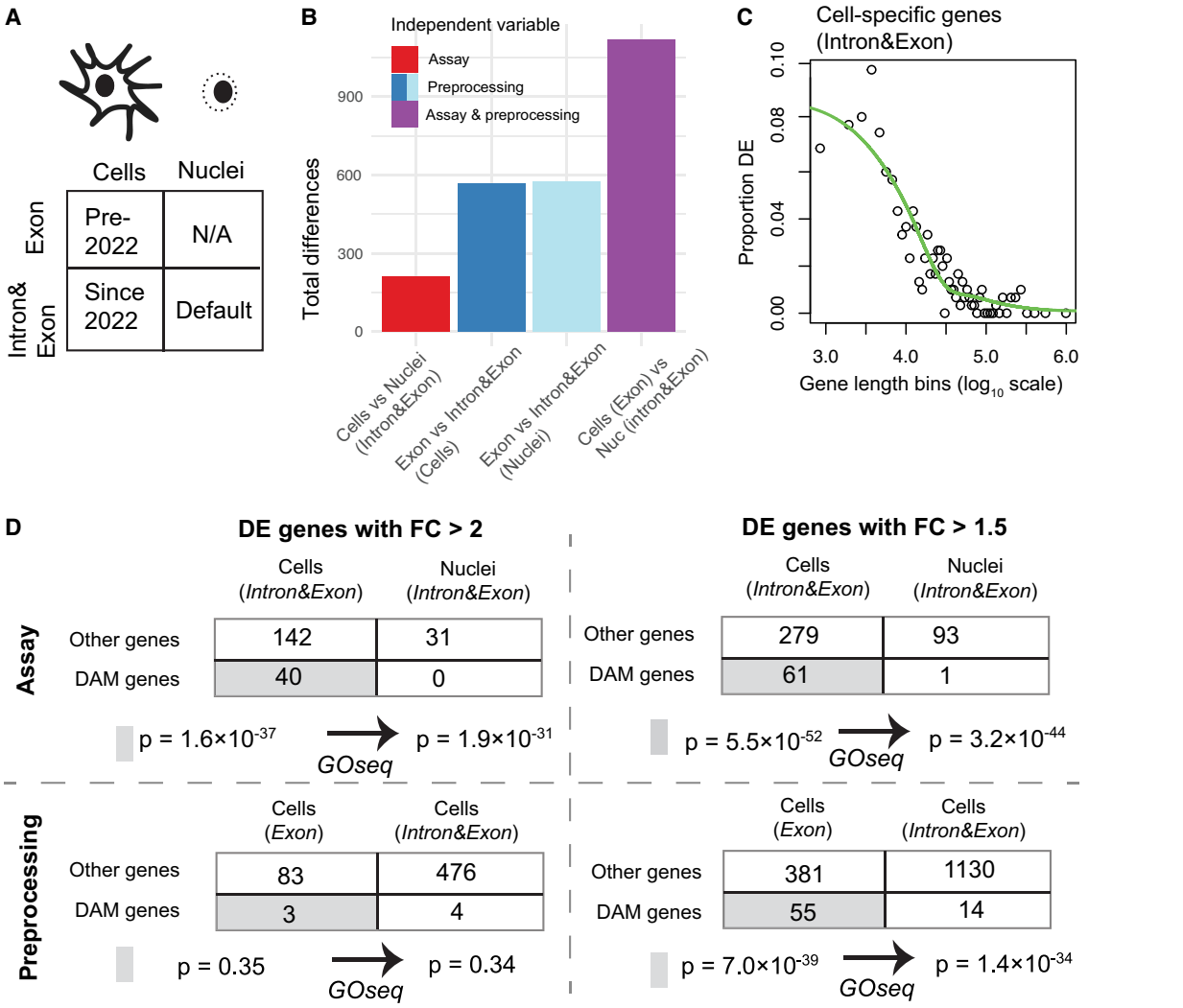

Figure 5. Impact of differential bioinformatic preprocessing on comparison within and between experiments. (*A*) Recommended bioinformatic processing of scRNA-seq and snucRNA-seq has converged over time according to CellRanger documentation. (*B*) The processing method (blue bars) has a larger effect on differential expression testing than does assay choice (red bar) in microglia. Historical paradigm in which nuclei and cells use two different processing methods maximizes apparent differences (purple). (*C*) GOseq analysis of genes significantly more abundant in microglial cells than nuclei with *Intron&Exon* quantification (FC > 1.5). Genes are grouped into bins of size 300 based on their length. Note that the correlation is negative in this instance because cell-specific genes are depicted rather than nucleus-specific genes. (*D*) GOseq analysis of genes significantly more abundant (FC > 1.5 or FC > 2) in response to a change in assay (*top*) or preprocessing (*bottom*). In either case, DAM genes are enriched at the FC > 1.5 threshold, and GOseq reduces but does not remove the strong statistical significance. At FC > 2, DAM genes are not significantly enriched in response to preprocessing.

preprocessing. However, additional technical factors, including sample size or sequencing depth, may also contribute. For example, a metanalysis described by Martins-Ferreira et al. (2023) in a recent preprint reported successful identification of DAM genes in microglial nuclei aggregated across multiple studies with much larger sample sizes.

In summary, we have shown that failing to control for data preprocessing confounds the interpretation of scRNA-seq and snucRNA-seq and that GOseq is useful for disentangling the effect of data preprocessing from systematic sampling bias. Because of the shift in default behavior of CellRanger, the risk of repeating this type of error is likely to increase. Although the incompatibility is acknowledged in principle by the software developers, it is incumbent upon the analyst to properly account for it in practice. Our analysis also shows the utility of quantifying both *Exon* and *Intron&Exon* counts, and we encourage modifying the standard workflow to use both quantification strategies.

Discussion

Rapid development in single-cell and single-nucleus sequencing technologies, algorithms, and data workflows have complicated the establishment of best practices in experimental design and analysis. In this study, we have addressed the roles of assay choice (cell or nuclear preparation) and data preprocessing method (quantification with or without intronic reads, or gene length correction) on the consistency of results within and between data sets. We confirm that inclusion of intronic reads is useful for improving concordance between cells and nuclei, despite the introduction of gene length bias. However, the basis of this improvement is

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

complex. In isolation, pre-mRNA measurements are more similar than mRNA measurements, reflecting patterns of synthesis, localization, and degradation within the cell. Differential enrichment of pre-mRNAs between cells and nuclei then contributes to apparent differences via gene length bias. More fundamentally, these results show that bioinformatic choices depend on the specific aims of the analysis, particularly when comparing new to existing data. For example, previous studies have shown that incorporating intronic reads does not improve similarity to bulk RNA-seq (DelAguila et al. 2019), which suggests that the procedure may not be helpful for tasks such as the bulk sample deconvolution attempted in a recent preprint (Park et al. 2021). Similarly, the high consistency of pre-mRNA measurements implies that the best way to minimize apparent differences between cells and nuclei would be to exclude mRNA data entirely. This is obviously an untenable recommendation and suggests that orthogonal approaches are needed for evaluating preprocessing methods.

Collectively, these findings lead us to a series of recommendations. Foremost, the quantification method must be controlled for, as newly generated data are likely to be incompatible with previously published data sets. For example, an analysis of GTEx data by Eraslan et al. (2022) compared new snucRNA-seq to existing scRNA-seq without controlling for differences in preprocessing, although this had little bearing on their primary analysis. Fortunately, modifying the bioinformatic workflow to quantify gene expression both with and without introns is straightforward with contemporary tools, including CellRanger (Zheng et al. 2017), alevin (Srivastava et al. 2019), kallisto (Melsted et al. 2021), and STARsolo (Methods) (Kaminow et al. 2021). Alternatively, the spliced and unspliced counts already generated by the popular Velocyto method (La Manno et al. 2018) can be substituted. We anticipate that the conclusions will be broadly similar, namely, that the "unspliced" fraction is closely correlated with the "intron" fraction; nuances in these alternatives have been discussed previously (Soneson et al. 2021). We also suggest sharing both *Exon* and *Intron&Exon* matrices as these files are much smaller than the raw sequencing data. Separating *Exon* and *Intron* counts has also been used for other steps in single-cell analysis, such as ambient RNA decontamination (Alvarez et al. 2020), and in a recent preprint that describes mechanistic models of differential expression (Gorin and Pachter 2023a).

Although the high absolute levels of pre-mRNA in nervous tissue cells and nuclei appear to preclude the use of the Gupta et al. length-correction method, performance in other tissues is likely to be positive. Conversely, tissues such as muscle and heart tend to yield very little pre-mRNA, leading to a low level of gene length bias, which may not warrant correction. However, the variable pre-mRNA enrichment evident across cortex cell types indicates that the length-correction method will require validation with matched single-cell and single-nucleus data sets before using it in new contexts. At the same time, the role of pre-mRNA sampling in reducing apparent differences between assays does nothing to ameliorate gene length bias within a given assay. For example, marker genes from nuclei or high pre-mRNA tissue types will be artifactually longer than markers from low pre-mRNA samples. As such, we suggest use of GOseq to measure and correct for this bias when performing gene set enrichment analysis.

Our analysis is subject to a few limitations. Foremost, we relied on the mouse cortex cell type annotations generated by Yao et al. (2021) as these were based on the integrated analysis of seven data sets. Further effort is needed to assess the impact of data preprocessing and pre-mRNA content levels on cell type assignment. 10x Genomics have argued that the change in cell type assignment after incorporating introns is negligible in PBMCs (https://www.10xgenomics.com/support/single-cell-gene-expression/documentation/steps/sequencing/interpreting-single-cell-gene-expression-data-with-and-without-intronic-reads), whereas Kuo et al. (2023) did observe differences among nuclei. However, the lack of ground truth labels prevents us from including a proper benchmark. Related work on ambient RNA contamination has found that marker gene analysis is much more sensitive than cell type assignment (Janssen et al. 2023). Also, we limited our main analysis to data generated with the 10x Genomics 3′ platform. Although this is currently the most popular option, a number of alternatives are available, and platform-specific differences in pre-mRNA sampling are possible. For example, a recent preprint reported increased intronic read fraction and stronger gene length bias in a new single-cell platform developed by Parse Biosciences (Xie et al. 2023).

Additional limitations pertain to the Gupta et al. method, and we did not alter the parameterization of their approach. In particular, the definition of an internal priming site is specified arbitrarily, but it directly determines the reduction in total *Intron* counts. Currently, *Intron* counts are divided by the expected number of A(15) motifs in the gene, given its length. Relaxing the motif definition would result in a substantially greater reduction in total counts and vice versa. *Exon* and *Intron* quantification also do not perfectly correspond to pre-mRNA and mRNA or their capture modes, as described in a preprint by He et al. (2023), nor is gene length bias fully explainable by internal priming sites alone (Kuo et al. 2023). We speculate that the Gupta et al. length-correction method could be improved by instead dividing counts into poly(A) tail and internally primed, deriving the scaling factor empirically, and flooring the scalar at one to avoid variance inflation of short genes. As a more direct estimation of pre-mRNA content, binarizing by priming site type approach would also be useful for understanding the technical and/or biological factors contributing to variation across cell types and to gene length bias in *Exon* counts. This could reflect differences in RNA processing time among genes or reflect additional technical artifacts. Beyond computational solutions, these issues would likely be substantially diminished through improved single-nucleus protocols that retain more nuclear-associated mRNAs during dissociation (Drokhlyansky et al. 2020).

The description of gene length bias in single-cell and single-nucleus sequencing is not a novel report, but solutions remain almost absent in practice. Meanwhile, discussions have tended to emphasize open questions in so-called "downstream" analysis steps of single-cell data science (Lähnemann et al. 2020; Heumos et al. 2023). However, upstream and downstream analyses are clearly not independent tasks, and further effort and attention are needed to establish unifying principles of data analysis and interpretation as the field continues its rapid expansion (Svensson et al. 2020).

Methods

sc/snucRNA-seq data preprocessing

We analyzed the 10x Genomics V3 data subset from Yao et al. (2021) generated at the Allen Institute for Brain Science; a complementary data set generated at the Broad Institute was excluded because it included nuclei only. Raw sequence data and metadata

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

were downloaded from http://data.nemoarchive.org/biccn/lab/zeng/transcriptome/. Thrupp et al. data were acquired from the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) accession numbers GSE153807 and GSE137444. Preprocessed data and metadata from the human fetal atlas study (Cao et al. 2020) were downloaded from NCBI GEO accession number GSE156793.

Alignment and gene expression quantification were performed using STARsolo version 2.7.3a (Kaminow et al. 2021) with the option "--soloFeatures Gene GeneFull," which corresponds to the *Exon* and *Intron&Exon* schemes. We used the GRCm38 and GRCh38 reference genomes for the mouse and human samples, respectively, to be consistent with the versions used in the original analyses of these public data sets and with the standard reference packages provided by 10x Genomics. We used Ensembl version 99 gene annotations filtered by biotype based on 10x Genomics guidelines and the "cellranger mkgtf" utility. Because of erroneous resequencing of some original mouse cortex libraries by Yao et al. (2021), we used seqtk (https://github.com/lh3/seqtk) to filter out read pairs with truncated barcodes (shorter than the intended 28 bp), as the mixture of barcode read lengths (26 bp and 28 bp) was incompatible with STARsolo. Resulting abundance estimates were nearly identical to CellRanger values provided by the investigators, which is expected by design.

Microglia snucRNA-seq data contained a mixture of cell types, but barcode annotations were not provided, so we replicated the Seurat-based clustering analysis as closely as possible based on the description by Thrupp et al. (2020). We identified two clusters that corresponded to microglia based on marker genes described by Thrupp et al. (2020) including *P2RY12*. The single-cell data from this study contain only microglia as they were FACS-purified before sequencing, obviating the need for cluster analysis.

Statistical analyses

Primary analyses were conducted in Rstudio version 2021.09.0 (RStudio Team 2020) and R version 4.1.1 (R Core Team 2024) using the Seurat v4.1.1 (Butler et al. 2018) and tidyverse v1.3.2 packages (Wickham et al. 2019a,b). We performed library size normalization with the Seurat NormalizeData function and "relative counts" mode, which converts raw UMI counts to CPM within each cell. At the cell type level, we depict gene abundance as the average $\log_{10}$ CPM. To test for technical differences in gene abundance between assay- and/or preprocessing for a specific cell type, the relevant Seurat objects were merged as needed and the FindAllMarkers function was applied, with either the "preprocessing" or "assay" supplied as the identity class and Wilcoxon rank-sum as the statistical test. We note that this analysis differs from the typical biological usage of the term "marker gene" as a cell type–defining gene, and we were only comparing two groups here, despite the name of the software function. An adjusted *P*-value threshold of 0.05 was used to select significant genes in all comparisons, based on the Bonferroni correction built into Seurat.

We also performed a typical marker gene analysis, which identifies genes that are significantly more abundant in one cell type compared with the rest, again using the FindAllMarkers function independently on the cell and nuclear Seurat objects. We used the original cell type labels as identity classes and used a fold change threshold of 1.5. The number of A(15) motifs in the mouse genome was calculated with the Biostrings v2.62.0 (https://bioconductor.org/packages/release/bioc/html/Biostrings.html) and GenomicRanges v1.46.1 (Lawrence et al. 2013) R packages. Down-sampling was performed with the DownSampleMatrix function from DropletUtils v1.14.2 (Lun et al. 2019). L5 IT cells were randomly down-sampled to 61% of their *Intron&Exon* counts, and nuclei were down-sampled to 34%.

Estimation of pre-mRNA content and scaling approach

We inferred pre-mRNA content from the difference in total transcripts detected with or without the use of intronic reads. STARsolo refers to these approaches as "Gene" (*Exon*) and "GeneFull" (*Intron&Exon*). We filtered the "raw" matrices to include only the cell barcodes defined by Yao et al. (2021). We reimplemented the Gupta et al. length-correction approach as described in their paper and code. *Intron* counts were taken as the difference between *Intron&Exon* and *Exon*, floored at zero. *Intron* counts were then divided by a gene-specific scalar, the expected number of A(15) motifs per gene; given gene length; and added back to *Exon*. We calculated a scalar of .27 motifs per kilobase of gene in mouse.

GOseq analysis

We performed three analyses with the GOseq R package (Young et al. 2010). First, we assessed genes significantly more abundant in L5 IT nuclei compared with cells (fold change >1.5). Second, we tested marker genes for each cell type in the cortex nucleus data subset (fold change >1.5). Finally, we examined the purported enrichment of DAM genes in human microglial cells by applying GOseq to the cell-enriched genes (repeated with fold change >1.5 and >2). As input for the background gene set, we used genes that were detected with at least one CPM in both cells and/or nuclei in their respective quantification scheme. The same approach was used for each preprocessing mode. We supplied gene length on $\log_{10}$ scale as the bias term, as opposed to transcript length used by the developers.

Software availability

All analysis code is publicly available at GitHub (https://github.com/johnchamberlin/internalpriming) and as Supplemental Code.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

J.T.C. was supported by a National Institutes of Health National Library of Medicine T15 training grant in biomedical informatics, project number 5T15LM007124.

Author contributions: J.T.C. conceived of the study with supervision from Y.L. A.R.Q. supervised the primary analysis and writing. G.T.M. provided additional supervision and assistance with writing.

References

Alvarez M, Rahmani E, Jew B, Garske KM, Miao Z, Benhammou JN, Ye CJ, Pisegna JR, Pietiläinen KH, Halperin E, et al. 2020. Enhancing droplet-based single-nucleus RNA-seq resolution using the semi-supervised machine learning classifier DIEM. *Sci Rep* **10:** 11019. doi:10.1038/s41598-020-67513-5

Bakken TE, Hodge RD, Miller JA, Yao Z, Nguyen TN, Aevermann B, Barkan E, Bertagnolli D, Casper T, Dee N, et al. 2018. Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PLoS One* **13:** e0209648. doi:10.1371/journal.pone.0209648

Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36:** 411–420. doi:10.1038/nbt.4096

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Cao J, O'Day DR, Pliner HA, Kingsley PD, Deng M, Daza RM, Zager MA, Aldinger KA, Blecher-Gonen R, Zhang F, et al. 2020. A human cell atlas of fetal gene expression. *Science* **370:** eaba7721. doi:10.1126/science.aba7721

Del-Aguila JL, Li Z, Dube U, Mihindukulasuriya KA, Budde JP, Fernandez MV, Ibanez L, Bradley J, Wang F, Bergmann K, et al. 2019. A single-nuclei RNA sequencing study of Mendelian and sporadic AD in the human brain. *Alzheimers Res Ther* **11:** 71. doi:10.1186/s13195-019-0524-x

Drokhlyansky E, Smillie CS, Van Wittenberghe N, Ericsson M, Griffin GK, Eraslan G, Dionne D, Cuoco MS, Goder-Reiser MN, Sharova T, et al. 2020. The human and mouse enteric nervous system at single-cell resolution. *Cell* **182:** 1606–1622.e23. doi:10.1016/j.cell.2020.08.003

Eraslan G, Drokhlyansky E, Anand S, Fiskin E, Subramanian A, Slyper M, Wang J, Van Wittenberghe N, Rouhana JM, Waldman J, et al. 2022. Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science* **376:** eabl4290. doi:10.1126/science.abl4290

Fazal FM, Han S, Parker KR, Kaewsapsak P, Xu J, Boettiger AN, Chang HY, Ting AY. 2019. Atlas of subcellular RNA localization revealed by APEX-seq. *Cell* **178:** 473–490.e26. doi:10.1016/j.cell.2019.05.027

Gorin G, Pachter L. 2023a. Distinguishing biophysical stochasticity from technical noise in single-cell RNA sequencing using Monod. bioRxiv doi:10.1101/2022.06.11.495771

Gorin G, Pachter L. 2023b. Length biases in single-cell RNA sequencing of pre-mRNA. *Biophys Rep* **3:** 100097. doi:10.1016/j.bpr.2022.100097

Grubman A, Chew G, Ouyang JF, Sun G, Choo XY, McLean C, Simmons RK, Buckberry S, Vargas-Landin DB, Poppe D, et al. 2019. A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. *Nat Neurosci* **22:** 2087–2097. doi:10.1038/s41593-019-0539-4

Gupta A, Shamsi F, Altemose N, Dorlhiac GF, Cypess AM, White AP, Yosef N, Patti ME, Tseng Y-H, Streets A. 2022. Characterization of transcript enrichment and detection bias in single-nucleus RNA-seq for mapping of distinct human adipocyte lineages. *Genome Res* **32:** 242–257. doi:10.1101/gr.275509.121

Habib N, Avraham-Davidi I, Basu A, Burks T, Shekhar K, Hofree M, Choudhury SR, Aguet F, Gelfand E, Ardlie K, et al. 2017. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat Methods* **14:** 955–958. doi:10.1038/nmeth.4407

He D, Soneson C, Patro R. 2023. Understanding and evaluating ambiguity in single-cell and single-nucleus RNA-sequencing. bioRxiv doi:10.1101/2023.01.04.522742

Heumos L, Schaar AC, Lance C, Litinetskaya A, Drost F, Zappia L, Lücken MD, Strobl DC, Henao J, Curion F, et al. 2023. Best practices for single-cell analysis across modalities. *Nat Rev Genet* **24:** 550–572. doi:10.1038/s41576-023-00586-w

Hu P, Fabyanic E, Kwon DY, Tang S, Zhou Z, Wu H. 2017. Dissecting cell-type composition and activity-dependent transcriptional state in mammalian brains by massively parallel single-nucleus RNA-seq. *Mol Cell* **68:** 1006–1015.e7. doi:10.1016/j.molcel.2017.11.017

Hu P, Liu J, Zhao J, Wilkins BJ, Lupino K, Wu H, Pei L. 2018. Single-nucleus transcriptomic survey of cell diversity and functional maturation in postnatal mammalian hearts. *Genes Dev* **32:** 1344–1357. doi:10.1101/gad.316802.118

Janssen P, Kliesmete Z, Vieth B, Adiconis X, Simmons S, Marshall J, McCabe C, Heyn H, Levin JZ, Enard W, et al. 2023. The effect of background noise and its removal on the analysis of single-cell expression data. *Genome Biol* **24:** 140. doi:10.1186/s13059-023-02978-x

Kaminow B, Yunusov D, Dobin A. 2021. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. bioRxiv doi:10.1101/2021.05.05.442755

Keren-Shaul H, Spinrad A, Weiner A, Matcovitch-Natan O, Dvir-Szternfeld R, Ulland TK, David E, Baruch K, Lara-Astaiso D, Toth B, et al. 2017. A unique microglia type associated with restricting development of Alzheimer's disease. *Cell* **169:** 1276–1290.e17. doi:10.1016/j.cell.2017.05.018

Kuo A, Hansen KD, Hicks SC. 2023. Quantification and statistical modeling of droplet-based single-nucleus RNA-sequencing data. *Biostatistics* kxad010. doi:10.1093/biostatistics/kxad010

Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, Vallejos CA, Campbell KR, Beerenwinkel N, Mahfouz A, et al. 2020. Eleven grand challenges in single-cell data science. *Genome Biol* **21:** 31. doi:10.1186/s13059-020-1926-6

Lake BB, Codeluppi S, Yung YC, Gao D, Chun J, Kharchenko PV, Linnarsson S, Zhang K. 2017. A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA. *Sci Rep* **7:** 6031. doi:10.1038/s41598-017-04426-w

Lake BB, Chen S, Sos BC, Fan J, Kaeser GE, Yung YC, Duong TE, Gao D, Chun J, Kharchenko PV, et al. 2018. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol* **36:** 70–80. doi:10.1038/nbt.4038

La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, Lidschreiber K, Kastriti ME, Lönnerberg P, Furlan A, et al. 2018. RNA velocity of single cells. *Nature* **560:** 494–498. doi:10.1038/s41586-018-0414-6

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9:** e1003118. doi:10.1371/journal.pcbi.1003118

Lun ATL, Riesenfeld S, Andrews T, Dao TP, Gomes T; participants in the 1st Human Cell Atlas Jamboree; Marioni JC. 2019. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol* **20:** 63. doi:10.1186/s13059-019-1662-y

Martins-Ferreira R, Calafell-Segura J, Leal B, Rodríguez-Ubreva J, Mereu E, Pinho e Costa P, Ballestar E. 2023. The human microglia atlas (HuMicA) unravels changes in homeostatic and disease-associated microglia subsets across neurodegenerative conditions. bioRxiv doi:10.1101/2023.08.01.550767

Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, Menon M, He L, Abdurrob F, Jiang X, et al. 2019. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570:** 332–337. doi:10.1038/s41586-019-1195-2

Melsted P, Booeshaghi AS, Liu L, Gao F, Lu L, Min KHJ, da Veiga Beltrame E, Hjörleifsson KE, Gehring J, Pachter L. 2021. Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nat Biotechnol* **39:** 813–818. doi:10.1038/s41587-021-00870-2

Mereu E, Lafzi A, Moutinho C, Ziegenhain C, McCarthy DJ, Álvarez-Varela A, Batlle E, Sagar, Grün D, Lau JK, et al. 2020. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat Biotechnol* **38:** 747–755. doi:10.1038/s41587-020-0469-4

Park Y, He L, Davila-Velderrain J, Hou L, Mohammadi S, Mathys H, Peng Z, Bennett D, Tsai L-H, Kellis M. 2021. Single-cell deconvolution of 3,000 post-mortem brain samples for eQTL and GWAS dissection in mental disorders. bioRxiv doi:10.1101/2021.01.21.426000

Phipson B, Zappia L, Oshlack A. 2017. Gene length and detection bias in single cell RNA sequencing protocols. *F1000Res* **6:** 595. doi:10.12688/f1000research.11290.1

R Core Team. 2024. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

RStudio Team. 2020. *RStudio: integrated development for R*. RStudio, PBC, Boston. http://www.rstudio.com/.

Sala Frigerio C, Wolfs L, Fattorelli N, Thrupp N, Voytyuk I, Schmidt I, Mancuso R, Chen W-T, Woodbury ME, Srivastava G, et al. 2019. The major risk factors for Alzheimer's disease: age, sex, and genes modulate the microglia response to Aβ plaques. *Cell Rep* **27:** 1293–1306.e6. doi:10.1016/j.celrep.2019.03.099

Selewa A, Dohn R, Eckart H, Lozano S, Xie B, Gauchat E, Elorbany R, Rhodes K, Burnett J, Gilad Y, et al. 2020. Systematic comparison of high-throughput single-cell and single-nucleus transcriptomes during cardiomyocyte differentiation. *Sci Rep* **10:** 1535. doi:10.1038/s41598-020-58327-6

Soneson C, Srivastava A, Patro R, Stadler MB. 2021. Preprocessing choices affect RNA velocity results for droplet scRNA-seq data. *PLoS Comput Biol* **17:** e1008585. doi:10.1371/journal.pcbi.1008585

Srivastava A, Malik L, Smith T, Sudbery I, Patro R. 2019. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. *Genome Biol* **20:** 65. doi:10.1186/s13059-019-1670-y

Svensson V, da Veiga Beltrame E, Pachter L. 2020. A curated database reveals trends in single-cell transcriptomics. *Database* **2020:** baaa073. doi:10.1093/database/baaa073

Svoboda M, Frost HR, Bosco G. 2022. Internal oligo(dT) priming introduces systematic bias in bulk and single-cell RNA sequencing count data. *NAR Genom Bioinform* **4:** lqac035. doi:10.1093/nargab/lqac035

Thrupp N, Sala Frigerio C, Wolfs L, Skene NG, Fattorelli N, Poovathingal S, Fourne Y, Matthews PM, Theys T, Mancuso R, et al. 2020. Single-nucleus RNA-seq is not suitable for detection of microglial activation genes in humans. *Cell Rep* **32:** 108189. doi:10.1016/j.celrep.2020.108189

Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, et al. 2019a. Welcome to the tidyverse. *J Open Source Softw* **4:** 1686. doi:10.21105/joss.01686

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, et al. 2019b. Welcome to the tidyverse. https://tidyverse.tidyverse.org/articles/paper.html [accessed July 1, 2022].

Downloaded from genome.cshlp.org on June 16, 2026 . Published by Cold Spring Harbor Laboratory Press

Wu H, Kirita Y, Donnelly EL, Humphreys BD. 2019. Advantages of single-nucleus over single-cell RNA sequencing of adult kidney: rare cell types and novel cell states revealed in fibrosis. *J Am Soc Nephrol* **30:** 23–32. doi:10.1681/ASN.2018090912

Xie Y, Chen H, Chellamuthu VR, Lajam ABM, Albani S, Low AHL, Petretto E, Behmoaras J. 2023. Comparative analysis of single-cell RNA sequencing methods with and without sample multiplexing. bioRxiv doi:10.1101/2023.06.28.546827

Yao Z, Liu H, Xie F, Fischer S, Adkins RS, Aldridge AI, Ament SA, Bartlett A, Behrens MM, Van den Berge K, et al. 2021. A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature* **598:** 103–110. doi:10.1038/s41586-021-03500-8

Young MD, Wakefield MJ, Smyth GK, Oshlack A. 2010. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* **11:** R14. doi:10.1186/gb-2010-11-2-r14

Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, et al. 2017. Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8:** 14049. doi:10.1038/ncomms14049

Received July 7, 2023; accepted in revised form February 1, 2024.