Variable RNA sampling biases mediate concordance of single-cell and nucleus sequencing across cell types

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

1. Department of Biomedical Informatics, University of Utah, Salt Lake City, USA
2. Seoul National University, College of Veterinary Medicine, Seoul, South Korea
3. Department of Human Genetics, Utah Center for Genetic Discovery, University of Utah, Salt Lake City, USA

* to whom correspondence should be addressed

Abstract

The capacity for nuclear RNA measurements to recapitulate results from whole cells is essential to the utility of single-nucleus RNA-seq. Early studies argued that nuclear samples could yield comparable results to single-cell RNA-seq if intronic reads from pre-mRNAs were included in the analysis of both assays. While pre-mRNA sampling has since been acknowledged to be subject to sampling bias related to gene length, the impact of this phenomenon across cell types has been largely ignored. Here, we describe the contrasting effects of mRNA and pre-mRNA sampling on the concordance of gene expression estimates between cells and nuclei. We also address the generalizability of a recently published normalization method intended to maximize assay similarity by removing gene length bias from pre-mRNA sampling. Comparing nuclei to cells among cell types of the cortex, we show that pre-mRNA (intron) abundances are much more similar than mRNA (exon) abundances. When comparing overall gene expression, the magnitude of gene length bias reflects the relative enrichment of pre-mRNAs in nuclei, which varies considerably among cell types of the cortex. This variability leads to unreliable performance of the normalization method, which emphasizes mRNA measurements by downweighting pre-mRNA measurements according to gene length. As a potential alternative, we demonstrate adaptation of an existing method for removing systematic bias from gene set enrichment analysis results. Broadly, our analysis provides a mechanistic explanation for variation in assay similarity across cell types and argues for the application of *post hoc* normalization approaches as an avenue to improved biological interpretation.

Background

Single-cell RNA sequencing (scRNA-seq) quantifies gene expression in individual cell types, as opposed to the tissue-wide measurements obtained from conventional "bulk" RNA-seq. Applications range from broad “cell atlas" projects[1,2] to studies of specific diseases such as Alzheimer’s[3–5]. When dissociation of intact single cells is difficult, as with frozen nervous tissue

samples, researchers often instead use single-*nucleus* sequencing (snucRNA-seq), where nuclei are isolated from cell lysate.

Although nuclei contain only a subset of cellular RNA, they are enriched for pre-mRNAs which results in a high fraction of intronic read alignments. Initial publications argued that elevated data sparsity could be mitigated by incorporating these intronic reads, assumed to reflect unspliced pre-mRNAs, during data preprocessing[6–10]. Subsequent studies extended this strategy to whole cells when analyzed alongside nuclei[11,12]. Most recently, the widely-used CellRanger software was updated to incorporate intronic reads by default for both cell and nuclear experiments[13]. While inclusion of intronic reads has been reported to improve similarity of scRNA-seq and snucRNA-seq data[10], the extent and generality of this improvement across cell types has not been systematically addressed in the literature.

In principle, assays such as the 10x Genomics system utilize barcoded poly(dT) primers to assign at most one unique molecular identifier (UMI) to each polyadenylated mRNA, irrespective of length[14]. Sequence reads are then generated near the 3' end of the cDNA, i.e., shortly upstream of the priming site. This model led some authors to ascribe the preponderance of intronic reads in nuclei (**Figure 1A**) to the capture of polyadenylated but incompletely spliced RNAs[6]. It is now appreciated that primer hybridization also occurs readily at internal adenosine homopolymers[12,15–17], which primarily reside in introns, i.e., pre-mRNAs[18]. So-called *internal priming*[19] introduces a gene length-associated sampling bias when intronic reads are included in the analysis[15,20,21]: whereas an mRNA can have only one poly(A) tail, the number of internal priming sites in a pre-mRNA is approximately a function of its transcribed length (**Figure 1B**).

Given that the large majority of internal priming sites occur in introns (i.e., not in mRNAs)[20,22], the overall magnitude of gene length bias becomes a function of the proportion of pre-mRNA captured in the sample. As such, nuclei exhibit stronger gene length bias than cells because they are enriched for pre-mRNA[6]. In other words, improved similarity of cell and nuclear data through incorporation of pre-mRNA signal is achieved despite implicit introduction of a differential gene length bias. Past studies have occasionally commented on this bias but have made no attempt to address it[15].

Recently, Gupta et al 2022 introduced a normalization scheme designed to account for discrepant gene length bias in cell-vs-nucleus comparisons[21]. Specifically, they propose separating total transcript counts (i.e., observed UMIs per gene per cell) into *exon*- and *intron*-derived components and scaling solely the *intron* UMI counts by gene length multiplied by the transcriptome-wide rate of internal priming site occurrence (**Figure 1C**). In practice, this equates to dividing the *intron* UMI counts from a gene by the expected number of internal priming sites given its length, and then adding the scaled result to the unaltered *exon* UMI counts. Gupta et al show that their method minimizes gene length bias and leads to both improved correlation of gene expression estimates and reduced number of seemingly differentially-expressed genes between human preadipocyte cells and nuclei. However, the authors do not evaluate performance in other cell types, which is necessary for understanding the generalizability of the method, since pre-mRNA content (i.e., intronic read fraction) can vary

widely[20]. For example, two early snucRNA-seq studies reported just 16% in mouse heart versus 50% in mouse brain[23,24], which suggests that the incentive to account for bias is highly tissue-specific.

Here, we address this gap by re-analyzing a comprehensive dataset from the mouse motor cortex[25]. The cortex is a useful case study because nervous tissues are commonly subjected to snucRNA-seq and tend to exhibit high intronic read fraction, which suggests that it is a priority candidate for the application of gene length normalization methods.

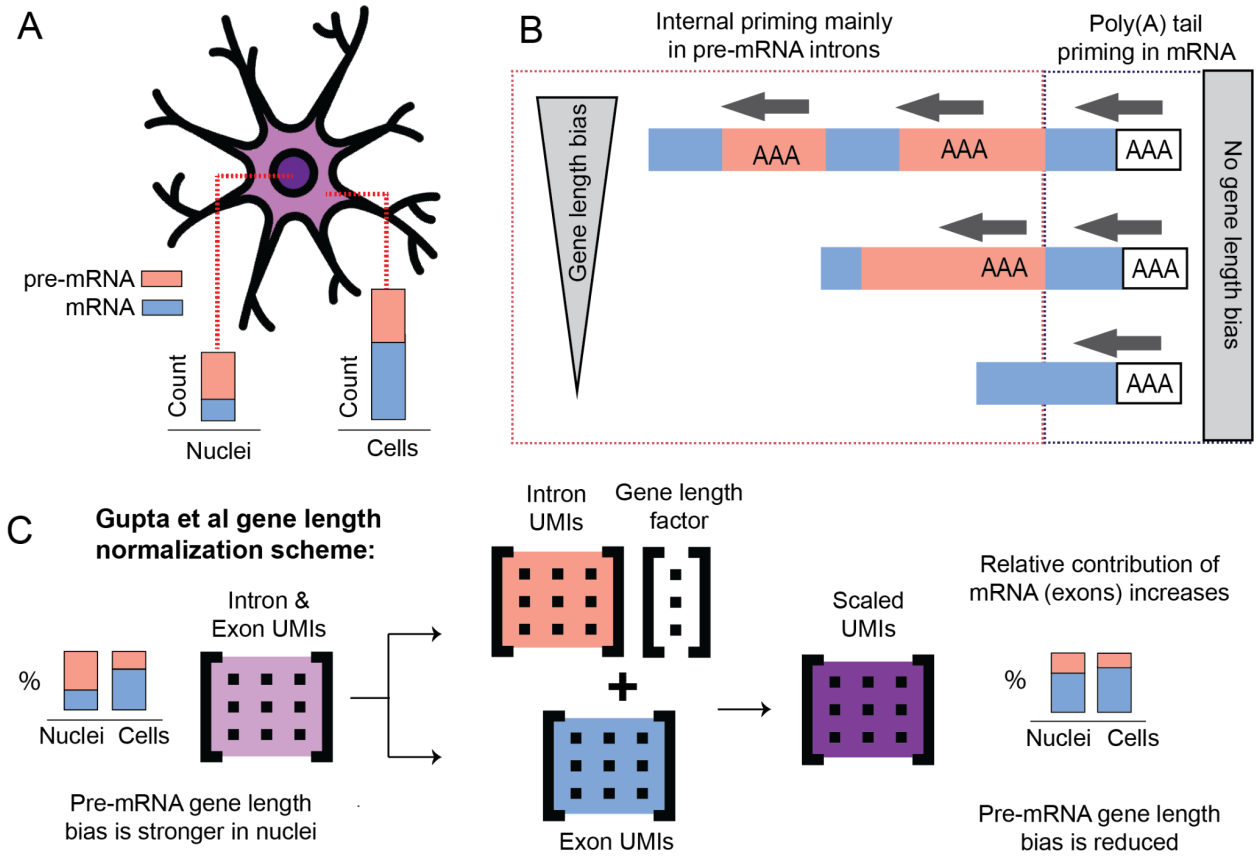

Figure 1: (A) Nuclei contain less total RNA but are enriched for pre-mRNA compared to cells (B) Theoretical ideals of RNA sampling: pre-mRNA tends to generate intronic reads via internal priming in a gene-length associated matter. mRNA generates exonic reads from priming at the poly(A) tail, irrespective of length. (C) Outline of Gupta et al normalization approach. Total UMIs in both cells and nuclei are separated into *intron* and *exon*-derived components, and *intron* counts are divided by a factor of gene length. *Exon* and adjusted *intron* counts are then summed. Because *intron* counts are reduced, the relative contribution of *exon* counts increases.

Results and Discussion

With respect to this analysis, we note that typical studies do not generate matched single-cell and nucleus RNA-seq datasets, are narrow in scope, or are otherwise unsuitable for a systematic evaluation. As such, we first confirmed the expected relationship between intronic read fraction and gene length bias (**Supplementary Figure 1**) using metadata from a multi-organ snucRNA-seq atlas of human fetuses[26]. We then centered our primary analysis on a well-annotated dataset from the mouse motor cortex which includes matched cell- and nuclear-data generated with the popular 10x Genomics V3 assay[25]. We used STARsolo[27] to compute gene abundance estimates with and without intronic regions. Quantification scheme nomenclature are not standardized in the literature; herein we use the terms *exon* (UMI counts from exonic alignments), *intron* (UMI counts not from exonic alignments), and *intron&exon* (UMI counts from genic alignments). The *de facto* standard *intron&exon* workflow does not allow the user to distinguish between the two components - *intron* counts must be inferred by separately generating and then subtracting *exon* counts. We refer to the result of the Gupta et al normalization scheme as *scaled* counts to avoid confusion with the ubiquitous practice of log normalization of gene abundances. *Scaled* counts were generated by dividing *intron* counts for each gene by a factor based on the rate of internal priming sites, defined as 15 or more consecutive adenosines (0.27 sites per kilobase of gene in mouse) and adding them back to *exon* counts.

We began by analyzing L5 IT neurons, the most abundant (42% of cells, 29% of nuclei) annotated cell type in the cortex dataset. As expected, intron content was higher (67% vs 41% of UMIs), *intron&exon* gene abundances were more positively correlated with gene length in nuclei than in cells (**Figure 2A**), and genes enriched in nuclei were significantly longer than those in cells **(Figure 2B)**. *Intron* gene length bias was similarly strong between the two assays, indicating that pre-mRNA sampling patterns are comparable. While the overall level of intronic reads was higher in cortex than in preadipocytes, we found the inter-assay ratio to be comparatively low. Specifically, introns contributed 67% and 41% of counts in L5 IT nuclei and cells, respectively, while Gupta et al reported intronic read fractions of 40% in nuclei vs just 9% in cells. The relative enrichment values (1.5-fold in neurons, 4.4 in preadipocytes) suggest that gene length bias has a weaker effect on between-assay differences in L5 IT neurons than in preadipocytes despite stronger bias within either assay individually.

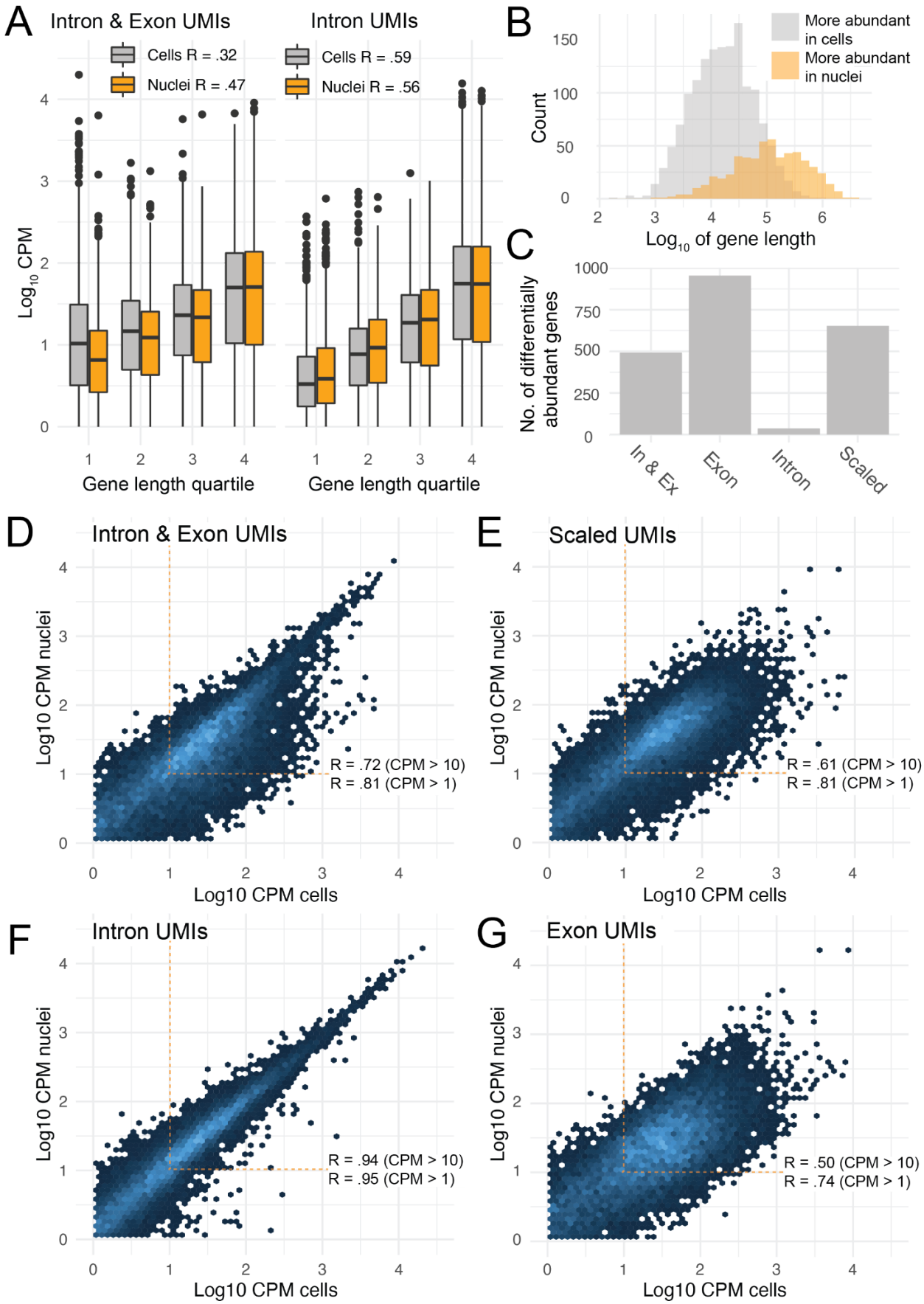
A
Intron & Exon UMIs
Intron UMIs
Cells R = .32
Nuclei R = .47
Cells R = .59
Nuclei R = .56
Log10 CPM
Gene length quartile
Gene length quartile
B
More abundant in cells
More abundant in nuclei
Count
Log10 of gene length
C
No. of differentially abundant genes
In & Ex
Exon
Intron
Scaled
D
Intron & Exon UMIs
Log10 CPM nuclei
Log10 CPM cells
R = .72 (CPM > 10)
R = .81 (CPM > 1)
E
Scaled UMIs
Log10 CPM nuclei
Log10 CPM cells
R = .61 (CPM > 10)
R = .81 (CPM > 1)
F
Intron UMIs
Log10 CPM nuclei
Log10 CPM cells
R = .94 (CPM > 10)
R = .95 (CPM > 1)
G
Exon UMIs
Log10 CPM nuclei
Log10 CPM cells
R = .50 (CPM > 10)
R = .74 (CPM > 1)

Figure 2. (A) *Intron&exon* gene length bias is stronger in nuclei (orange) than in cells (gray), but *intron* bias is similar. Units are log10 of average counts per million (CPM). (B) Log_{10} length of genes which differ in *intron&exon* abundance between cells and nuclei ($log_2FC > 1$, FDR < .05). Genes enriched in nuclei are significantly longer (C) Total number of differentially abundant genes per counting scheme ($log_2FC > 1$, FDR < .05). The scaling approach worsens the apparent similarity with respect to *intron&exon*. (D) Correlation of *intron&exon* abundances for genes above 1 CPM. (E) *Scaled* abundances are no better correlated than the baseline result, and are less well-correlated among highly expressed genes. (F) Correlation of *intron* abundances is very high. (G) Correlation of *exon* abundances is less strong, particularly for abundant genes.

Gupta et al reported improvements to assay similarity after applying their normalization method in terms of both reduced number of differentially abundant genes and increased correlation of gene expression: the number of differentially abundant genes decreased from 1061 to 631 and Pearson correlation of gene expression improved from 0.5 to 0.6. In neurons, *intron&exon* gene expression estimates were comparatively well-correlated even without adjustment (**Figure 2D**, R = .81, mean abundance of genes greater than 1 CPM). However, *scaled* abundances offered no further improvement: while R remained at .81 (**Figure 2E)**, differential expression testing resulted in 33% *more* apparent differences (**Figure 2C**, 654 vs 492 genes with log_2 fold change > 1), in contrast to the 40% reduction reported by Gupta et al. The increase in the number of differential genes likely reflects worsened correlation of *scaled* abundances among more-abundant genes that are statistically powered to appear significantly different.

To understand the basis of this discrepancy, we compared *intron* and *exon* abundances separately (**Figure 2F,G**). *Intron* counts resulted in a very high correlation of R=.95 and just 37 differentially expressed genes, whereas *exon* counts showed the weakest correlation at R=.72 and greatest number of differential genes (955, **Figure 2C).** Downsampling confirmed that the high correlation of *intron* counts was not a function of sequencing depth, indicating that differences in mRNA abundances primarily reflect biased cytosolic localization and/or post-transcriptional regulation of mRNAs. These patterns are explanatory if we consider the reduction in data that occurs upon scaling: the *intron* contribution (% of total of *intron&exon* vs *scaled*) decreased from 40% to 25% in cells and from 67% to 50% in nuclei. As such, amelioration of gene length bias in L5 IT neurons through scaling of *intron* UMI counts is counteracted by increased emphasis on dissimilarities in mRNA abundance.

We next extended our analysis to the remaining cortex cell types. While we expected the pre-mRNA content to vary, we were surprised to find that a higher intronic UMI fraction in nuclei of a given type did not necessarily equate to a higher level in cells (**Figure 3A,** p = .45). This is important because it implies that the magnitude of differential gene length bias, or, the potential impact of the normalization method, is not only variable but also unpredictable from snucRNA-seq data alone. Differential expression tests on *intron&exon* counts for each cell type revealed a positive trend between the number of differentially-abundant genes and the ratio of median intronic UMI content between nuclei and cells (**Figure 3B,** Spearman's rho = .92), supporting the premise that assay differences are driven in part by the discrepancy in gene length bias between cells and nuclei from the same cell type.

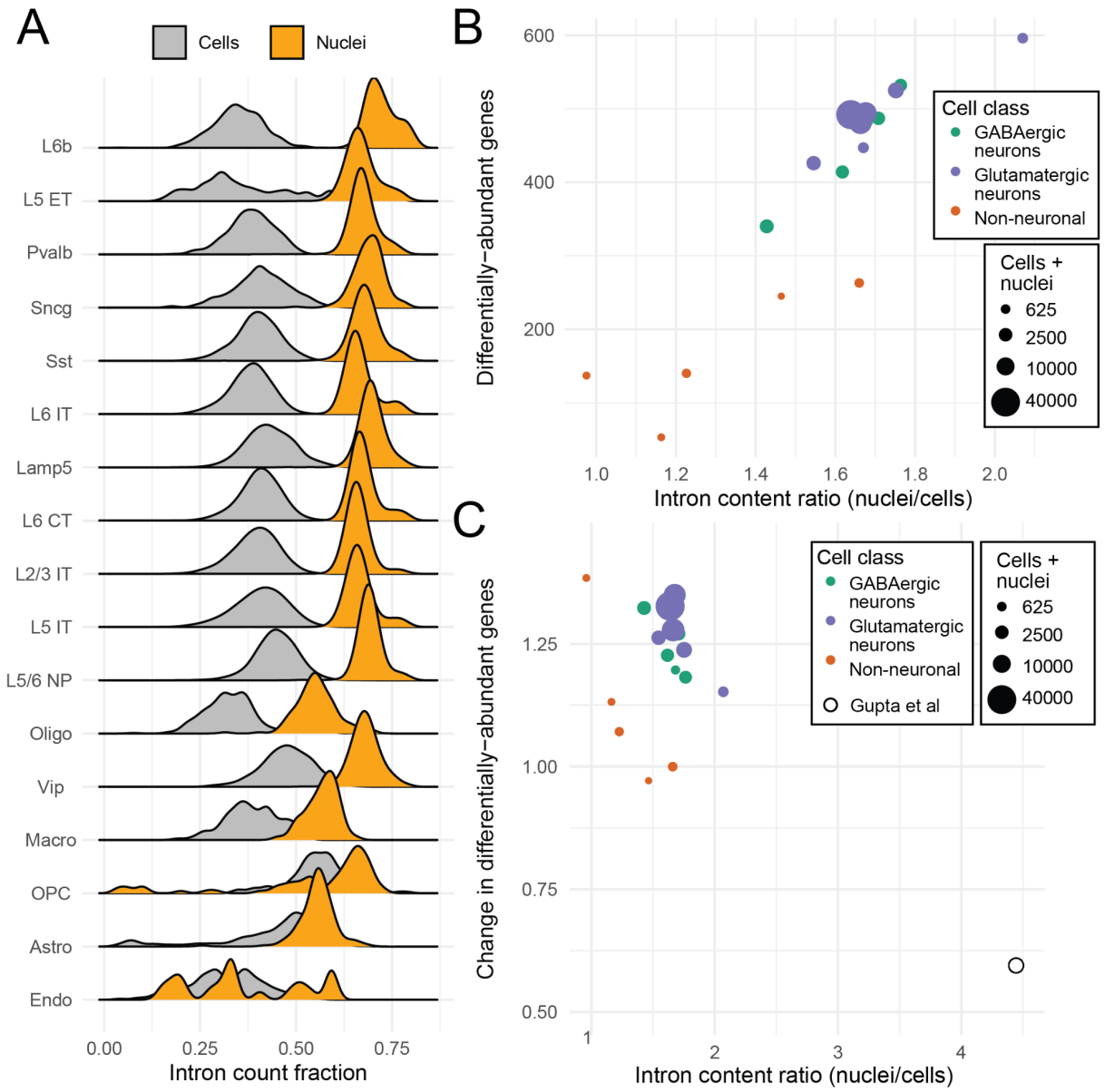

Figure 3. (A) Intron content distribution (total *intron* UMIs divided by total *intron&exon* UMIs per cell or nucleus) for each cell type. Mean intron content in cells does not increase with mean intron content in nuclei (rho = .19, p = .46). Non-neuronal cells are abbreviated: astro, astrocytes; endo, endothelial cells; oligo, oligodendrocytes; OPC, oligodendrocyte precursor cells; macro, macrophages. Remaining labels specify types of neuronal cells. (B) Cell type-specific differential expression of cells vs nuclei. The number of differential genes (log2FC > 1) increases with the ratio of mean intron content in the cell type (rho = .92). Cells are colored by "class label" assigned by primary authors (C) Fold change in number of differentially abundant genes for cells vs nuclei after applying the Gupta et al normalization procedure. Red point shows the result reported by Gupta et al for white preadipocytes. Values greater than 1 indicate adverse performance of the normalization method. Linear modeling (excluding preadipocyte) identifies cell class and intron content ratio as significant predictors of method performance (R^2 = .78, p = .0001).

We repeated the differential expression and correlation tests after applying the normalization method to the full dataset. The number of significant differences increased for 15 of the 17 cell types, indicating adverse performance (**Figure 3C**), though performance tended to improve with the intron content ratio after accounting for cell class as a covariate. A similar pattern was observed in Pearson correlations (**Figure S2**); however, the exact values are sensitive to the specific gene expression or fold change threshold. We emphasize that these results are not incompatible with the values reported by Gupta et al (depicted as an open circle in **Figure 3C**), where pre-mRNA enrichment (i.e., intron content ratio) between preadipocyte nuclei and cells was more than twice the maximum of any cortex type. However, it suggests that the utility of the method is constrained to scenarios where gene length bias is known to be highly discordant between cell and nucleus assays *a priori*. While this limitation may be surmountable with refinements to the normalization method, we also suggest that *post hoc* normalization methods are a potential alternative.

As a simple demonstration of post hoc normalization, we tested the GOseq algorithm, originally developed for removing transcript length bias from gene set enrichment analysis in conventional fragmentation-based RNA-seq[28,29]. We applied the algorithm to the set of genes that were enriched in L5 IT nuclei compared to cells (result from **Figure 2B**), with gene length as the bias term. Without 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 purportedly of the same type (**Supplementary Table 1)**. After correcting for gene length, top terms were consistent with patterns observed from localization assays, such as “RNA splicing” (**Supplementary Table 2**)[30]: Specifically, Fazal et reported that “mRNAs enriched in nuclear locations tend to code for proteins enriched in nuclear speckles and nucleoplasm.” An equivalent pattern was reported by Bakken et al 2018 when comparing *intron&exon* results to *exon* results using a technically distinct scRNA-seq assay[7]. This demonstrates how a biologically plausible interpretation can be achieved without manipulation of raw gene expression values and without access to raw sequence data, which may be restricted.

Broadly, these findings demonstrate considerable variation in the equivalence of cell- and nuclear measurements across cell types, with consequential variation in performance of the proposed scaling-based normalization scheme. Because pre-mRNA enrichment is difficult to predict, the capacity for snucRNA-seq to recapitulate whole-cell results is also difficult to predict *a priori*. Whereas the reason for the qualitative difference in pre-mRNA enrichment between cells and nuclei is clear, the basis for variation among cell types is not necessarily known. The patterns we observed in the cortex likely reflect the interaction of biological and technical factors, including cellular morphology and transcriptional activity level; sample preservation and dissociation protocol; and absolute RNA content and levels of ambient RNA contamination. The impact of this variation on accurate cell type assignment remains an open question. This variation is also relevant to analyses such as bulk sample deconvolution[31], because it implies that cell types with a higher pre-mRNA content and/or longer marker genes will appear to be

less abundant in bulk samples due to overestimation of those genes' expression in snucRNA-seq.

More fundamentally, our results beg the question of how well mRNA or protein abundances can be predicted from pre-mRNA abundances, irrespective of gene length bias. In fact, others have already reported that the inclusion of intronic reads in snucRNA-seq worsens correlation with mRNA abundances from bulk RNA-seq[5,32]. Similarly, Thrupp et al 2020 reported that snucRNA-seq was not able to recapitulate a microglial gene signature defined from single-cell mRNA abundances[12]. These findings suggest that maximization of single-cell/nucleus assay similarity is not an appropriate aim for all analyses. Further effort is needed to determine the degree to which these are basic limitations imparted by post-transcriptional regulation vs the effect of pre-mRNA sampling bias which can be modeled and corrected for directly, a la Gupta et al.

Our analysis is subject to a number of limitations which are also pertinent to the design of the Gupta et al method. Foremost, we did not alter the parameterization of the method. In particular, the definition of internal priming site was specified arbitrarily, but it directly determines the reduction in total *intron* UMI counts. Relaxing the motif definition (e.g., A(10)) would result in a substantially greater reduction in counts but weaker reduction in gene length bias, and vice versa. Similarly, the majority of genes were amplified, perhaps unintentionally, because they contain fewer than one expected priming site. The intron-exon dichotomization of UMIs is also imperfect: others have shown that gene length bias is not fully explainable by internal priming sites alone, and is also present in exonic reads to a minor extent[33]. True biological differences, or additional technical artifacts[20] may also contribute. We suggest that the method could be improved by instead dividing UMIs into poly(A) tail- and internally-primed; deriving the scale factor empirically; and flooring the scalar at 1 to avoid variance inflation of short genes. Finally, we assumed that the cell type assignments from the Yao et al cortex dataset were correct, and we did not implement any additional steps such as ambient RNA decontamination, as these were not part of the original study. Beyond computational solutions, these issues would likely be circumvented to a large extent through the adoption of improved single-nucleus dissociation protocols which retain more nuclear-associated mRNAs[34].

Conclusion

A mechanistic understanding of the biological and technical factors which influence the equivalence (or lack thereof) of single cell- and single nucleus transcript measurements is essential to the design and interpretation of experiments. Pre-mRNA and mRNA are subject to distinct sampling biases, but the feature-of-origin of the resulting gene expression data is obfuscated by typical workflows. By reprocessing published data, we have shown here that pre-mRNA abundances (i.e., intron-derived UMIs) are more similar than mRNA abundances (i.e., exon-derived UMIs) between cells and nuclei of the same type, independent of sampling depth. Overall similarity (pre-mRNA+mRNA) is moderated by gene length bias which reflects cell type-specific differences in pre-mRNA enrichment. Attempted removal of the gene length

bias through scaling of pre-mRNA UMI counts is counteracted to a varying extent by a consequential emphasis on the unaltered mRNA counts. Because snucRNA-seq is often used in lieu of scRNA-seq, it is difficult to determine if removal of gene length bias during pre-processing will more closely recapitulate what would have been found in whole cells, so we conclude that *post hoc* normalization is preferable in the absence of contrary evidence. Foundational to this analysis is the computation of per-cell pre-mRNA (intron) content as a bias metric. Published literature typically provides pre-mRNA content either at the sample-wide level or not at all; this precludes the analyst from anticipating adverse impacts of high or variable pre-mRNA abundances on comparisons between cell types and assays. To this end, we encourage the estimation of gene expression both with and without intronic reads included, which is an available option in all commonly-used preprocessing software.

Methods

All data described in this analysis are publicly available. We analyzed only the 10x Genomics v3 data subset from Yao et al[25] generated at the Allen Institute for Brain Studies; a complementary dataset generated at the Broad Institute was excluded because it sampled nuclei only. Raw sequence data and metadata were downloaded from http://data.nemoarchive.org/biccn/lab/zeng/transcriptome/

Alignment and gene expression quantification were performed using STARsolo[27] version 2.7.3a with option "--soloFeatures Gene GeneFull", which corresponds to the *exon* and *intron&exon* schemes. We used the GRCm38 reference genome and Ensembl version 99 gene annotation filtered by biotype per 10x Genomics guidelines and the "cellranger mkgtf" utility. Due to erroneous resequencing of some original cortex libraries, we used seqtk[35] to filter out read pairs with truncated barcodes (read1 shorter than the intended 28 base pairs), as the mixture of barcode read lengths was incompatible with STARsolo. This procedure affected the total depth, but not the relative abundances. Resulting abundance estimates were nearly identical to CellRanger values provided by the authors, which is the stated aim of STARsolo.

Preprocessed data and metadata from the human fetal atlas study[26] were downloaded from NCBI GEO accession GSE156793.

Primary analyses were conducted in Rstudio using Seurat[36] and tidyverse packages[37]. The number of A(15) motifs in the mouse genome was calculated with Biostrings[38] and GenomicRanges[39] R packages.

References

1. Regev, A. *et al.* Science forum: the human cell atlas. *Elife* **6**, e27041 (2017).
2. Quake, S. R. A decade of molecular cell atlases. *Trends Genet.* (2022) doi:10.1016/j.tig.2022.01.004.
3. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).
4. Grubman, A. *et al.* A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. *Nat. Neurosci.* **22**, 2087–2097 (2019).
5. Del-Aguila, J. L. *et al.* A single-nuclei RNA sequencing study of Mendelian and sporadic AD in the human brain. *Alzheimers. Res. Ther.* **11**, 71 (2019).
6. Lake, B. B. *et al.* Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **36**, 70–80 (2018).
7. Bakken, T. E. *et al.* Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PLoS One* **13**, e0209648 (2018).
8. Lake, B. B. *et al.* A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA. *Sci. Rep.* **7**, 6031 (2017).
9. Selewa, A. *et al.* Systematic Comparison of High-throughput Single-Cell and Single-Nucleus Transcriptomes during Cardiomyocyte Differentiation. *Sci. Rep.* **10**, 1535 (2020).
10. Wu, H., Kirita, Y., Donnelly, E. L. & Humphreys, B. D. 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 (2019).
11. Mereu, E. *et al.* Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat. Biotechnol.* **38**, 747–755 (2020).

12. Thrupp, N. *et al.* Single-Nucleus RNA-Seq Is Not Suitable for Detection of Microglial Activation Genes in Humans. *Cell Rep.* **32**, 108189 (2020).
13. Release notes for Cell Ranger 7.0.0 (May 17, 2022): -Software -Single Cell Gene Expression -Official 10x Genomics Support. https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/release-notes.
14. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
15. Eraslan, G. *et al.* Single-nucleus cross-tissue molecular reference maps to decipher disease gene function. *bioRxiv* 2021.07.19.452954 (2021) doi:10.1101/2021.07.19.452954.
16. Truong, D. D. *et al.* Dissociation Protocols used for Sarcoma Tissues Bias the Transcriptome observed in Single-cell and Single-nucleus RNA sequencing. *bioRxiv* 2022.01.21.476982 (2022) doi:10.1101/2022.01.21.476982.
17. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
18. Svoboda, M., Robert Frost, H. & Bosco, G. Internal oligo(dT) priming in bulk and single cell RNA sequencing. *bioRxiv* 2021.09.24.461289 (2021) doi:10.1101/2021.09.24.461289.
19. Nam, Lee, Zhou & Cao. Oligo (dT) primer generates a high frequency of truncated cDNAs through internal poly (A) priming during reverse transcription. *Proc. Estonian Acad. Sci. Biol. Ecol.*
20. Interpreting Intronic and Antisense Reads in 10x Genomics Single Cell Gene Expression Data. 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.
21. Gupta, A. *et al.* Characterization of transcript enrichment and detection bias in

single-nucleus RNA-seq for mapping of distinct human adipocyte lineages. *Genome Res.* (2022) doi:10.1101/gr.275509.121.

22. Phipson, B., Zappia, L. & Oshlack, A. Gene length and detection bias in single cell RNA sequencing protocols. *F1000Res.* **6**, 595 (2017).
23. Hu, P. *et al.* Single-nucleus transcriptomic survey of cell diversity and functional maturation in postnatal mammalian hearts. *Genes Dev.* **32**, 1344–1357 (2018).
24. Hu, P. *et al.* Dissecting Cell-Type Composition and Activity-Dependent Transcriptional State in Mammalian Brains by Massively Parallel Single-Nucleus RNA-Seq. *Mol. Cell* **68**, 1006–1015.e7 (2017).
25. Yao, Z. *et al.* A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature* **598**, 103–110 (2021).
26. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**, (2020).
27. Kaminow, B., Yunusov, D. & Dobin, A. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. *bioRxiv* 2021.05.05.442755 (2021) doi:10.1101/2021.05.05.442755.
28. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* **11**, R14 (2010).
29. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. goseq: Gene Ontology testing for RNA-seq datasets. *R Bioconductor* **8**, 1–25 (2012).
30. Fazal, F. M. *et al.* Atlas of Subcellular RNA Localization Revealed by APEX-Seq. *Cell* **178**, 473–490.e26 (2019).
31. Park, Y. *et al.* Single-cell deconvolution of 3,000 post-mortem brain samples for eQTL and GWAS dissection in mental disorders. *bioRxiv* 2021.01.21.426000 (2021) doi:10.1101/2021.01.21.426000.
32. Svoboda, M., Frost, H. R. & Bosco, G. Internal oligo(dT) priming introduces systematic bias

in bulk and single-cell RNA sequencing count data. *NAR Genom Bioinform* **4**, lqac035 (2022).

33. Kuo, A., Hansen, K. D. & Hicks, S. C. Quantification and statistical modeling of Chromium-based single-nucleus RNA-sequencing data. *bioRxiv* 2022.05.20.492835 (2022) doi:10.1101/2022.05.20.492835.
34. Drokhlyansky, E. *et al.* The Human and Mouse Enteric Nervous System at Single-Cell Resolution. *Cell* **182**, 1606–1622.e23 (2020).
35. Li, H. *seqtk: Toolkit for processing sequences in FASTA/Q formats*. (Github).
36. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
37. Welcome to the Tidyverse. https://tidyverse.tidyverse.org/articles/paper.html.
38. Pages, Aboyoun, Gentleman & DebRoy. Biostrings: String objects representing biological sequences, and matching algorithms. *R package version*.
39. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).

Acknowledgements

Funding

JC was supported by an NLM T15 training grant in biomedical informatics, project number 5T15LM007124-25.

Availability of Data and Materials

All data analyzed in this manuscript were acquired from public repositories. Analysis code is available from https://github.com/johnchamberlin/internalpriming

Author Information

Affiliations

Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA
John Chamberlin

Department of Human Genetics and Utah Center for Genetic Discovery, University of Utah, Salt Lake City, UT, USA
Gabor Marth and Aaron Quinlan

Seoul National University, College of Veterinary Medicine, Seoul, South Korea
Younghee Lee

Corresponding author

Correspondence to Aaron Quinlan, aquinlan@genetics.utah.edu

Contributions

JC conceived and conducted the study and wrote the manuscript. YL supervised the conceptualization of the study. GM and AQ supervised the analysis. All authors reviewed and approved the manuscript.

Ethics Declaration

The authors declare that they have no competing interests.

Supplementary Materials

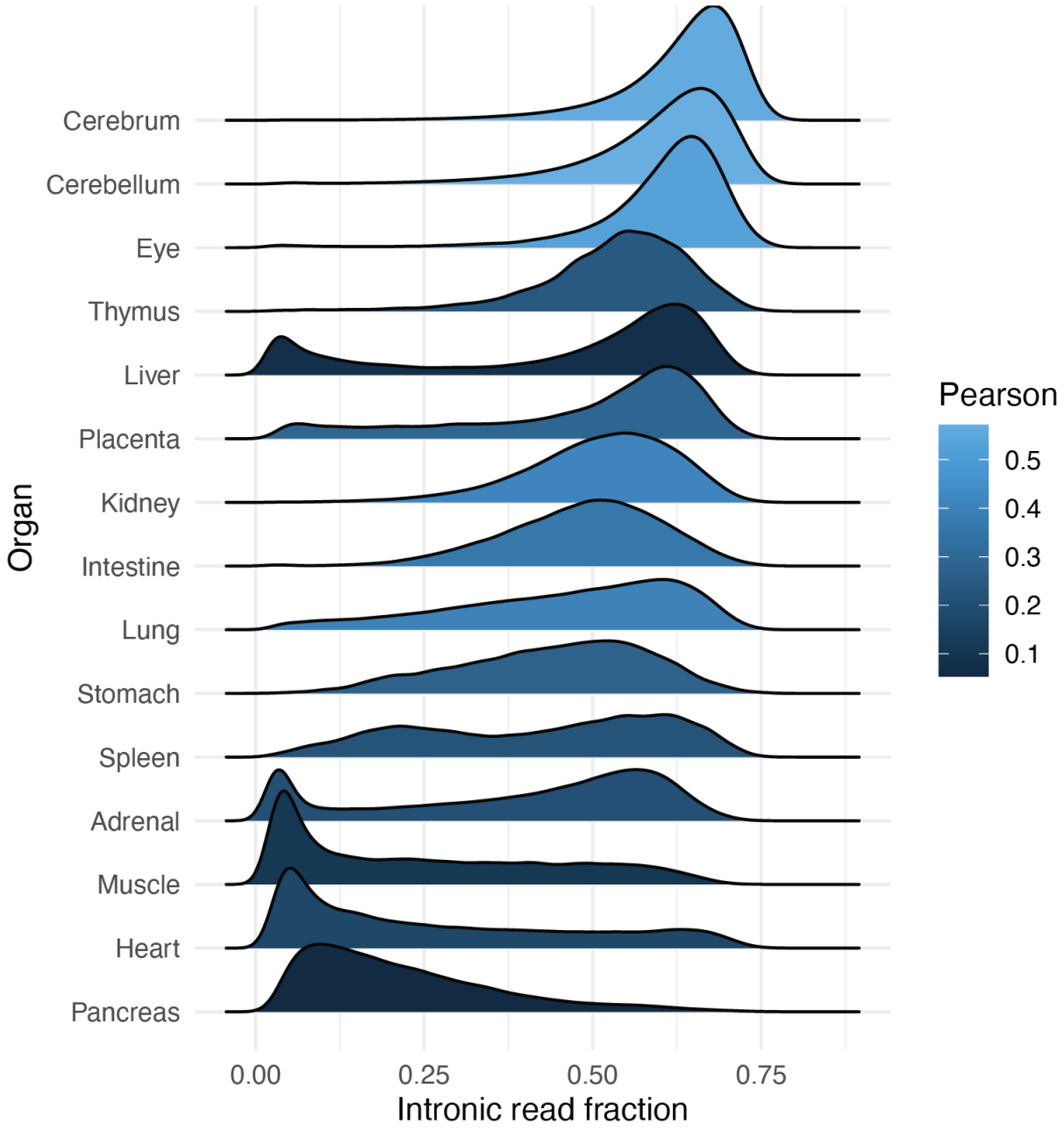

Figure S1: Variation in intronic read fraction reported by the Human Fetal Atlas(Cao et al. 2020). Tissues are ordered by median fraction; nervous tissues exhibit the strongest correlations between average *intron&exon* abundance (log of average CPM) and gene length. Some tissues show stark bimodality, such as Liver and Adrenal gland.

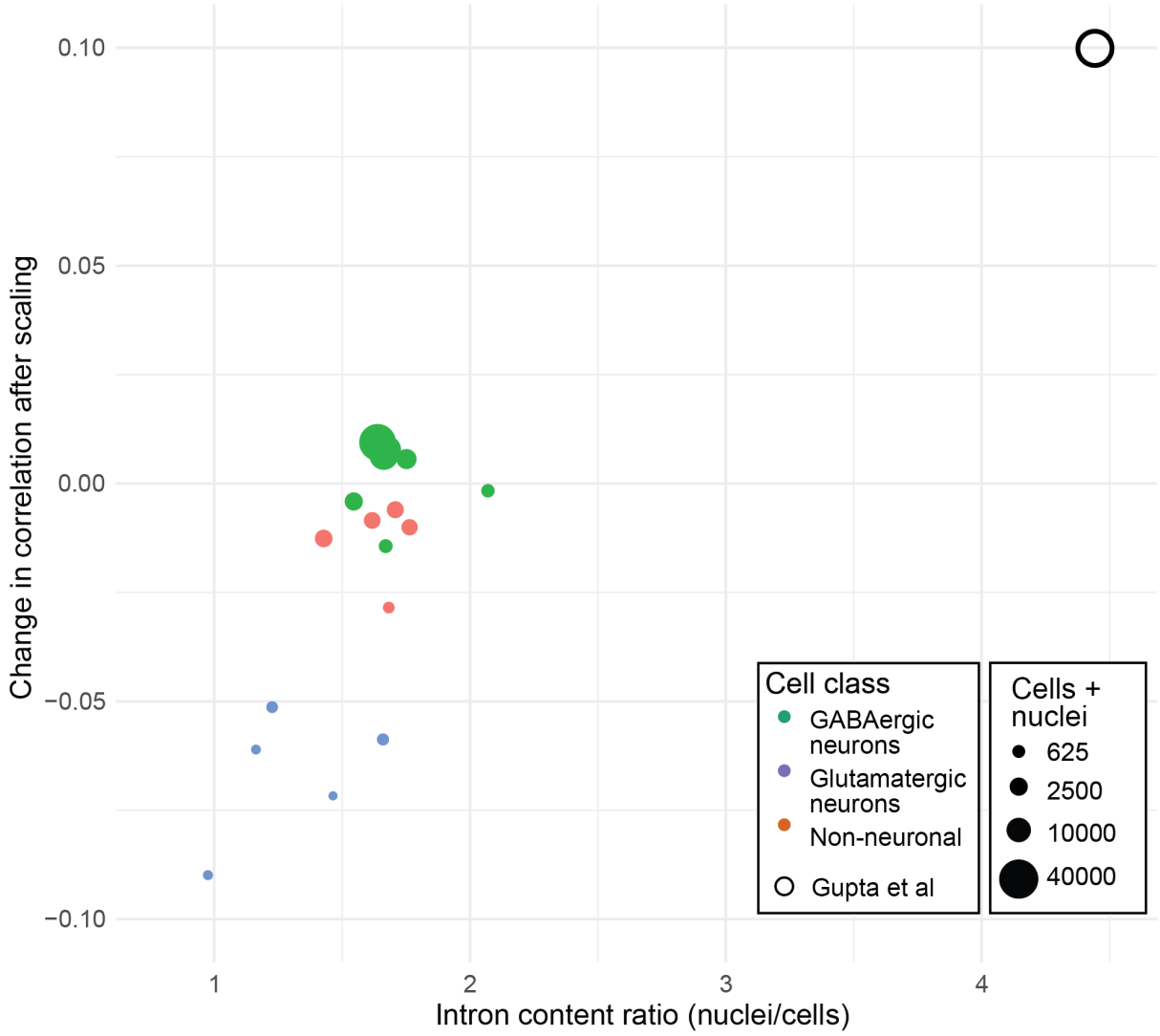

Figure S2: Effect of normalization on correlation between cell and nuclear abundances in mouse motor cortex, for genes greater than 1 CPM. Points are colored by cell category and sized by the total number in both assays. Result reported by Gupta et al is shown as an open circle.

category	over_represented_pvalue	under_represented_pvalue	numDEInCat	numInCat	term	ontology
GO:0050808	3.77039638293979E-12	0.999999999999007	46	440	synapse organization	BP
GO:0045202	8.01153097969425E-12	0.99999999999691	83	1156	synapse	CC
GO:0098978	9.46943054430809E-12	0.999999999997853	37	308	glutamatergic synapse	CC
GO:0097060	1.17882803092381E-11	0.999999999997177	39	341	synaptic membrane	CC
GO:0098794	1.49603995787744E-11	0.999999999995504	52	562	postsynapse	CC
GO:0007399	2.48652253888647E-11	0.999999999988735	117	1944	nervous system development	BP
GO:0048812	4.2095391595161E-11	0.999999999986785	53	596	neuron projection morphogenesis	BP
GO:0034330	7.32671892294605E-11	0.999999999976065	55	642	cell junction organization	BP
GO:0120039	9.40173671755274E-11	0.999999999969759	53	609	plasma membrane bounded cell projection morphogenesis	BP
GO:0048858	1.19737674469223E-10	0.999999999961201	53	613	cell projection morphogenesis	BP

Table S1: Top 10 terms enriched in L5 IT nuclei without correcting for gene length

category	over_represented_pvalue	under_represented_pvalue	numDEInCat	numInCat	term	ontology
GO:0008380	4.12526228435935E-08	0.999999989935242	27	355	RNA splicing	BP
GO:0016071	6.25651838365051E-06	0.999997605075098	34	593	mRNA metabolic process	BP
GO:0006396	9.03004233333777E-06	0.999996271788817	38	807	RNA processing	BP
GO:0006397	1.21023199290547E-05	0.999995878535751	26	419	mRNA processing	BP
GO:0000375	9.20122176629977E-05	0.999972776048697	17	244	RNA splicing, via transesterification reactions	BP
GO:0000377	9.20122176629977E-05	0.999972776048697	17	244	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	BP
GO:0000398	9.20122176629977E-05	0.999972776048697	17	244	mRNA splicing, via spliceosome	BP
GO:0016604	0.000132107144066925	0.999939282078959	36	642	nuclear body	CC
GO:0005681	0.000136172221548681	0.999970688099691	11	182	spliceosomal complex	CC
GO:0002940	0.000221336572067742	1	2	2	tRNA N2-guanine methylation	BP

Table S2: Top 10 terms enriched in L5 IT nuclei after correcting for gene length with GOseq