Biostatistics (2024) **25**, 3, *pp.* 801–817
https://doi.org/10.1093/biostatistics/kxad010
Advance Access publication on May 31, 2023

Quantification and statistical modeling of droplet-based single-nucleus RNA-sequencing data

ALBERT KUO

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N Wolfe St, Baltimore, MD 21205, USA

KASPER D. HANSEN

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N Wolfe St, Baltimore, MD 21205, USA and Department of Genetic Medicine, Johns Hopkins School of Medicine, 733 N Broadway, Baltimore, MD 21205, USA

STEPHANIE C. HICKS*

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N Wolfe St, Baltimore, MD 21205, USA

shicks19@jhu.edu

Summary

In complex tissues containing cells that are difficult to dissociate, single-nucleus RNA-sequencing (snRNA-seq) has become the preferred experimental technology over single-cell RNA-sequencing (scRNA-seq) to measure gene expression. To accurately model these data in downstream analyses, previous work has shown that droplet-based scRNA-seq data are not zero-inflated, but whether droplet-based snRNA-seq data follow the same probability distributions has not been systematically evaluated. Using pseudonegative control data from nuclei in mouse cortex sequenced with the 10x Genomics Chromium system and mouse kidney sequenced with the DropSeq system, we found that droplet-based snRNA-seq data follow a negative binomial distribution, suggesting that parametric statistical models applied to scRNA-seq are transferable to snRNA-seq. Furthermore, we found that the quantification choices in adapting quantification mapping strategies from scRNA-seq to snRNA-seq can play a significant role in downstream analyses and biological interpretation. In particular, reference transcriptomes that do not include intronic regions result in significantly smaller library sizes and incongruous cell type classifications. We also confirmed the presence of a gene length bias in snRNA-seq data, which we show is present in both exonic and intronic reads, and investigate potential causes for the bias.

Keywords: Binomial distribution; Gene expression; Negative binomial distribution; Poisson distribution; Single-cell RNA-sequencing; Single-nucleus RNA-sequencing; Zero inflation.

*To whom correspondence should be addressed.

1. Introduction

Single-nucleus RNA-sequencing (snRNA-seq) is a common experimental technology to profile gene expression in frozen cells or cells that are hard to dissociate, such as in brain tissue (Lake *and others*, 2016; Slyper *and others*, 2020). Previous studies have shown that snRNA-seq offers substantial advantages over single-cell RNA-sequencing (scRNA-seq), including reduced dissociation bias (Habib *and others*, 2016; Bakken *and others*, 2018) and the ability to capture rare cell types (Wu *and others*, 2019). However, several questions remain on the degree to which existing tools used to analyze scRNA-seq data can be used in application of snRNA-seq data, including, (i) what is an appropriate genomic unit (e.g., exonic regions, intronic regions, etc.) to quantify reads for downstream analysis and (ii) what are appropriate probability distribution(s) to model measurement error, such as sampling variability, or noise. We begin by discussing these two topics in greater detail.

A standard approach to remove ribosomal RNA (rRNA) from scRNA-seq and snRNA-seq protocols, such as droplet-based technologies with unique molecular identifiers (UMIs) from the Drop-Seq (Macosko *and others*, 2015) or 10x Genomics Chromium (Zheng *and others*, 2017) systems, is to select polyadenylated RNA (polyA) transcripts using oligo (dT) primers. During the process of transcription, the gene is converted into a precursor mRNA (referred to as pre-mRNA) that contains both exonic and intronic regions. Mature mRNA is formed after the intronic regions have been spliced out of pre-mRNA, leaving only exonic regions. RNA processing happens in the nucleus so we expect mRNA existing outside the nucleus to be without introns, which have been experimentally verified (Cooper and Hausman, 2007; Ding *and others*, 2020; Lee *and others*, 2020). Hence, raw sequencing reads from scRNA-seq protocols are typically quantified using reference genomes or transcriptomes with only exonic regions, which has been shown to be sufficient for downstream analyses such as accurately classifying cell types (Habib *and others*, 2017; Bakken *and others*, 2018; Ding *and others*, 2020). However, it is unclear whether it is also sufficient to quantify reads from snRNA-seq protocols to only exonic regions for downstream analyses. If not, then how should intronic regions be incorporated, that is, whether reads should be quantified with exonic and intronic regions separately or mapped to full-length spliced and unspliced transcripts (pre-mRNA) in snRNA-seq data. For example, previous work has shown these choices need to be carefully considered for RNA velocity with scRNA-seq (Soneson *and others*, 2021) and may be also necessary for obtaining high-quality results downstream with snRNA-seq data (Bakken *and others*, 2018).

Furthermore, when considering quantified reads from intronic regions with snRNA-seq data, it has been previously suggested that there is a gene length bias (Chamberlin and Quinlan, 2020; Chamberlin *and others*, 2022). The source of this bias may lie with the enriched pre-mRNA in snRNA-seq data, as scRNA-seq data sequenced under UMI-based protocols, such as CEL-Seq, SMARTer, and CEL-Seq with InDrop, are not believed to exhibit a length bias (Phipson *and others*, 2017); we note this has not been specifically investigated with the Drop-Seq or 10x Genomics Chromium systems. However, the extent to which reads from intronic versus exonic regions contribute to the length bias and the potential causes of the bias are not well understood. In addition, it is unclear whether the length bias is a function of the full length of the unspliced transcript, consisting of both exonic and intronic regions, or the length of the spliced transcripts, consisting of exonic regions.

Next, given that an appropriate unit of quantification has been determined, another question is the choice of probability distributions to model snRNA-seq data. Much progress has been made on investigating the appropriateness of distributions to model scRNA-seq data (Hafemeister and Satija, 2019; Townes *and others*, 2019; Svensson, 2020; Choi *and others*, 2020; Ahlmann-Eltze and Huber, 2020; Sarkar and Stephens, 2021; Jiang *and others*, 2022; Choudhary and Satija, 2022) and an open question is whether similar distributions can be used to model measurement error or noise in droplet-based scRNA-seq data. In the context of scRNA-seq, previous work has argued that intronic reads

represent experimental and transcriptional noise (Harati *and others*, 2014) and are not usable in gene quantification (Zhao *and others*, 2018). As snRNA-seq enriches for transcripts in the nucleus with both mRNA and pre-mRNA, and scRNA-seq enriches for mostly mature mRNA, it is unclear if the measurement error in observed in snRNA-seq data is likewise affected by the inclusion of the intronic reads.

Furthermore, the consequences on the choice of appropriate probability distributions to model measurement error are not well understood. Previous work has demonstrated that droplet-based scRNA-seq data with UMIs are not zero-inflated and can be accurately modeled using Poisson, negative binomial (NB), or multinomial distributions (Townes *and others*, 2019; Hafemeister and Satija, 2019; Svensson, 2020). This was demonstrated using negative control data, where no biological heterogeneity is expected, by adding a controlled amount of RNA to each droplet. However, to the best of our knowledge, there has been no comparable analysis for the analysis of droplet-based snRNA-seq data while mapping reads to both introns and exons.

In this article, we first evaluate the choice of probability distributions to model measurement error in snRNA-seq data by creating pseudonegative control data sets. Next, we evaluate reference transcriptomes, which differ in how to include exonic and intronic regions, used in quantification mapping tools and consider the impact on cell type classification in downstream analyses. Then, we investigate and confirm the existence of a gene length bias in both intronic and exonic reads.

2. Results

Throughout, we used snRNA-seq data from two mouse cortices (Ding *and others*, 2020) measured on the 10x Genomics Chromium platform (Zheng *and others*, 2017), but we found consistent results with two other droplet-based snRNA-seq data sets, described below. We begin by creating pseudonegative control data sets (Figure S1 of the supplementary material available at *Biostatistics* online) by working with subsets of mouse cortex cell types to identify more homogenous populations of cells, where we expect less biological variation within a cell type than across cell types (Figure S2 of the supplementary material available at *Biostatistics* online). Throughout, we use *italicized* font when referring to data sets or different types of reference transcriptomes, referred to as *transcripts*, *preandmrna*, *intronicollapse*, and *intronseparate*. These reference transcriptomes differ in how to include exonic and intronic regions used in the `salmon alevin` (Srivastava *and others*, 2019) tool to perform quantification mapping of raw sequencing reads.

2.1. *Droplet-based single-nucleus RNA-seq data are not zero-inflated*

We begin by considering a pseudonegative control data set made with *preandmrna* reference transcriptome, which uses mRNA and pre-mRNA (Table S1 of the supplementary material available at *Biostatistics* online). However, at the end of this section, we show consistent results with other reference transcriptomes. Here, we use the pseudonegative control data to investigate whether the same probability distributions used to model measurement error in droplet-based scRNA-seq can be used for droplet-based snRNA-seq.

Common distributions to model measurement error in scRNA-seq with UMI counts include the binomial, Poisson, and NB distributions (Grün *and others*, 2014; Vieth *and others*, 2017; Svensson, 2020; Choudhary and Satija, 2022). Historically, it has been argued the scRNA-seq are "zero-inflated", where the fraction of observed zeros (or for a particular gene, the fraction of zeros across cells) in the single-cell counts is larger than what is expected under a specific distribution, such as the NB (Pierson and Yau, 2015; Risso *and others*, 2018; Jiang *and others*, 2022). The NB distribution

has two parameters: a mean (or rate) parameter (μ) and a dispersion parameter (ϕ). This dispersion parameter can be estimated as an overall dispersion parameter for a given data set or it can be estimated for each gene (or feature). Historically in application of bulk RNA-sequencing data, the interpretation of the ϕ parameter is to quantify how much extra biological variation is observed on top of technical variation (Robinson *and others*, 2010). However, in our application of snRNA-seq negative control data, we aim to estimate the parameters μ and ϕ to investigate if the NB distribution can be used to accurately capture the measurement error or technical variation observed from snRNA-seq data. For a random variable X that follows a NB with a given μ and a data set-specific ϕ, the variance is

$$\text{Var}[X] = \mu + \frac{1}{\phi} \cdot \mu^2$$

and the probability of observing a zero count is given by

$$P(X = 0|\mu, \phi) = 1 - \frac{\phi}{\phi + \mu}.$$

For a given cell type, we can compute for each gene (i) the empirical mean and empirical variance (commonly referred to as the "mean–variance relationship") (Chen *and others*, 2014) and (ii) the empirical mean and observed fraction of zero counts. We compare these observed quantities, the empirical variance and the observed fraction of zero counts, to what we expect them to be under the binomial, Poisson, and NB distributions. We also consider a NB distribution with a gene-specific dispersion parameter ϕ. Considering all four of these distributions, we calculate the Bayesian information criterion (BIC) (Burnham and Anderson, 2004) to assess the best model fit, where we sum up the log-likelihood across genes and assume the genes are independent. Lastly, we create quantile–quantile plots from a Pearson's goodness-of-fit test under the Poisson model to assess the fit of the Poisson distribution to the observed counts.

Using these pseudonegative control data sets with the excitatory neurons (Figure 1a–d), inhibitory neurons (Figure 1e–h) and astrocytes (Figure 1i–l), we found the empirical behavior of droplet-based snRNA-seq UMI counts is closely approximated by standard probability distributions (Figure 1a, e, i) and are not zero-inflated (Figure 1b, f, j). Most genes exhibit variances and fraction of zero counts that can be approximated by a binomial or Poisson distribution, but the NB distribution provides the best fit using BIC (Figure 1c, g, k). As noted in Figure 1a and b, the binomial and Poisson distributions provide nearly indistinguishable theoretical fits, but differ from the NB distribution at higher empirical means. However, the difference in BIC between the NB distribution and the binomial or Poisson distribution is driven primarily by a few genes (Figure S3 of the supplementary material available at *Biostatistics* online). The largest BIC values, and therefore the poorest fit, was found for the NB distribution with gene-specific overdispersion parameters ("G-S NB"), which indicates that the gain in likelihood from having an overdispersion parameter for every gene is outweighed by the BIC penalty on the increase in the number of parameters. Using a Pearson's goodness-of-fit test under a Poisson model, we also found that the gene counts exhibit deviations from the Poisson distribution (Figure 1d, h, l), which further suggest that the Poisson distribution does not adequately capture the technical variation in the counts. The Poisson model does provide a closer fit in the case of astrocytes (Figure 1l) and one potential reason is that they exhibit less heterogeneity in comparison to, for example, the inhibitory neurons, where it has been previously reported to have eight different subtypes of inhibitory neurons (compared to two astrocyte subtypes) in the human amygdala (Tran *and others*, 2021).

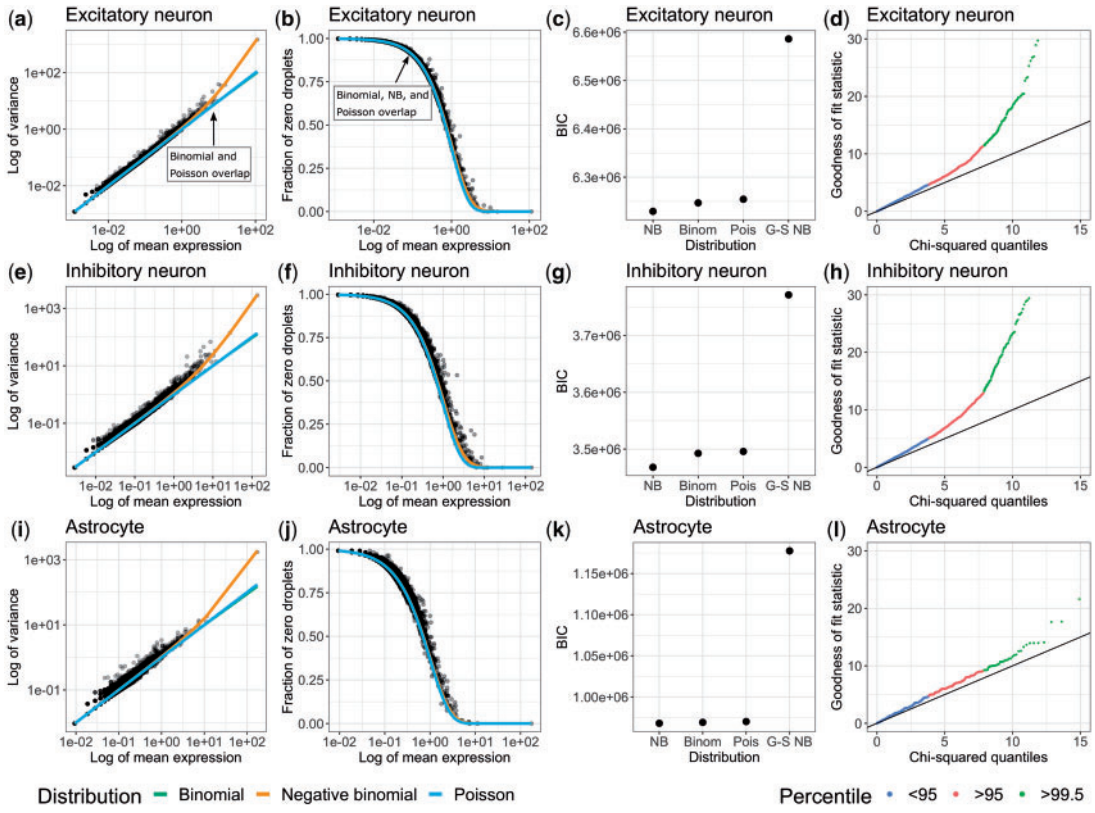

Fig. 1. **Droplet-based single-nucleus RNA-seq data is not zero-inflated.** Using subsets of cell types, **(a–d)** Excitatory neurons, **(e–h)** Inhibitory neurons, and **(i–l)** Astrocytes, from Cortex 1 in the mouse cortex data set (Ding *and others*, 2020), the leftmost column shows the log-transformed empirical mean (x-axis) and variance (y-axis) for each gene (dots) with the theoretical variance (lines). The second column shows the log-transformed empirical mean (x-axis) and observed fraction of zeros (y-axis) for each gene (dots) with the expected fraction of zeros under each distribution (lines). The third column shows BIC value across all genes (assuming the genes are independent) for each distribution: NB with one overdispersion parameter estimated for the data set, binomial (Binom), Poisson (Pois), and NB with gene-specific overdispersion parameters (G-S NB). The rightmost column shows the quantile–quantile plot under the Poisson model with the theoretical chi-squared quantile on the x-axis and the observed chi-squared statistic on the y-axis. Reads were quantified with the *preandmrna* reference transcriptome.

In addition to cell types, we found that these results also hold true across different biological replicates (the two mouse cortices) and reference transcriptomes (Figures S4–S6 and Table S2 of the supplementary material available at *Biostatistics* online) and for additional droplet-based snRNA-seq data sets sequenced using the 10x Genomics Chromium (Figure S7a–d of the supplementary material available at *Biostatistics* online) and the sNucDrop-Seq system (Figure S7e–h of the supplementary material available at *Biostatistics* online).

2.2. *Reference transcriptomes with intronic regions increases the total number of mapped reads*

We continue with the same snRNA-seq data from the mouse cortex (Ding *and others*, 2020), but here we consider four reference transcriptomes (*transcripts*, *preandmrna*, *introncollapse*, and *intronseparate*), which differ in how to include exonic and intronic regions, used for quantification

mapping with the `salmon alevin` (Srivastava *and others*, 2019) tool. The *transcripts* reference uses only the spliced transcripts as target sequences, while the other three quantification references additionally incorporate intronic regions as target sequences in different ways (Table S1 of the supplementary material available at *Biostatistics* online). We evaluate how the choice of the reference transcriptome can impact the total number of mapped reads, and the number of mapped reads to subsets of genes including protein coding genes and pseudogenes.

We found that the estimated library size for the *transcripts* reference was smaller than the estimated library sizes for the other three references (Figure 2a). We observe a similar disparity between references for the number of reads in protein-coding genes (Figure 2b). Interestingly, we found a decrease in the number of reads mapping to processed pseudogenes in the references with intronic reads (Figure 2c). This may occur if true pre-mRNA reads are mapped to other regions like processed pseudogenes in the *transcripts* reference due to the lack of intronic target sequences, but are correctly mapped to the intronic regions of genes in the other references. However, only minor differences in the number of mapped reads appear between the three references that incorporate intronic regions (*preandmrna*, *introncollapse*, and *intronseparate*). Finally, we also found an increase mapped reads to long non-coding RNA and antisense for the references that included intronic regions (Figure S8 of the supplementary material available at *Biostatistics* online). This suggests that for snRNA-seq, the primary difference with respect to total mapped reads is driven by the incorporation of intronic regions in target sequences or not, rather than the specific ways in which they are incorporated.

2.3. *Reference index impacts cell type classification*

An example of how the lower mapping rate in the *transcripts* reference affects downstream analysis is in application of cell type classification. To demonstrate this, we used the reference-based cell type annotation algorithm `SingleR` (Aran *and others*, 2019) to identify cell types using each of the four reference transcriptomes. We compared these cell type classification labels to the labels provided by the authors of the original article (Ding *and others*, 2020) (Figure 3a). We found that for the most common cell types (as classified by Ding *and others* (2020)), there is a high level of agreement with the `SingleR` cell type classification across most reference transcriptomes. However, we observe a higher discordance to the cell type classified by Ding *and others* (2020) in the *transcripts* reference, especially for certain cell types. For instance, among the nuclei labeled as astrocytes by the authors, more nuclei are labeled as quiescent neural stem cells (qNSCs) instead of astrocytes by `SingleR` in the *transcripts* reference compared to the others (*preandmrna*, *introncollapse*, and *intronseparate*). Because `SingleR` is a reference-based algorithm, it depends on reads mapping to known marker genes to accurately classify cell types. Upon further inspection of the marker genes for astrocytes and qNSCs, we found more reads mapping to astrocyte marker genes compared to qNSC marker genes for only the references that include intronic reads, among the nuclei labeled as astrocytes by the authors (Figure 3b). In contrast, using the *transcripts* reference, the ratio of counts in astrocyte marker genes compared to qNSC marker genes is close to 1 on average for the astrocyte nuclei, resulting in some of these nuclei being classified as astrocytes and others classified as qNSCs by `SingleR`. This result indicates that the increase in mapping rate with the inclusion of intronic reads, as described in the previous section, does not occur uniformly across all marker genes, and real biological signal may be lost when intronic reads are not included.

We also observed major differences in cell type classification between references for the oligodendrocyte progenitor cell (OPC) cell type. In the *transcripts* reference, most of the OPC nuclei are classified into one of several cell types, while with the other references, nearly half of the OPC nuclei were not assigned a label. Cells are not assigned a label by `SingleR` when there is not enough signal to unambiguously assign a cell type classification, for example, when a given cell or nuclei has an

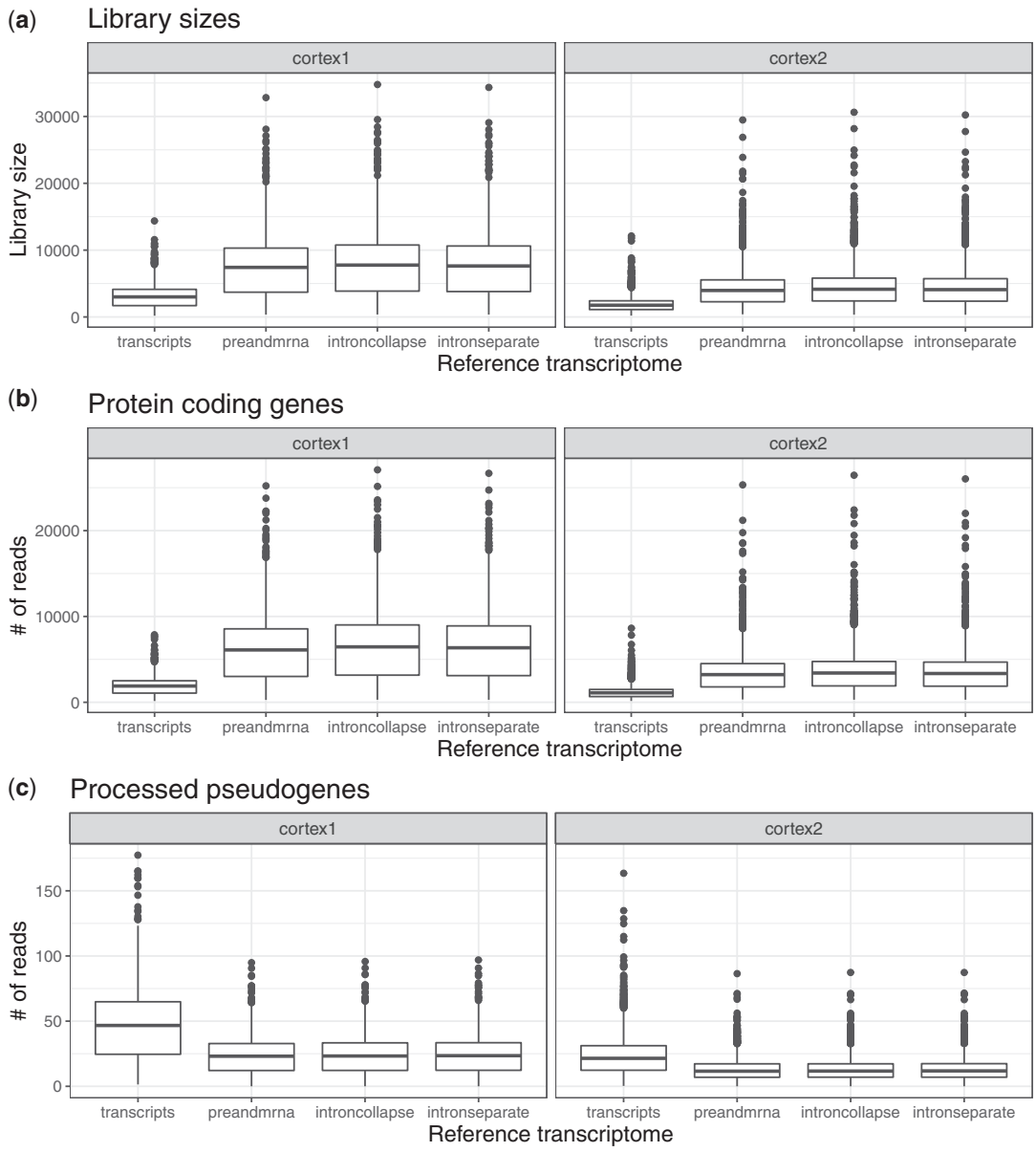

Fig. 2. **Incorporating intronic regions into reference transcriptomes leads to larger total mapped reads and reads mapping to protein coding genes.** Left and right columns represent two biological replicates [mouse Cortex1 (left) and Cortex2 (right)] that were sequenced and quantified using four reference transcriptomes (does not include intronic reads: *transcripts*; includes intronic reads: *preandmrna*, *introncollapse*, *intronseparate*). Boxplots of the number of **(a)** total UMIs for each nuclei (or library sizes), **(b)** reads mapped to protein-coding genes, and **(c)** reads mapped to processed pseudogenes.

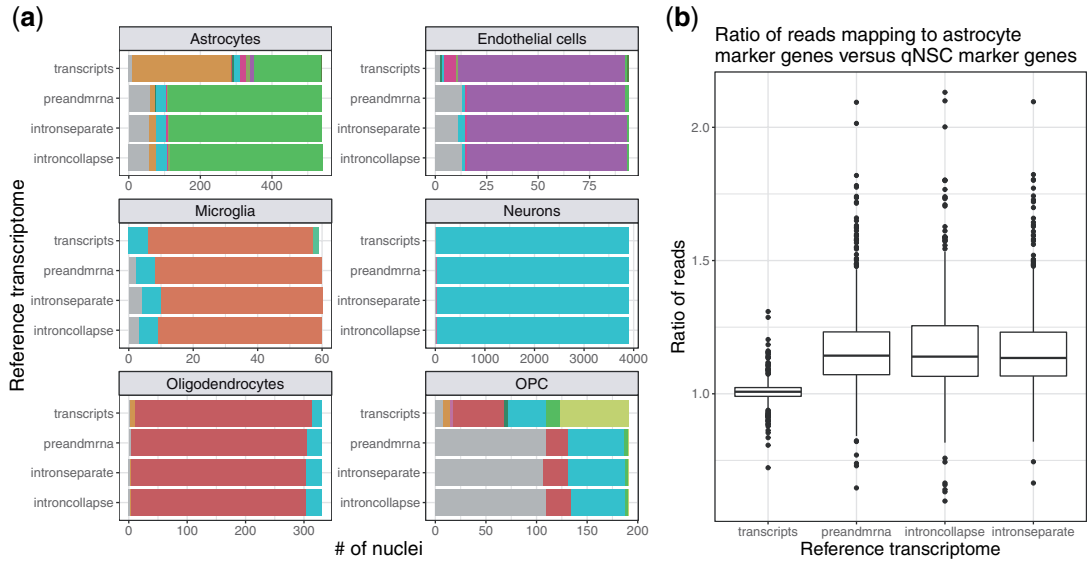

Fig. 3. **Choice of reference index can impact the classification of cell types in scRNA-seq data. (a)** For each cell type [labeled facets, classified by Ding *and others* (2020), the bar plots show the number of nuclei that are assigned to different cell types by the reference-based `SingleR` annotation algorithm in each reference (*transcripts*, *preandmrna*, *introncollapse*, or *intronseparate*). Excitatory neurons and inhibitory neurons are combined into one cell type named "Neurons" as the training data set used in `SingleR` does not distinguish between them. **(b)** For the nuclei classified as astrocytes by Ding *and others* (2020), the ratio of UMI counts in astrocyte marker genes to UMI counts in qNSC marker genes is elevated in reference indices that incorporate intronic regions (*preandmrna*, *introncollapse*, *intronseparate*) versus those that do not (*transcripts*).

expression profile equally similar to two or more cell types (Aran *and others*, 2019). Therefore, we observe that quantification choices can also influence cell type classification in less apparent ways, namely by determining whether a cell type label is assigned or not.

Similar to the number of mapped reads, the major differences in cell type classification remain between the choice of reference transcriptome to include intronic regions or not, with more minor differences among using references that include intronic regions. This shows that the specific ways of defining intronic regions in the reference transcriptome is less consequential than the choice to include or not include intronic regions.

2.4. *Droplet-based snRNA-seq data exhibit a gene length bias*

In this section, we continue with the same snRNA-seq data (Ding *and others*, 2020), but here we show that there is a gene length bias, namely we observe a higher level of expression for genes that are longer compared to genes that are shorter. This is surprising because it is assumed that scRNA-seq and snRNA-seq UMI-based protocols, unlike full-length transcript protocols, do not exhibit a gene length bias due to the polyA selection on the 3′ end of the mRNA molecule (Phipson *and others*, 2017; Vallejos *and others*, 2017; Zheng *and others*, 2017). Nevertheless, a length bias has been previously described in snRNA-seq data (Chamberlin and Quinlan, 2020; Chamberlin *and others*, 2022) and has been suggested to be caused by internal poly-A priming (below).

We group genes into ten bins by their preandmrna length, with each bin containing the same number of genes. The "preandmrna" length refers to the full gene length, which includes both the

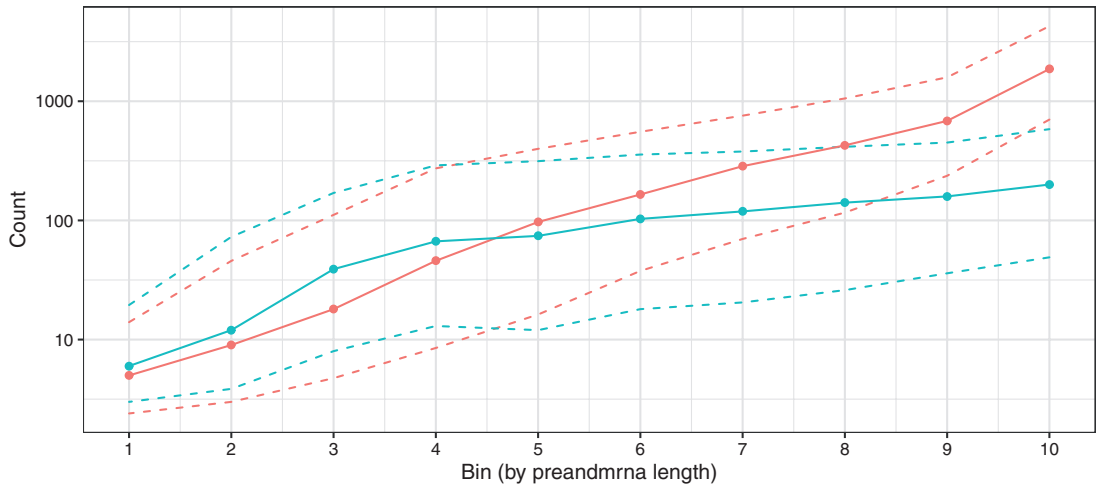

Fig. 4. **Droplet-based snRNA-seq data exhibit a gene length bias.** Genes are binned by their length into 10 equally sized bins (*x*-axis) where the smallest bin number corresponds to the shortest genes and the largest bin number corresponds to the longest genes. Using gene counts derived from either the *preandmrna* (red) or *transcripts* (blue) reference transcriptome, the distribution of gene counts across nuclei (*y*-axis) are shown with the median (solid points), and the 25th and 75th percentile (dashed lines). Genes are binned (*x*-axis) using the full gene length with both exons and introns, referred to as the "preandmrna length".

intronic and exonic regions of a gene (see Figure 9a for a comparison between "transcript" length and "preandmrna"). Most importantly, we observe a gene length bias, where the number of total counts per gene (sum of counts across all nuclei) increases with the gene length using both the *transcripts* and *preandmrna* references (Figure 4). The bias is stronger in the *preandmrna* reference, which suggests that the intronic reads play a major role in the length bias. We observe a similar trend using the *intronicollapse* and *intronseparate* references (Figure S9b and c of the supplementary material available at *Biostatistics* online). Furthermore, we confirm the result of a gene length bias using additional droplet-based snRNA-seq data sets sequenced using the 10x Genomics Chromium and sNucDrop-Seq system (Figure S10 of the supplementary material available at *Biostatistics* online) and show that it is therefore not unique to the 10x Genomics Chromium system.

Next, we explore potential causes for this bias. One previously described mechanism that could explain the length bias is internal priming (Chamberlin and Quinlan, 2020; 10x Genomics, 2021; Svoboda *and others*, 2022; Chamberlin *and others*, 2022). Here, the poly(dT) primer primes at an internal poly-A sequence rather than the poly-A mRNA tail. The end result is that a single transcript can erroneously get counted multiple times if there are multiple stretches of poly-A sequences. Now, internal poly-A sequences can theoretically occur in either exonic or intronic regions, but intronic regions are typically longer and thus more likely to contain internal poly-A sequences (Sakharkar *and others*, 2004). Thus, the effect of internal priming is likely to be stronger for intronic reads than exonic reads, which would then explain the result in Figure 4 with a greater length bias for the *preandmrna* reference than the *transcripts* reference.

However, we found that internal priming does not fully explain the observed length bias. Using the *preandmrna* reference (Figure 5a), we found that given the same preandmrna gene length, genes with at least one internal poly-A 8-mer have higher expression than genes without any poly-A 8-mers on average. This supports the idea that internal poly-A priming is driving at least some portion of

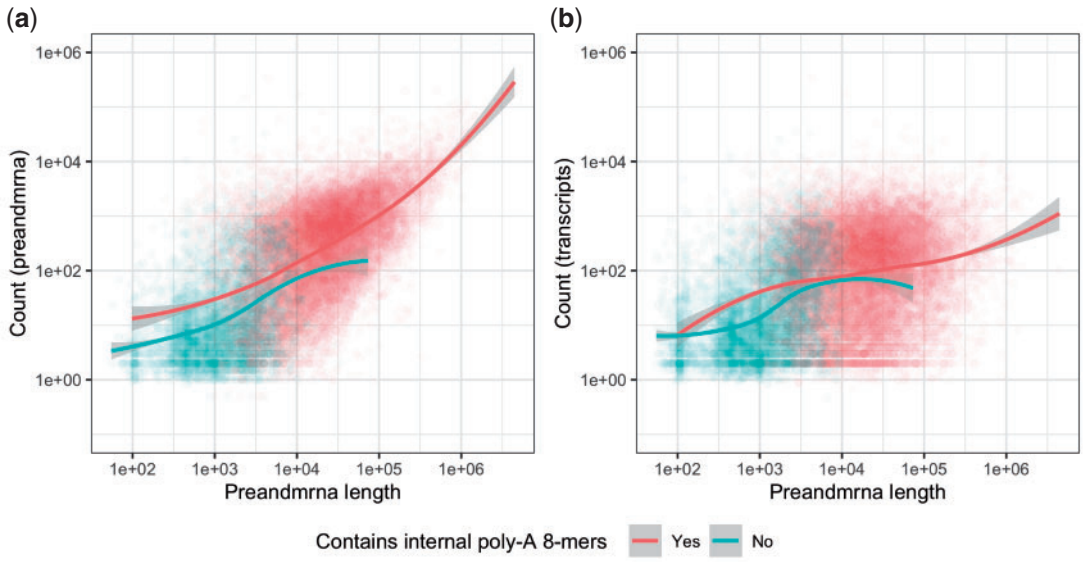

Fig. 5. **Comparison of gene length bias for genes with and without internal poly-A sequences.** Each point is a different gene and are colored red if they have at least one internal poly-A 8-mer and blue if they do not and a loess curve is drawn for each set of genes. The *x*-axis uses the full gene length with both exons and introns ("preandmrna" gene length). The *y*-axis plots the sum of reads across all nuclei (base-10 log scale) from the **(a)** *preandmrna* reference or **(b)** *transcripts* reference.

the gene length bias. However, we also see that among genes that do not have any internal poly-A 8-mers, a length bias can still be observed (blue line). This result holds true for different poly-A *n*-mer cut-offs (poly-A 6-mers, 10-mers, and 12-mers) (Figure S11 of the supplementary material available at *Biostatistics* online).

Applying the same analysis with the *transcripts* reference (Figure 5b), we observe that first, in comparison to the *preandmrna* reference, there is less of a difference between genes with internal poly-A 8-mers and genes without internal poly-A 8-mers. This suggests that, as expected, internal priming explains less of the bias among the exonic reads. Second, for the genes without any internal poly-A 8-mers, there is still a clear length bias, similar to what we saw in the *preandmrna* reference. As the *transcripts* reference only includes exonic regions, this suggests that a significant portion of the length bias observed in the *preandmrna* reference that is not explained by internal priming actually resides in the exonic reads.

To further investigate mechanisms, we compared the strength of the bias when using the preandmrna length, where including both intronic and exonic regions, versus the transcript length, where we only include the exonic regions. As the exonic region is a subset of the full gene, these two lengths are correlated (Figure S9a of the supplementary material available at *Biostatistics* online), hence we expect a length bias with both. The respective strengths of the bias, however, can tell us whether the causal mechanisms behind the bias is something we should expect to scale with the preandmrna length or the transcript length.

We found that the *preandmrna* reference exhibits a length bias that is more correlated with the preandmrna length than the transcript length. Under a base-10 log scale, the overall Pearson's correlation coefficient (r) between the counts and the preandmrna length is $r = 0.68$, while $r = 0.37$ between the counts and the transcript length (Figure S12a and b of the supplementary material

available at *Biostatistics* online). In comparison, with the *transcripts* reference, $r = 0.39$ between the counts and the preandmrna length and $r = 0.38$ between the counts and the transcript length (Figure S12c and d of the supplementary material available at *Biostatistics* online). This suggests that part of the length bias lies with the intronic reads in the *preandmrna* reference and is correlated with the length of the intronic region, something that a mechanism like internal priming would be consistent with. However, there is another part of the length bias that lies with the exonic reads and is less correlated with either the preandmrna or transcript length. This reinforces our previous conclusion that there are likely multiple sources for the length bias, which may be different for intronic reads versus exonic reads.

3. Discussion

Droplet-based snRNA-seq technologies are becoming the preferred technology to profile gene expression in frozen cells or cells that are hard to dissociate. With these new data come new statistical challenges that need to be addressed, including how to model these data. Here, we demonstrate that droplet-based snRNA-seq data are not zero-inflated and follow a NB distribution. These data can also be approximated by binomial and Poisson distributions. Our results suggest that statistical methods that depend on these assumptions, such as tools for batch correction (Satija *and others*, 2015) or differential expression analysis (Robinson *and others*, 2010; Anders and Huber, 2010; Risso *and others*, 2018) commonly used for scRNA-seq, can likewise be used for snRNA-seq. As a general example, our results demonstrate that a NB generalized linear model $g(Y) = X\beta$ can be used for snRNA-seq data, where Y are the counts and X are the variables of interest.

Furthermore, we show that choices in the reference transcriptomes used to perform quantification mapping of snRNA-seq data can impact both the fraction of reads mapped and downstream analyses, such as cell type classification. This is meaningful as different annotated cells can result in different biological interpretations of the same data. Standard quantification tools used for scRNA-seq are therefore not sufficient for analyzing snRNA-seq and the incorporation of intronic regions in the quantification of scRNA-seq data is an important consideration. In addition, we show that the choice of how intronic reads are included in quantification is less important than the choice to do so. Both in terms of library size and cell type classification, we found similar performance using reference transcriptomes that incorporate intronic regions.

With respect to cell type classification, we note that the higher agreement with cell type labels derived by the Ding *and others* (2020) authors in the reference transcriptomes that incorporate intronic regions (*preandmrna*, *intronicollapse*, and *intronseparate*) is not surprising given that they also include intronic regions in their quantification tool (Ding *and others*, 2020). However, we do not claim that concordance between the cell type labels imply that our assigned cell type labels are "correct." Instead, we simply demonstrate that there are significant differences in cell type classification that arise from the choice to include or not include intronic regions in quantification. Since the disparities in mapping rate between references that do not include intronic regions (*transcripts*) compared to those that do (*preandmrna*, *intronicollapse*, and *intronseparate*) lead to differences in counts that do not occur uniformly across all marker genes.

Across the references, we also observe a gene length bias in snRNA-seq. This bias is strongest when intronic regions are included, which may partly explain why a length bias has not been previously described with scRNA-seq data using the same sequencing technology. However, we also showed how the bias is not limited to intronic regions and is present in exonic regions, as well. Previously proposed mechanisms like internal priming do not fully explain the observed bias, particularly for exonic regions, and we leave further investigation of potential causes to future work.

Our work comes with limitations. First, as we use experimental mouse cortex nuclei data for our analysis, we can only create "pseudonegative" control data sets that are not completely biologically homogeneous. For this reason, our results can be considered as a maximum bound on the amount of overdispersion. Since we have found that droplet-based snRNA-seq data follow a NB distribution and can be approximated by the binomial or Poisson distribution in many cases, we expect that the true measurement error should be even lower than what we have found. Negative controls of technical replicates have previously been generated to study technical variability in scRNA-seq data (Zheng *and others*, 2017) and a similar data set for snRNA-seq can, in principle, be used to verify our conclusions.

Other limitations are that we only investigated two droplet-based high-throughput experimental technologies to capture gene expression. However, we do not expect our results on the distributions for measurement error of snRNA-seq data set to change with droplet-based protocols with UMIs, as similar analyses for scRNA-seq found consistent results across three platforms (Svensson, 2020).

We note that the magnitude and nature of the effect of quantification choices on downstream analyses may vary depending on the data set. As we observed in our data set, some cell type classifications appear to be more affected by the inclusion of intronic reads than others. However, since we also observed large disparities in library size depending on whether intronic reads are included in the reference transcriptome and this phenomenon is likely to be agnostic to different snRNA-seq data sets, our results suggest that quantification choices will be informative for downstream analyses of snRNA-seq data. Given the relative ease of including introns in quantification with tools like `salmon alevin` and the significant loss in information when they are not included, the inclusion of intronic reads is a crucial step in the analysis of snRNA-seq data.

4. METHODS

4.1. *Data*

The mouse whole cortex nuclei data set was generated by Ding *and others* (2020) using the 10x Genomics Chromium platform (Zheng *and others*, 2017). Two experiments were performed, resulting in two biological replicates (Cortex1 and Cortex2). Each of these experiments was run on a platform with two flow cells with four lanes each, resulting in eight SRA files for each experiment. After running `salmon alevin` (Srivastava *and others*, 2019) to map the reads with the different transcriptome indices described in the next section, we read in the nucleus $\times$ gene UMI count matrices as a `SingleCellExperiment` object using the `tximeta` (Love *and others*, 2020) and `SingleCellExperiment` R packages (Amezquita *and others*, 2020).

Two additional snRNA-seq data sets were used to validate the results from this study. UMI count matrices were downloaded for mouse brain nuclei isolated with 10x Genomics Chromium and processed using Cell Ranger (10x Genomics, 2022). Mouse kidney nuclei generated using the sNuc-Dropseq protocol were aligned with STAR (Wu *and others*, 2019). The number of genes $\times$ number of nuclei is 32 885 $\times$ 7377 for the 10x Genomics mouse brain data set and 19 713 $\times$ 3011 for the sNuc-Dropseq mouse kidney data set.

4.2. *Quantification with four sets of reference indices*

We started with Sequence Read Archive (SRA) data downloaded from the Gene Expression Omnibus with accession number https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE132044 and converted them into FASTQ files using the SRA toolkit. The FASTQ files, along with a reference transcriptome, are fed into `alevin` for quantification (Srivastava *and others*, 2019). To create the

reference transcriptome, we processed Gencode reference files, in particular the GRCm38 primary assembly FASTA file and the Gencode vM25 gene annotation GTF file.

In total, four reference transcriptomes were created. Each of these reference transcriptomes differ in that they incorporate intronic regions into its target sequences in different ways. We define the following types of target sequences. First, we start with the transcript sequences, which are defined from the downloaded genome sequence and GTF file and consist of the exonic regions for each transcript. These are what we call "spliced transcripts," and each transcript is a separate target sequence. We can create "unspliced transcripts" as target sequences by re-adding the intronic regions between any two exonic regions in a transcript. The length of each unspliced transcript must therefore be greater than or equal to its corresponding spliced transcript. Lastly, intronic regions for a given transcript or gene can themselves be used as target sequences. Based on the work of Soneson *and others* (2021), we define the introns in two ways: "separate" or "collapse." In the "separate" approach, the intron target sequences are defined as the intronic regions from a transcript of a given gene. In the "collapse" approach, the intron target sequences are defined to be the intronic regions of a gene that are not exonic in any isoforms of the gene. Thus, while the "separate" approach allows for intron target sequences to overlap with exonic regions of other transcripts, the "collapse" approach does not. For all intronic target sequences, a flanking length of 50 bp (read length) is also added to account for reads that map to exon/intron junctions.

The different target sequences used to create the reference transcriptome form the basis of our four distinct quantification reference transcriptomes (Table S1 of the supplementary material available at *Biostatistics* online). In all four references, the complete genome sequence was also added to the reference transcriptome to create a decoy-aware transcriptome and minimize the spurious mapping of reads to intergenic regions.

4.3. *Preprocessing and quality control*

For quality control, we use `perCellQCMetrics` from the `scater` R package (McCarthy *and others*, 2017). The quality control procedure removes nuclei with low library sizes or few expressed genes and discards genes with zero counts across all nuclei. After these steps, the number of genes $\times$ number of nuclei is 27 651 $\times$ 5612 for the *transcripts* reference, 31 701 $\times$ 5680 for the *preandmrna* reference, 31 670 $\times$ 5686 for the *introncollapse* reference, and 31 661 $\times$ 5686 for the *intronseparate* reference. The quantification and preprocessing steps are summarized in Figure S1 of the supplementary material available at *Biostatistics* online.

For exploratory analysis, we run principal component analysis (PCA) on the log normalized counts. The counts are normalized using size factors computed by `calculateSumFactors` from the `scran` R package (Lun *and others*, 2016), and PCA was performed using the `scater` R package (McCarthy *and others*, 2017).

4.4. *Description of methods for distribution plots*

For a description of the methods used for distribution plots, see Note S1 of the supplementary material available at *Biostatistics* online.

4.5. *Cell type classification*

For our analysis on distributions, we separate the mouse cortex nuclei using cell type labels computationally generated by Ding *and others* (2020). We compare these cell type labels to cell type labels generated using the `SingleR` R package and the built-in reference `MouseRNAseqData()`

(Aran *and others*, 2019). Specifically, `SingleR` cell type labels are generated by comparing gene expression to the expression profile of the reference cells across marker genes. Each mouse cortex nuclei is assigned a cell type label based on similarity with Spearman correlation and labels are pruned by discarding ambiguous labels.

4.6. *Gene length bias*

We define the gene length to be either the "preandmrna" length, which is the full-length transcript with both exonic and intronic regions, or the "transcript" length, which is the transcript with only exonic regions. For each gene, we calculate the sum of the expression counts across all nuclei in the data set.

To count the number of poly-A sequences for each gene, we first count the number of poly-A n-mers in every transcript for a given gene. Then we summarize at the gene-level by taking the maximum number of poly-A n-mers across all transcripts (spliced and unspliced) for a given gene. This allows us to separate genes into two groups, that is, genes that do not have any poly-A n-mers for any of its transcripts and genes that have at least one poly-A n-mer in at least one of its transcripts.

Code and software availability

All analyses and code were conducted in the `R` programming language. Code for reproduction of all plots in this manuscript is available at https://github.com/stephaniehicks/quantify-snrna.

Supplementary material

Supplementary material is available online at http://biostatistics.oxfordjournals.org.

Acknowledgments

The authors would like to thank the Joint High Performance Computing Exchange (JHPCE) for providing computing resources. set COI statement below Acknowledgments
Conflict of Interest: None declared.

Funding

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number (R01GM121459); the National Human Genome Research Institute of the National Institutes of Health under the award number (R00HG009007 and CZF2019-002443); and the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation (CZF2018-183446).

Author contributions

A.K. and SCH performed all the data analyses. AK and SCH wrote the manuscript. All authors read and approved the final manuscript.

References

10x Genomics. (2021). Technical note - interpreting intronic and antisense reads in 10x genomics single cell gene expression data, document number cg000376. https://www.10xgenomics.com/support/single-cell-gene-expression/documentation/steps/sequencing/interpreting-intronic-and-antisense-reads-in-10-x-genomics-single-cell-gene-expression-data.

10x Genomics. (2022). 5k Adult Mouse Brain Nuclei Isolated with Chromium Nuclei Isolation Kit, 7.0.0, Single Cell Gene Expression Dataset by Cell Ranger. https://www.10xgenomics.com/resources/datasets/5k-adult-mouse-brain-nuclei-isolated-with-chromium-nuclei-isolation-kit-3-1-standard.

Ahlmann-Eltze, C. and Huber, W. (2020). glmGamPoi: fitting Gamma-Poisson generalized linear models on single cell count data. *Bioinformatics* **36**(24), 5701–5702.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology* **11**, R106.

Amezquita, R., Lun, A., Becht, E., Carey, V., Carpp, L., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C., *and others.* (2020). Orchestrating single-cell analysis with Bioconductor. Nature Methods **17**, 137–145. https://www.nature.com/articles/s41592-019-0654-x.

Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R. P., Wolters, P. J., Abate, A. R., Butte, A. J. *and others*. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology* **20**, 163–172.

Bakken, T. E., Hodge, R. D., Miller, J. A., Yao, Z., Nguyen, T. N., Aevermann, B., Barkan, E., Bertagnolli, D., Casper, T., Dee, N. *and others*. (2018). Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PLoS One* **13**, e0209648.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: understanding aic and bic in model selection. *Sociological Methods & Research* **33**, 261–304.

Chamberlin, J. and Quinlan, A. (2020). *Systematic Gene Detection Bias in Singlenucleus RNA-seq*. Biological Data Science at Cold Spring Harbor Laboratories.

Chamberlin, J. T., Lee, Y., Marth, G. T. and Quinlan, A. R. (2022). Variable RNA sampling biases mediate concordance of single-cell and nucleus sequencing across cell types. bioRxiv doi: https://doi.org/10.1101/2022.08.01.502392.

Chen, Y., Lun, A. TL and Smyth, G. K. (2014). *Differential Expression Analysis of Complex RNA-seq Experiments using edgeR*. Springer. https://link.springer.com/chapter/10.1007/978-3-319-07212-8_3.

Choi, K., Chen, Y., Skelly, D. A. and Churchill, G. A. (2020). Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics. *Genome Biology* **21**, 183.

Choudhary, S. and Satija, R. (2022). Comparison and evaluation of statistical error models for scRNA-seq. *Genome Biology* **23**, 27.

Cooper, G. M. and Hausman, R. E. (2007). *The Cell: A Molecular Approach* 2nd Edition. Sunderland (MA): Sinauer Associates.

Ding, J., Adiconis, X., Simmons, S. K., Kowalczyk, M. S., Hession, C. C., Marjanovic, N. D., Hughes, T. K., Wadsworth, Marc H., Burks, T., Nguyen, L. T., *and others*. (2020). Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nature Biotechnology* **38**, 737–746.

Grün, D., Kester, L. and Van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. *Nature Methods* **11**, 637–640.

Habib, N., Avraham-Davidi, I., Basu, A., Burks, T., Shekhar, K., Hofree, M., Choudhury, S. R., Aguet, F., Gelfand, E., Ardlie, K. *and others*. (2017). Massively parallel single-nucleus RNA-seq with dronc-seq. *Nature Methods* **14**, 955–958.

Habib, N., Li, Y., Heidenreich, M., Swiech, L., Avraham-Davidi, I., Trombetta, J. J., Hession, C., Zhang, F. and Regev, A. (2016). Div-Seq: single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science (New York, N.Y.)* **353**, 925–928.

Hafemeister, C. and Satija, R. (2019, 12). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology* **20**, 296.

Harati, S., Phan, J. H. and Wang, M. D. (2014). Investigation of factors affecting RNA-seq gene expression calls. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society* **2014**, 5232–5235.

Jiang, R., Sun, T., Song, D. and Li, J. J. (2022). Statistics or biology: the zero-inflation controversy about scRNA-seq data. *Genome Biology* **23**, 31.

Lake, B. B., Ai, R., Kaeser, G. E., Salathia, N. S., Yung, Y. C., Liu, R., Wildberg, A., Gao, D., Fung, H.-L., Chen, S., *and others*. (2016). Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science (New York, N.Y.)* **352**, 1586–1590.

Lee, S., Zhang, A. Y, Su, S., Ng, A. P., Holik, A. Z., Asselin-Labat, M.-L., Ritchie, M. E. and Law, C. W. (2020). Covering all your bases: incorporating intron signal from RNA-seq data. *NAR Genomics and Bioinformatics* **2**, lqaa073.

Love, M. I., Soneson, C., Hickey, P. F., Johnson, L. K., Pierce, N. T., Shepherd, L., Morgan, M. and Patro, R. (2020). Tximeta: reference sequence checksums for provenance identification in RNA-seq. *PLoS Computational Biology* **16**, e1007664.

Lun, A. T. L., McCarthy, D. J. and Marioni, J. C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research* **5**, 2122.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., *and others*. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214.

McCarthy, D. J., Campbell, K. R., Lun, A. T. L. and Wills, Q. F. (2017, 04). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179–1186.

Phipson, B., Zappia, L. and Oshlack, A. (2017, April). Gene length and detection bias in single cell RNA sequencing protocols. *F1000Research* **6**, 595.

Pierson, E. and Yau, C. (2015). Zifa: dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology* **16**, 1–10.

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. and Vert, J.-P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications* **9**, 284.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140.

Sakharkar, M. K., Chow, V. T. K. and Kangueane, P. (2004). Distributions of exons and introns in the human genome. *In Silico Biology* **4**, 387–393.

Sarkar, A. and Stephens, M. (2021). Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nature Genetics* **53**, 770–777.

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* **33**, 495–502.

Slyper, M., Porter, C. B. M., Ashenberg, O., Waldman, J., Drokhlyansky, E., Wakiro, I., Smillie, C., Smith-Rosario, G., Wu, J., Dionne, D., *and others*. (2020). A single-cell and single-nucleus RNA-seq toolbox for fresh and frozen human tumors. *Nature Medicine* **26**, 792–802.

Soneson, C., Srivastava, A., Patro, R. and Stadler, M. B. (2021). Preprocessing choices affect RNA velocity results for droplet scRNA-seq data. *PLoS Computational Biology* **17**, e1008585.

Srivastava, A., Malik, L., Smith, T., Sudbery, I. and Patro, R. (2019). Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. *Genome Biology* **20**, 65.

Svensson, V. (2020). Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology* **38**, 147–150.

Svoboda, M., Frost, H. R. and Bosco, G. (2022). Internal oligo(dT) priming introduces systematic bias in bulk and single-cell RNA sequencing count data. *NAR Genomics and Bioinformatics* **4**, lqac035.

Townes, F. W., Hicks, S. C., Aryee, M. J. and Irizarry, R. A. (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology* **20**, 295.

Tran, M. N., Maynard, K. R., Spangler, A., Huuki, L. A., Montgomery, K. D., Sadashivaiah, V., Tippani, M., Barry, B. K., Hancock, D. B., Hicks, S. C. *and others*. (2021). Single-nucleus transcriptome analysis reveals cell-type-specific molecular signatures across reward circuitry in the human brain. *Neuron* **109**, 3088–3103.

Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S. and Marioni, J. C. (2017). Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nature Methods* **14**, 565–571.

Vieth, B., Ziegenhain, C., Parekh, S., Enard, W. and Hellmann, I. (2017). powsimR: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics* **33**, 3486–3488.

Wu, H., Kirita, Y., Donnelly, E. L. and Humphreys, B. D. (2019). Advantages of single-nucleus over single-cell RNA sequencing of adult kidney: rare cell types and novel cell states revealed in fibrosis. *Journal of the American Society of Nephrology* **30**, 23 LP – 32.

Zhao, S., Zhang, Y., Gamini, R., Zhang, B. and von Schack, D. (2018). Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Scientific Reports* **8**, 4781.

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., *and others*. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8**, 14049.

[*Received May 20, 2022; revised March 22, 2023; accepted for publication April 19, 2023*]