Accurate quantification of single-nucleus and single-cell RNA-seq transcripts

Kristján Eldjárn Hjörleifsson*[1], Delaney K. Sullivan*[2,3], Guillaume Holley[5], Páll Melsted[4,5], Lior Pachter[2]

[1] Department of Computing and Mathematical Sciences, California Institute of Technology
[2] Division of Biology and Biological Engineering and Department of Computing and Mathematical Sciences, California Institute of Technology
[3] UCLA-Caltech Medical Scientist Training Program, David Geffen School of Medicine, University of California, Los Angeles
[4] School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland
[5] deCODE Genetics/Amgen Inc., Reykjavik, Iceland

* Equal contribution

Address correspondence to lpachter@caltech.edu

Abstract

The presence of both nascent and mature mRNA molecules in single-cell RNA-seq data leads to ambiguity in the notion of a “count matrix”. Underlying this ambiguity, is the challenging problem of separately quantifying nascent and mature mRNAs. We address this problem by relating *k*-mer assignment to read assignment in the context of different classes of molecules, and describe a unified approach to quantifying both single-cell and single-nucleus RNA-seq.

Introduction

The utility of single-cell RNA-seq measurements for defining cell types has represented a marked improvement over bulk RNA-seq, and is driving rapid adoption of single-cell RNA-seq assays (Zeng 2022). Another application of single-cell RNA-seq that is not possible with bulk RNA-seq is the study of cell transitions and transcription dynamics, even via snapshot single-cell RNA-seq experiments (Gorin et al. 2022). This novel application of single-cell RNA-seq is based on the quantification of both nascent and mature mRNAs (Figure 1a), lending import to the computational problem of accurately quantifying these two modalities. The challenge of quantifying nascent mRNAs in addition to mature mRNAs has also been brought to the fore with single-nucleus RNA-seq (Kuo, Hansen, and Hicks 2022).

The difficulty in quantifying both nascent and mature transcripts from single-cell RNA-seq data stems from the fact that sequenced reads are typically much shorter than transcripts, and therefore there can be ambiguity in classification of reads as originating from mature mRNAs vs. their nascent precursors (Figure 1b). Reads that span a junction, i.e. cover two exons separated

by an intron, must originate from a mature mRNA (Ⓜ), whereas reads containing sequence unique to an intron must originate from a nascent mRNA (Ⓝ), however there are many reads for which it is impossible to know whether they originated from a nascent or mature transcript (Ⓝ|Ⓜ). Furthermore, the way in which these cases are resolved depends on whether reads have been mapped to a whole genome, or directly to transcript sequences derived from annotations of the genome. The former approach lends itself well to identifying reads originating from nascent transcripts, as the genome includes all non-coding sequences, however the latter approach is superior for identifying spliced reads, because the sequence of processed transcripts is present in the reference being mapped to. Furthermore, methods that rely on *k*-mer matching to speed up alignment must account for the distinction between *k*-mer ambiguity and read ambiguity (Figure 1c), and this distinction has not been carefully accounted for in existing *k*-mer based single-cell RNA-seq pre-processing workflows (Melsted et al. 2021; He et al. 2022).

As a result of these complexities, existing single-cell RNA-seq quantification algorithms provide a smorgasbord of options for users, but confusing, or at times contradictory, guidance on how to quantify single-cell RNA-seq, with unfortunate implications for analysis (Soneson et al. 2021). For example,the popular Cell Ranger software for quantifying single-cell RNA-seq generated with 10x Genomics machines, in its first six releases only quantified mature RNAs, and not nascent transcripts, and a separate program was required for generating nascent molecule counts (La Manno et al. 2018). Moreover, the Cell Ranger quantification is based on an assumption that all reads that are ambiguous as to their origin from nascent or mature transcripts, are always counted as being derived from mature transcripts. Alevin-fry offers a large number of different quantification modes (He et al. 2022), and there are significant asymmetries in the quantification of single-cell vs. single-nucleus RNA-seq. For single-nucleus RNA-seq, typically all reads mapping to gene bodies are included in generation of a single count matrix, regardless of whether the reads originate from mature or nascent transcripts. For single-cell RNA-seq, great care is taken to avoid the counting of reads definitely originating from nascent transcripts, and only reads originating from mature transcripts, or ambiguous reads, are included in count matrices (Kaminow, Yunusov, and Dobin 2021; He et al. 2022).

We address these shortcomings and inconsistencies based on a novel *k*-mer based method we develop for resolving reads as to their originating source. By utilizing *k*-mers, our approach has the benefit of being efficient as it is compatible with pseudoalignment, and we show via an implementation in the kallisto software (Bray et al. 2016; Melsted et al. 2021) that it yields a fast and efficient approach for quantification. Crucially, we introduce an approach to quantification of single-nucleus RNA-seq that focuses on the nascent transcripts, thereby mirroring the approach for quantifying single-cell RNA-seq that focuses on mature transcripts. This places the two assays on a (computationally) level playing field.

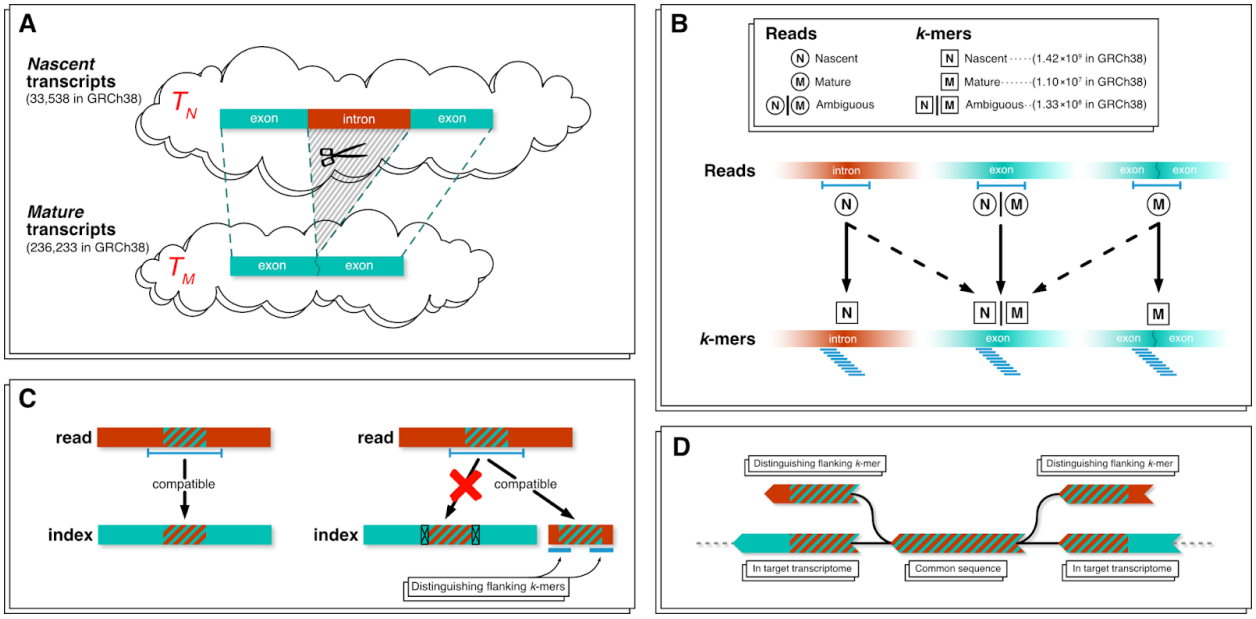

Figure 1. A. In any cell, there exist two sets of transcripts: the *nascent* set from which the introns have not yet been spliced, and the *mature* one, from which they have. The number of nascent and mature transcripts was obtained from version 104 of the Ensembl annotation of the human genome (Cunningham et al. 2022). **B.** Each read in an RNA-seq experiment is either explicitly Ⓝ, an expression of a nascent transcript, if it contains at least a part of a non-retained intronic sequence, Ⓜ, explicitly an expression of a mature transcript, if it contains an exon-exon boundary, or Ⓝ|Ⓜ, ambiguous, if it occurs in the interior of an exon. Furthermore, the *k*-mers that constitute an Ⓝ read may be [N], nascent or [N]|[M], ambiguous, if they also occur in an exon, and the *k*-mers that constitute a Ⓜ, mature read may be [M], mature if they contain the exon-exon junction or [N]|[M], ambiguous, if they are an interior *k*-mer in an exon. The number of unique [N], [M], and [N]|[M] *k*-mers in the human transcriptome was calculated from Ensembl version 104. **C.** A non-transcriptomic read containing a subsequence of length greater than *k,* which also occurs in a transcript in the target transcriptome will get attributed to that transcript. Distinguishing flanking *k*-mers (DFKs) can be used to determine whether a read compatible with a reference transcriptome may have originated from elsewhere in the genome. Using the entire set of GRCh38 scaffolds to construct a D-list for a kallisto index built from the mature transcripts in version 104 of the Ensembl annotation yields 7,192,804 DFKs. **D.** A *de Bruijn* graph representation of DFKs.

Results

To facilitate quantification of both nascent and mature mRNA transcripts, we distinguished *k*-mers into three categories, analogous to the three categories used for reads: [N], [M] or [N]|[M] (Figure 2b). A read can be classified as Ⓝ, Ⓜ or Ⓝ|Ⓜ based on the classification of its constituent *k*-mers (Methods). In order to classify *k*-mers without indexing all non-coding sequences, we utilized a D-list (Methods), which consists of distinguishing flanking *k*-mers (DFKs) that can be used to definitively assign a *k*-mer to a category (Methods). A read is classified as Ⓝ if it has at least one [N] *k*-mer, as Ⓜ if it has at least one [M] *k*-mer, and as Ⓝ|Ⓜ

otherwise. Note that a read cannot have both 🅝 and 🅜 *k*-mers, and that the DFKs suffice to classify all relevant *k*-mers for classifying a read (Methods).

To validate our method for classifying reads, we generated 5,000,000 mature reads and 5,000,000 nascent reads directly from the respective transcripts without error, and assessed whether we could correctly assign reads when using kallisto with a D-list included in the index (Figure 2, Methods). We found that using a D-list, we can reliably classify reads as nascent, mature or ambiguous, the latter one by identifying reads that are classified as both mature and nascent. The alevin-fry method failed to classify all reads correctly. In particular, it assigned all Ⓝ|Ⓜ ambiguous reads, e.g. reads that are contained within a single exon, to the "spliced" count matrix corresponding which counts mature molecules. This leads to an asymmetry between quantifications of single-cell RNA-seq data and single-nucleus RNA-seq data, where for the latter, alevin-fry (He et al. 2022), STARsolo (Kaminow, Yunusov, and Dobin 2021) and Cell Ranger (unpublished) always classify ambiguously mapped reads as mature regardless of assay. We benchmarked alevin-fry, and STARsolo in both "Gene/GeneFull"-mode and "Velocyto"-mode. STARsolo quantification in "Velocyto"-mode of the simulated data (Figure 2C) deviated from the ground truth to such an extent that comparisons with kallisto and alevin-fry were meaningless. When processing reads from mature transcripts, STARsolo in "Velocyto"-mode only mapped 825,141 (82.5%) of the reads. When processing reads from nascent transcripts STARsolo only mapped 847,392 (84.7%) of the reads, and of those only 85.4% were mapped correctly.

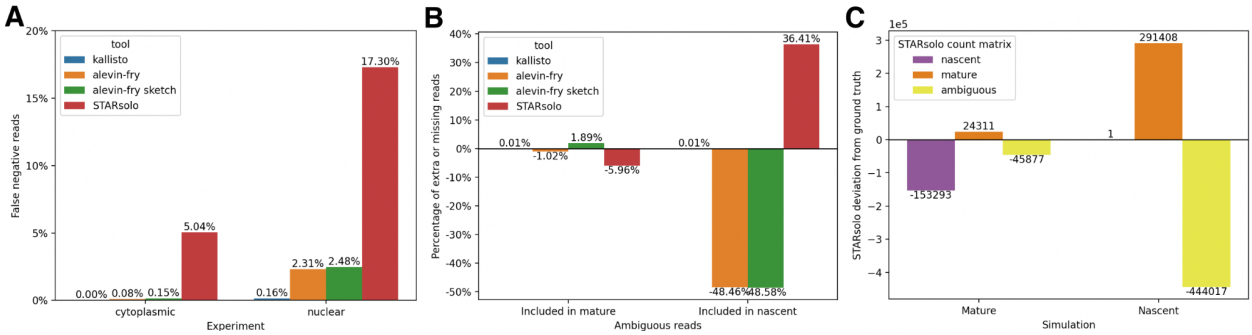

Figure 2. A. Left: the percentage of Ⓜ reads, not correctly identified as mature by kallisto, alevin-fry, and STARsolo in "Gene"-mode in a simulated single-cell RNA-seq experiment. Right: the percentage of Ⓝ reads, not correctly identified as nascent by kallisto, alevin-fry, and STARsolo in "GeneFull"-mode in a simulated single-nucleus RNA-seq experiment. **B.** Alevin-fry attributes ambiguous reads, e.g. reads that are contained within a single exon, to mature transcripts, leading to an asymmetry between quantifications of simulated single-cell, and single-nucleus RNA-seq reads. **C.** STARsolo quantification in "Velocyto"-mode of the simulated data. Left: simulated experiment with 1,000,000 reads containing no errors, from mature transcripts, 59,850 of which were ambiguous, occurring in both nascent and mature transcripts. Right: simulated experiment with 1,000,000 reads containing no errors, from nascent transcripts, of which 583,914 were ambiguous.

To better understand the performance of kallisto on data that includes errors, we assessed its performance using a simulation framework developed by the authors of STARsolo (Kaminow, Yunusov, and Dobin 2021). In that simulation framework, errors were introduced into reads at 0.5% error rate, and reads were simulated from both coding and non-coding genomic sequence to mimic the presence of both nascent and mature transcripts in single-cell RNA-seq experiments. We also followed the assessment framework of (Kaminow, Yunusov, and Dobin 2021), and compared the results of kallisto to STARsolo and alevin-fry. We found that kallisto performed similarly to STARsolo in a simulation containing multi-mapping reads, i.e. reads that align well to two or more distinct transcripts, and outperformed alevin-fry (Figure 3.B). In another simulation, containing no multi-mapping reads, STARsolo performed marginally better than both kallisto and alevin-fry (Figure 3.A). The poor performance of kallisto without the D-list is due to the unreasonable assessment procedure of (Kaminow, Yunusov, and Dobin 2021), which omitted assessment of true negatives (Methods). Note that correct calculations of the Spearman coefficient yield a negligible difference between kallisto, STARsolo, and alevin-fry (Supplementary table 2).

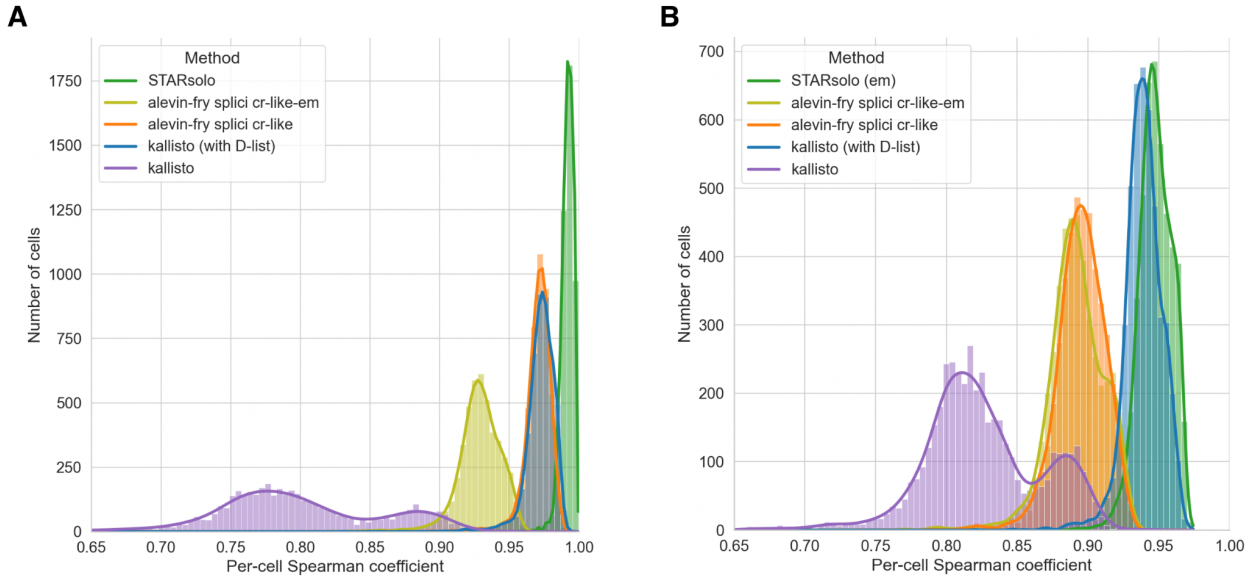

Figure 3. The Spearman coefficient for the correlation between the simulation ground truth expression and the expression quantified by kallisto, alevin-fry, and STARsolo in the simulation framework developed by Kaminow, Yunusov, and Dobin (2021). **A.** Reads from 4,548 cells, containing no multi-gene reads. **B.** Reads from 4543 cells, including multi-gene reads.

To understand the implications of correct classification of reads into the nascent transcript, mature transcript, or ambiguous categories, we examined the correlation in gene counts with and without the use of a D-list on both single-cell RNA-seq and single-nucleus RNA-seq data (Methods). The overall result was not materially different for single-cell RNA-seq (Figure 4A), with the Pearson correlation between kallisto and kallisto with the D-list at 0.99, corroborating the results of (Sina Booeshaghi and Pachter 2021). Similarly, kallisto with the D-list is highly similar to alevin-fry (Figure 4B) and STARsolo (Figure 4c). Even though quantification with the

D-list does not affect quantifications much, its use does not affect running times much (Figure 5A) so it can be used routinely anyway.

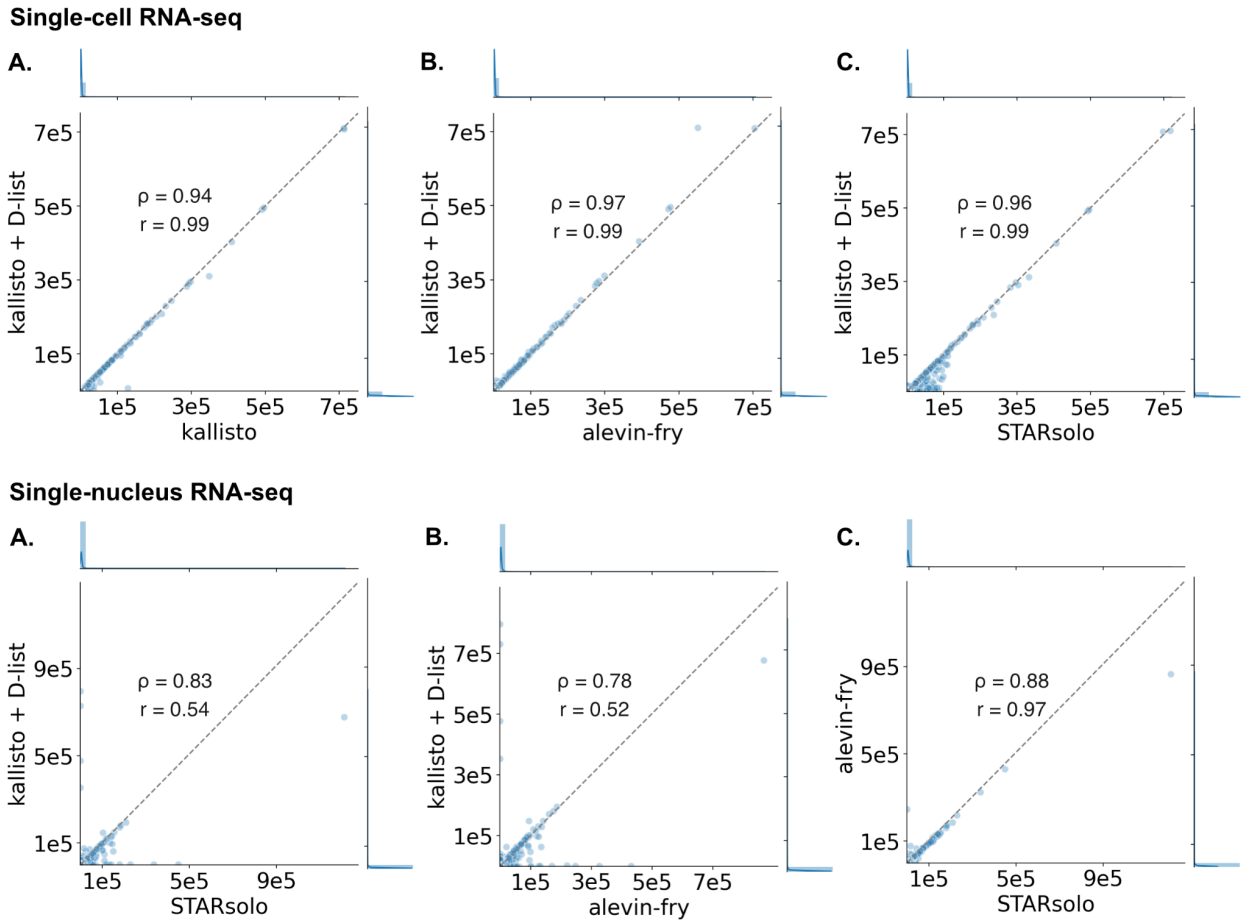

Figure 4. Comparison of kallisto with alevin-fry and STARsolo on 10x Genomics single-cell RNA-seq and single-nucleus data (Methods). **A.** Assessment of the effect of using the D-list with kallisto. **B.** Comparison of kallisto to alevin-fry on single-cell RNA-seq **C.** Comparison of kallisto to STARsolo on single-cell RNA-seq **D.** The difference between kallisto's quantification of nascent transcripts from single-nucleus RNA-seq, and alevin-fry's quantification of single-nucleus RNA-seq. **E.** The difference between kallisto's quantification of nascent transcripts from single-nucleus RNA-seq, andSTARsolo's quantification of single-nucleus RNA-seq. **F.** The similarity of alevin-fry and STARsolo's quantification of single-nucleus RNA-seq. In all plots Spearman correlation is shown with ρ and Pearson correlation with r.

More interesting, is the quantification of single-nucleus RNA-seq, for which nascent RNAs can be quantified by generating a D-list based on DFKs in mature transcripts. This biologically motivated approach to quantification of nascent RNAs, that counts only reads that are definitively nascent or ambiguous (but not mature mRNAs), is practical for biological data, and produces results that are significantly different from current approaches that quantify single-nucleus RNA-seq by agglomerating counts for nascent and mature transcripts together. This is reflected in a comparison of kallisto with the D-list to alevin-fry (Figure 4C) and to

STARsolo (Figure 4D). STARsolo and alevin-fry, which both quantify single-nucleus by mixing up mature and nascent transcripts, are more similar to each other (Figure 4E).

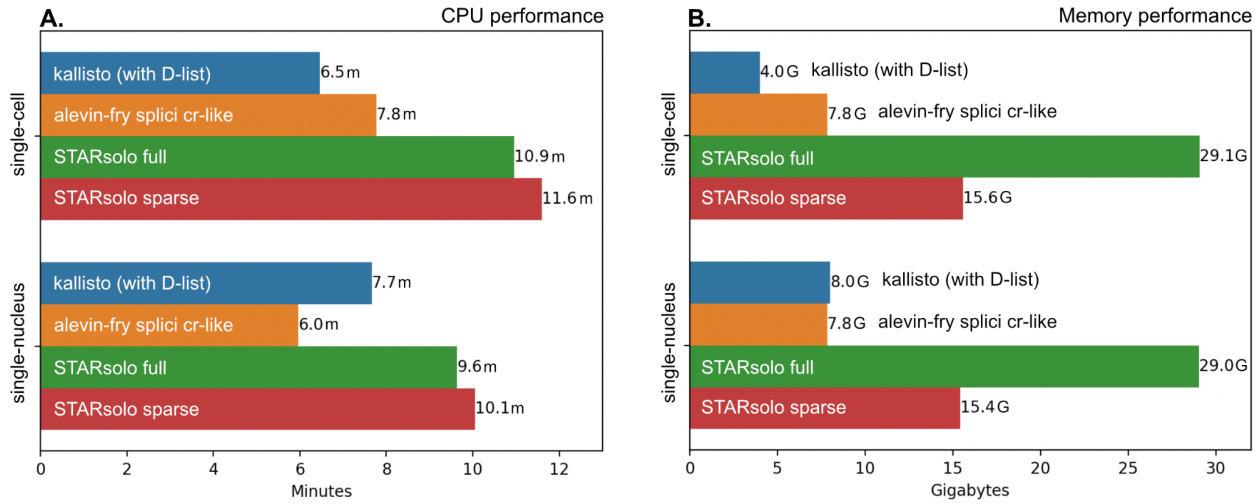

Figure 5. Running time and peak memory usage comparisons during quantification of 204,596,690 single-cell and single-nucleus RNA-seq reads from 7,377 mouse brain cells. All programs were allocated 20 threads for quantification (Methods).

Discussion

The use of single-nucleus RNA-seq in lieu of single-cell RNA-seq has been increasingly popular, primarily because nuclei don't need to be dissociated(Habib et al. 2017; Cervantes-Pérez et al. 2022; Liang et al., n.d.; Al-Dalahmah et al. 2020). Yet, despite the increasing preponderance of single-nucleus RNA-seq data and fundamental differences between it and single-cell RNA-seq data, the two data types have not been treated similarly for quantification purposes. While mature transcripts are quantified for single-cell RNA-seq, single-nucleus RNA-seq is quantified by combining quantifications of both nascent and mature transcripts. Our approach to quantification offers flexibility for teasing apart the counts of nascent vs. mature mRNAs from such data, and highlights the possibility for quantifying nascent transcripts from single-nucleus RNA-seq, just as mature transcripts are quantified from single-cell RNA-seq. Our method is based on an efficient and accurate algorithm, and is implemented in software that can form part of reproducible workflows that have modest hardware requirements. Furthermore, the generation of nascent and mature transcripts counts will prove invaluable for methods that rely on such counts for downstream analysis. Arguably, all single-nucleus and single-cell RNA-seq should first involve integrating the nascent and mature transcript modalities using a biophysical model of transcription (Gorin and Pachter 2022b), although investigation of that hypothesis is beyond the scope of this work.

There are several limitations to the quantification framework we have proposed. In a cell, the set of nascent mRNAs at any given time is likely to include partially processed molecules (Pai et al. 2018), and in principle the complete splicing cascade must be understood and known in order to accurately quantify single-nucleus or single-cell RNA-seq data (Gorin and Pachter 2022a).

Furthermore, the use of ambiguous reads both for single-cell and single-nucleus RNA-seq is unsatisfactory. Ideally reads should be longer so that they can be uniquely classified, or they should be fractionally classified according to probability estimates of the ratio between nascent and mature transcripts. The latter approach is non-trivial due to variation in effective transcript lengths that will depend on library preparation and must be accounted for (Pachter 2011) , but this is an interesting direction of study.

Finally, we believe the memory efficiencies introduced with the update to kallisto prepared for this work (Figure 5B), will greatly extend its utility for a variety of other applications, such as single-cell genomics assays that generate reads that must be aligned to the genome (Gao and Pachter 2021) and metagenomics (Schaeffer et al. 2017).

Methods

The D-list

A D-list (distinguishing list) enables accurate quantification of RNA-seq reads in experiments where reads that are not an expression of the target transcriptome may still contain sequences, which do occur in the target transcriptome. Without the D-list, these reads may be erroneously quantified as transcripts in the target transcriptome, based on alignment of the common sequences. Thus, the D-list may contain any sequences that are not desired in the abundance matrix yielded by the quantification. They may be the transcriptomes of other organisms, in case the RNA sample is contaminated, they may consist of the nascent (unspliced) versions of the mature (spliced) transcripts in the target transcriptome, in case only the quantification of mature transcripts, or only the quantification of nascent transcripts is desired, or they may contain common transposable elements, such as Alu regions, which might otherwise introduce undesired noise. The D-list is incorporated into the index by finding all sequences, *k* base-pairs or longer, that occur in both the D-list and the target transcriptome. The first *k*-mer upstream and the first *k*-mer downstream of each such common sequence in the D-list are added to the index colored *de Bruijn* graph (Figure 1.C, Figure 1.B). We refer to these new vertices in the graph as *distinguishing flanking kmers* (DFKs). The DFK vertices are left uncolored in the index, such that during quantification, reads that contain them will be masked out, and go unaligned. Note that there is no need for processing sequences that map to the DFK vertices in special cases downstream.

As an illustration, when mapping a Ⓜ read containing both 🅜 and 🅝|🅜 *k*-mers to an index built from Ⓝ transcripts, the 🅝|🅜 *k*-mers will be found in the index, whereas the disambiguating 🅜 *k*-mers will not. The whole read will be erroneously mapped based on the 🅝|🅜 ambiguous *k*-mers that are present in the index. By finding all 🅝|🅜 *k*-mers in the nascent versions of the transcripts, and adding any distinguishing flanking 🅜 *k*-mers to the index, the Ⓜ read will be masked from mapping to a Ⓝ transcript.

Recent papers have discussed various ways of reducing the number of false positives in RNA quantification. The simplest way of doing so is to align the RNA-seq reads against both the target transcriptome and a secondary transcriptome containing undesired transcripts. This has been explored in (Srivastava et al. 2020), and while it may yield fewer false positives than aligning against the target transcriptome alone, it is memory-intensive and, depending on the method, potentially CPU intensive. Another approach, also explored in (Srivastava et al. 2020) is to introduce a preprocessing step, wherein sequences that are similar to those in the target transcriptome are extracted, using some heuristic for similarity. These sequences are then added to the index, and handled in special cases downstream, during alignment and quantification. Most recently, alevin-fry introduced the splici index which reduces the number of false positives while controlling peak memory usage better than previous approaches. However, indexing intronic sequences, while enabling workflows like RNA velocity, still incurs a significantly larger memory cost than indexing just the target transcriptome.

Our method has the distinct advantage of incorporating only the minimum amount of data, required to disambiguate common sequences, into the index. Therefore, the memory usage and runtime of kallisto using a D-list are on par with the memory usage and runtime of kallisto, without the use of a D-list.

Validation

In addition to validating our results on the simulation framework developed by Kaminow, Yunusov, and Dobin (2021) we simulated two error-free experiments using reads generated by BBMap (Bushnell 2014)**.** One simulation represents a single-cell RNA-seq experiment consisting of 4,000,000 mature RNA reads and 1,000,000 nascent RNA reads. The other one represents a single-nucleus RNA-seq experiment consisting of 4,000,000 nascent RNA reads and 1,000,000 mature RNA reads. The respective ratios of mature to nascent reads was estimated based on (Gorin, Yoshida, and Pachter 2022). The mature transcripts were obtained from version 104 of the Ensembl of GRCh38, and the nascent transcripts were taken to be the entire sequence from the start of the first exon through the end of the last exon. The kallisto quantification of the single-cell RNA-seq simulation was performed using an index built from the mature transcripts and a D-list constructed from all the GRCh38 scaffolds. For the single-nucleus RNA-seq simulation an index built from the nascent transcripts, with a D-list constructed from the nascent transcripts was used.

The STARsolo simulation

We obtained the simulation framework developed by (Kaminow, Yunusov, and Dobin 2021; He et al. 2022) from https://github.com/dobinlab/STARsoloManuscript/ and ran it as-is, substituting the deprecated “decoy”-mode of alevin-fry with its “splici”-mode replacement.

Correction of the STARsolo calculation of Spearman correlation

(Kaminow, Yunusov, and Dobin 2021), calculate the correlation between quantification and the ground truth expression of a cell, only between the elements of the gene/cell count matrices which are expressed in either the simulation or tool quantification. Thus, genes for which there is no expression in either quantification or ground truth, i.e. true negatives, are ignored in the calculation of the Spearman correlation coefficient. This leads to an artificial inflation of the effect of false positives and false negatives on the coefficient. For example, consider a cell with very few genes expressed in the simulation, and as a result 0 counts for almost all genes. Now suppose a method reports only 7 genes with 1 count (out of thousands), and in the ground truth of the simulation there are also 7 genes with 1 count, with a disagreement on 2 genes, i.e. method = (1,1,1,1,0,1,1,1) and simulation = (1,1,1,0,1,1,1,1). (Kaminow, Yunusov, and Dobin 2021) compute the Spearman correlation between these vectors, which is -0.1428571, for a cell where the method and simulation agree over thousands of genes. Supplementary Table 1 shows the quantiles of the Spearman coefficient between the ground truth of the simulation developed by Kaminow, Yunusov, and Dobin (2021) and each of the quantification tool, calculated only from true positives, false positives, and false negatives. Supplementary table 2 shows the quantiles when the calculation attributes for true negatives. It is evident that when calculated correctly, there is negligible difference between kallisto, alevin-fry, and STARsolo with respect to the ground truth in this simulation.

Analysis of single-cell and single-nucleus RNA-seq

We quantified a 10xV3 dataset containing 204,596,690 single-cell, and single-nucleus RNA-seq reads from 7,377 adult mouse brain cells via 10x Genomics: https://www.10xgenomics.com/resources/datasets/5k-adult-mouse-brain-nuclei-isolated-with-chromium-nuclei-isolation-kit-3-1-standard, using kallisto, alevin-fry (splici em-like), and STARsolo. For quantification of the single-cell data, we constructed the kallisto index from version 108 of the Ensembl mouse annotation (Cunningham et al., 2022), using the entire mouse genome to construct the D-list. For quantification of the single-nucleus data, we constructed the kallisto index from nascent versions of the mouse transcripts (Validation), using the mature transcripts to construct the D-list. We ran STARsolo in "Gene"-mode for the single-cell data, and in "GeneFull"-mode for the single-nucleus data. All benchmarks were performed on a computer with two Intel(R) Xeon(R) Gold 6152 CPUs (a total of 44 cores), 768GiB of DDR4/2666MHz RAM, and twelve 12TB SATA hard drives. Programs were allocated 20 threads for quantifying the data, and CPU and peak memory usage were obtained via /usr/bin/time -v. Each program was run separately to minimize the likelihood of I/O bottlenecks.

Memory improvements to kallisto

The *de Bruijn* graph implementation in kallisto was replaced with Bifrost (Holley and Melsted 2020), which employs a minimizer lookup table in lieu of a *k*-mer lookup table in order to achieve a lower memory footprint. Furthermore, since the set of minimizers in the graph is known at the time of quantification, we replaced the minimizer hash function with BBHash (Limasset et al. 2017), which implements a minimal perfect hash function. This enables us to shrink the minimizer hash table to capacity, saving memory. Lastly, we implemented dynamic allocation of Equivalence Classes (ECs) during quantification. Thus, an EC is only created once it has been found to be used by a read, whereas the previous paradigm preemptively created the ECs used by all the vertices in the graph, during indexing, regardless of whether or not they were ever used by a read.

Code availability

kallisto is available under the BSD-2-Clause license. The version used for this paper is available at https://github.com/pachterlab/kallisto-D. All code for simulations and downstream analyses is available at https://github.com/pachterlab/HSHMP_2022

Software versions

alevin-fry 0.8.0
BBMap v35.85
bustools 0.41.0
kallisto 0.49
STAR 2.7.9a

Supplementary Information

	Per-cell Spearman coefficient quantiles		
Tool	**25%**	**50%**	**75%**
STARsolo	0.9893	0.9922	0.9948
kallisto (with D-list)	0.9678	0.9735	0.9789
alevin-fry splici cr-like	0.9665	0.9717	0.9763
alevin-fry splici cr-like-em	0.9208	0.9291	0.9384
alevin-fry cr-like	0.8604	0.8824	0.9157
alevin-fry cr-like-em	0.8294	0.8559	0.8966
kallisto	0.7594	0.7901	0.8381
alevin-fry sketch cr-like	0.7527	0.7830	0.8337
alevin-fry sketch cr-like-em	0.6932	0.7282	0.7891

Supplementary Table 1. STARsolo, kallisto, and alevin-fry compared in the simulated single-cell RNA-seq experiment by Kaminow et al. containing no multimapping reads. The Spearman coefficient of the quantified abundances and the ground truth per cell was calculated and the quantiles reported. As per Kaminow et al. the true negatives were left out of the abundance vectors used to calculate the Spearman coefficient, which artificially inflates the effect of false negatives and false positives on the coefficient.

	Per-cell Spearman coefficient quantiles		
Tool	**25%**	**50%**	**75%**
STARsolo	0.9961	0.9969	0.9977
kallisto (with D-list)	0.9885	0.9903	0.9919
alevin-fry splici cr-like	0.9909	0.9922	0.9933
alevin-fry splici cr-like-em	0.9740	0.9761	0.9782
alevin-fry cr-like	0.9468	0.9539	0.9635
alevin-fry cr-like-em	0.9427	0.9503	0.9610
kallisto	0.8792	0.8939	0.9139
alevin-fry sketch cr-like	0.8740	0.8893	0.9104
alevin-fry sketch cr-like-em	0.8564	0.8737	0.8988

Supplementary Table 2. STARsolo, kallisto, and alevin-fry compared in the same simulated single-cell RNA-seq experiment as in Supplementary table 1. The Spearman coefficient of the quantified abundances and the ground truth per cell was calculated and the quantiles reported. Unlike in (Kaminow, Yunusov, and Dobin 2021), the true negatives were accounted for in the abundance vectors used to calculate the Spearman coefficient, which yields a meaningful measure of correlation.

References

Al-Dalahmah, Osama, Alexander A. Sosunov, A. Shaik, Kenneth Ofori, Yang Liu, Jean Paul Vonsattel, Istvan Adorjan, Vilas Menon, and James E. Goldman. 2020. “Single-Nucleus RNA-Seq Identifies Huntington Disease Astrocyte States.” *Acta Neuropathologica Communications* 8 (1): 19.

Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. “Near-Optimal Probabilistic RNA-Seq Quantification.” *Nature Biotechnology* 34 (5): 525–27.

Bushnell, Brian. 2014. “BBMap: A Fast, Accurate, Splice-Aware Aligner.” LBNL-7065E. Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States). https://www.osti.gov/biblio/1241166.

Cervantes-Pérez, Sergio Alan, Sandra Thibivillliers, Sutton Tennant, and Marc Libault. 2022. “Review: Challenges and Perspectives in Applying Single Nuclei RNA-Seq Technology in Plant Biology.” *Plant Science: An International Journal of Experimental Plant Biology* 325 (111486): 111486.

Cunningham, Fiona, James E. Allen, Jamie Allen, Jorge Alvarez-Jarreta, M. Ridwan Amode, Irina M. Armean, Olanrewaju Austine-Orimoloye, et al. 2022. “Ensembl 2022.” *Nucleic Acids Research* 50 (D1): D988–95.

Gao, Fan, and Lior Pachter. 2021. “Efficient Pre-Processing of Single-Cell ATAC-Seq Data.” *bioRxiv*. https://doi.org/10.1101/2021.12.08.471788.

Gorin, Gennady, Meichen Fang, Tara Chari, and Lior Pachter. 2022. “RNA Velocity Unraveled.” *PLoS Computational Biology* 18 (9): e1010492.

Gorin, Gennady, and Lior Pachter. 2022a. “Modeling Bursty Transcription and Splicing with the Chemical Master Equation.” *Biophysical Journal* 121 (6): 1056–69.

———. 2022b. “Monod: Mechanistic Analysis of Single-Cell RNA Sequencing Count Data.” *bioRxiv*. https://doi.org/10.1101/2022.06.11.495771.

Gorin, Gennady, Shawn Yoshida, and Lior Pachter. 2022. “Transient and Delay Chemical Master Equations.” *bioRxiv*. https://doi.org/10.1101/2022.10.17.512599.

Habib, Naomi, Inbal Avraham-Davidi, Anindita Basu, Tyler Burks, Karthik Shekhar, Matan Hofree, Sourav R. Choudhury, et al. 2017. “Massively Parallel Single-Nucleus RNA-Seq with DroNc-Seq.” *Nature Methods* 14 (10): 955–58.

He, Dongze, Mohsen Zakeri, Hirak Sarkar, Charlotte Soneson, Avi Srivastava, and Rob Patro. 2022. “Alevin-Fry Unlocks Rapid, Accurate and Memory-Frugal Quantification of Single-Cell RNA-Seq Data.” *Nature Methods* 19 (3): 316–22.

Holley, Guillaume, and Páll Melsted. 2020. “Bifrost: Highly Parallel Construction and Indexing of Colored and Compacted de Bruijn Graphs.” *Genome Biology* 21 (1): 249.

Kaminow, Benjamin, Dinar Yunusov, and Alexander Dobin. 2021. “STARsolo: Accurate, Fast and Versatile Mapping/quantification of Single-Cell and Single-Nucleus RNA-Seq Data.” *bioRxiv*. https://doi.org/10.1101/2021.05.05.442755.

Kuo, A., K. D. Hansen, and S. C. Hicks. 2022. “Quantification and Statistical Modeling of Chromium-Based Single-Nucleus RNA-Sequencing Data.” *bioRxiv*. https://www.biorxiv.org/content/10.1101/2022.05.20.492835.abstract.

La Manno, Gioele, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, et al. 2018. “RNA Velocity of Single Cells.” *Nature* 560 (7719): 494–98.

Liang, Qingnan, Rachayata Dharmat, Leah Owen, Akbar Shakoor, Yumei Li, Sangbae Kim, Albert Vitale, et al. n.d. “Single-Nuclei RNA-Seq on Human Retinal Tissue Provides Improved Transcriptome Profiling.” https://doi.org/10.1101/468207.

Limasset, Antoine, Guillaume Rizk, Rayan Chikhi, and Pierre Peterlongo. 2017. “Fast and

Scalable Minimal Perfect Hashing for Massive Key Sets." *arXiv [cs.DS]*. arXiv. http://arxiv.org/abs/1702.03154.

Melsted, Páll, A. Sina Booeshaghi, Lauren Liu, Fan Gao, Lambda Lu, Kyung Hoi Joseph Min, Eduardo da Veiga Beltrame, Kristján Eldjárn Hjörleifsson, Jase Gehring, and Lior Pachter. 2021. "Modular, Efficient and Constant-Memory Single-Cell RNA-Seq Preprocessing." *Nature Biotechnology* 39 (7): 813–18.

Pachter, Lior. 2011. "Models for Transcript Quantification from RNA-Seq." *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/1104.3889.

Pai, Athma A., Joseph M. Paggi, Paul Yan, Karen Adelman, and Christopher B. Burge. 2018. "Numerous Recursive Sites Contribute to Accuracy of Splicing in Long Introns in Flies." *PLOS Genetics*. https://doi.org/10.1371/journal.pgen.1007588.

Schaeffer, L., H. Pimentel, N. Bray, P. Melsted, and L. Pachter. 2017. "Pseudoalignment for Metagenomic Read Assignment." *Bioinformatics* 33 (14): 2082–88.

Sina Booeshaghi, A., and Lior Pachter. 2021. "Benchmarking of Lightweight-Mapping Based Single-Cell RNA-Seq Pre-Processing." *bioRxiv*. https://doi.org/10.1101/2021.01.25.428188.

Soneson, Charlotte, Avi Srivastava, Rob Patro, and Michael B. Stadler. 2021. "Preprocessing Choices Affect RNA Velocity Results for Droplet scRNA-Seq Data." *PLoS Computational Biology* 17 (1): e1008585.

Srivastava, Avi, Laraib Malik, Hirak Sarkar, Mohsen Zakeri, Fatemeh Almodaresi, Charlotte Soneson, Michael I. Love, Carl Kingsford, and Rob Patro. 2020. "Alignment and Mapping Methodology Influence Transcript Abundance Estimation." *Genome Biology* 21 (1): 239.

Zeng, H. 2022. "What Is a Cell Type and How to Define It?" *Cell* 185 (15): 2739–55.