TECHNICAL NOTE

Interpreting Intronic and Antisense Reads in 10x Genomics Single Cell Gene Expression Data

Introduction

Chromium Single Cell Gene Expression products enable the profiling of tens of thousands of cells, enhancing the understanding of complex biological systems and diseases. Analysis of whole transcriptome data derived from Chromium Single Cell 3' and 5' Gene Expression libraries provides key cellular insights and interpretation of these data may be impacted if intronic and antisense reads are included in the analysis. This Technical Note discusses potential mechanisms that can result in intronic and/or antisense read generation in Chromium Single Cell Gene Expression assays. The document also analyzes data supporting each mechanism and elaborates on the overall impact of intronic and antisense reads on gene expression data analysis and interpretation.

Chromium Single Cell Gene Expression data primarily includes reads mapped to exonic regions derived from mature spliced transcripts. Additionally, reads that map entirely or partially to intronic regions (Figure 1) and strand-specific antisense reads have also been observed by researchers in gene expression data (Mereu et al., Ding et al.). Priming poly-A tracts on RNA or first-strand cDNA has been suggested as a possible mechanism for intronic and antisense reads in Chromium Single Cell Gene Expression data (Ding et al., La Manno et al.). Nuclei contain a large fraction of unprocessed pre-mRNA molecules that include introns and therefore, it is common to count intronic UMIs to increase assay sensitivity (Peng et al., Kreimann et al.).

To understand the underlying reasons for the presence of intronic and antisense reads in Chromium Single Cell Gene Expression data, five potential mechanisms were tested (Figures 2A-2E). Data analyzed (Table 1) to test each mechanism are presented and discussed in the Results section of this Technical Note (Figures 3-8).

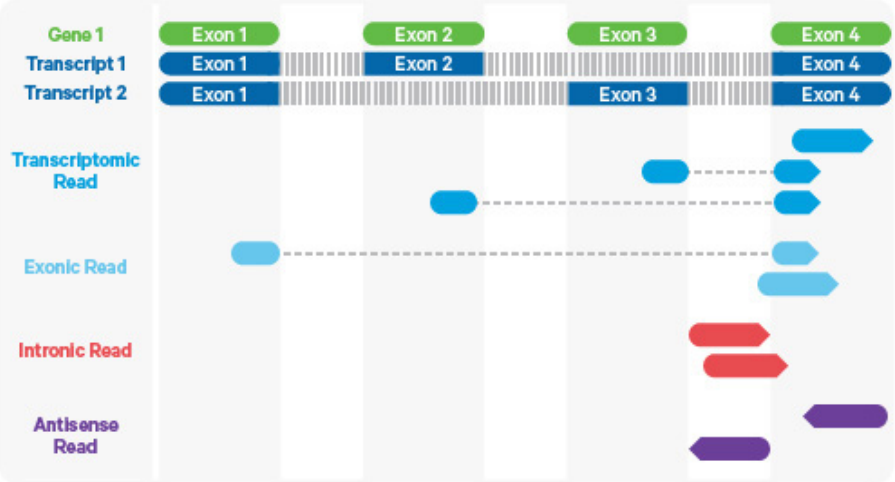

Figure 1. Gene organization and transcription. Genes include both exonic and intronic segments. When a pre-mRNA is processed, introns and possibly some exons are removed, which results in many possible processed transcripts. Reads map to these transcripts in a strand specific manner (for Single Cell 3' Gene Expression, sense reads face the end of the gene where a poly-A tail is appended to the transcript). Single cell gene expression data primarily includes sense reads that map to exons but may also include intronic reads and reads that map in an antisense orientation. By default, Cell Ranger only counts reads that are confidently mapped to an annotated transcript, in the sense orientation (transcriptomic reads).

Proposed Mechanisms

Proposed mechanisms for presence of intronic and antisense reads in gene expression data

1. Internal poly-A priming in Chromium Single Cell 3' assays

The Single Cell 3' Gel Bead poly(dT) primer (includes 10x Barcode and UMI) enables the production of barcoded, full-length cDNA by priming off the poly-adenylated (poly-A) mRNA tail. This poly(dT) primer can also potentially prime an internal poly-A stretch in the mRNA instead of the poly-A tail, resulting in a sense read pair located internally rather than at the end of the transcript (Figure 2A). This mechanism could result in reads occurring on exons or introns, although there are ~21.2 times as many poly-A stretches in introns compared to exons in the human genome and they occur at a rate ~1.7 times as large in introns compared to exons (counted as stranded, non-overlapping poly-A 7-mers).

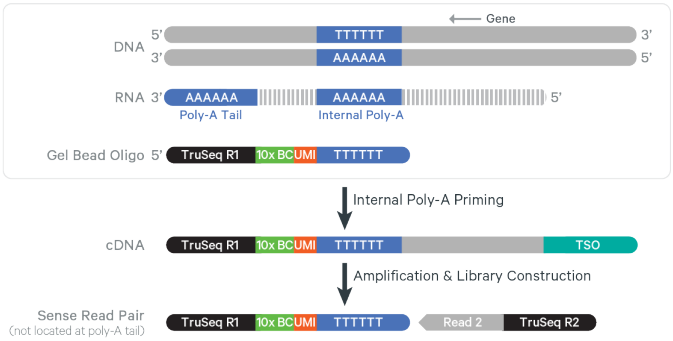

Figure 2A. Internal poly-A priming mechanism.

2. TSO priming in Chromium Single Cell 3' and 5' assays

Antisense reads can be potentially generated by aberrant TSO priming at homologous sites on the RNA. Such priming by the Single Cell 5' Gel Bead oligo that includes TSO, 10x Barcode, and UMI can result in amplifiable cDNA capable of producing both sense and antisense reads. In Single Cell 3' assays, an intermolecular template switch event from a poly-T stretch on the RNA to a poly-T stretch in the gel bead primer with 10x Barcode and UMI can result in an amplifiable antisense molecule (Figure 2B). There are ~25.1 times as many poly-T stretches in introns compared to exons in the human genome and they occur at a rate ~2.0 times as large in introns compared to exons (counted as stranded, non-overlapping poly-T 7-mers).

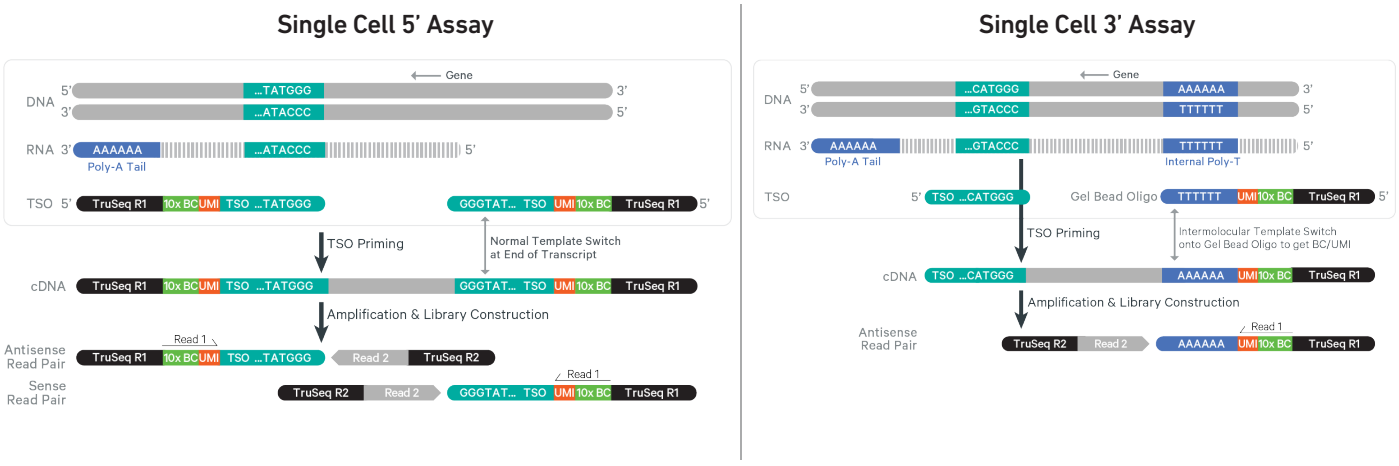

Figure 2B. TSO priming mechanism.

Contd.

Proposed Mechanisms

3. Poly(dT) primer strand invasion in Chromium Single Cell 3' assays

In Chromium Single Cell 3' assays, first-strand cDNA synthesis may be initiated normally or at an internal poly-A site but be interrupted by an intermolecular template switch from a poly-T stretch on the mRNA to the poly-T stretch on a Gel Bead oligo. As a result, an amplifiable cDNA with distinct UMIs on each end may be formed capable of generating both sense and antisense reads. If this occurs simultaneously with internal priming, multiple sense and antisense UMIs could be generated from distinct poly-A and poly-T sites on the same molecule (Figure 2C).

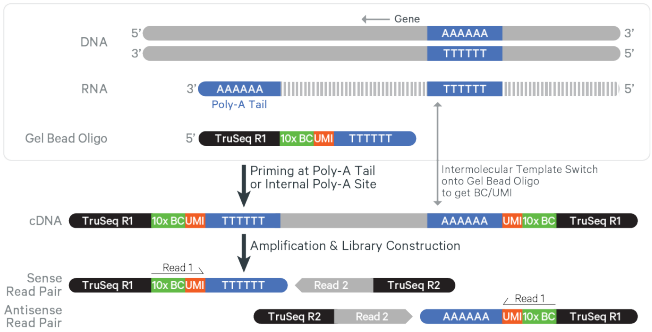

Figure 2C. Poly (dT) strand invasion mechanism.

4. First-strand cDNA priming in Chromium Single Cell 3' and 5' assays

After reverse transcription, RNA may be degraded or compromised, leaving the first-strand-cDNA exposed. The poly(dT) primer priming off a poly-A stretch on the cDNA can result in a cDNA with a sense 10x Barcode and UMI on one end and an antisense 10x Barcode and UMI on the other end. Since the barcode and UMIs on both ends include PCR primers, the construct can be amplified, fragmented, and sequenced and the antisense side of the construct can result in antisense read pairs (Figure 2D).

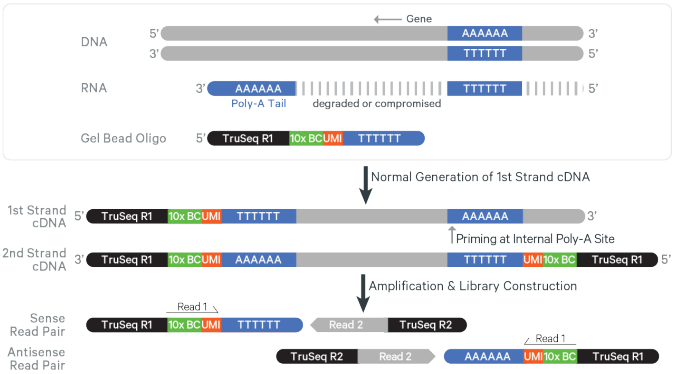

Figure 2D. First strand cDNA mechanism.

5. Sense-antisense fusion in Chromium Single Cell 3' and 5' assays

Mechanisms that produce sense-antisense fusion cDNAs have been proposed as a source of antisense artifacts (Perocchi et al., Houseley et al.). After first strand cDNA synthesis, if the RNA is degraded or the RNA-cDNA duplex is compromised, the cDNA can potentially form a hair-pin loop with itself at a short reverse-complementary motif and self-prime to create a hairpin cDNA molecule that is a fusion between sense (Figure 2E-black) and antisense (Figure 2E-red) sequence. At the initiation of PCR, the fusion molecule can denature and the second strand cDNA can be completed. This molecule can be amplified and fragmented generating an antisense read pair (Figure 2E).

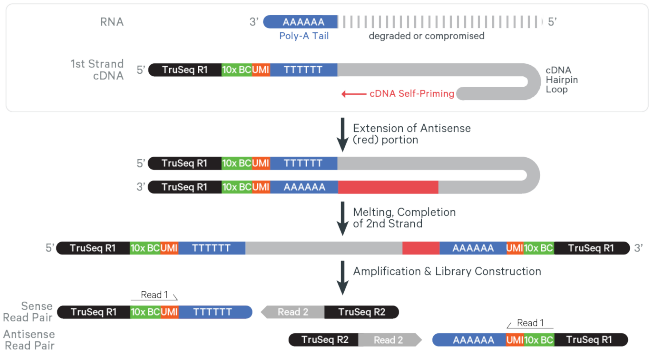

Figure 2E. Sense-antisense fusion mechanism.

Methods

Chromium Single Cell 3' Gene Expression (v3.1, v3.1 dual index) and 5' Gene Expression (v1.1, v2.0) libraries were generated from single cell or nuclei suspensions derived from various sample types and sequenced as described in the respective user guides.

Single Cell 3' Gene Expression libraries were also generated from nuclei using the Chromium Single Cell Multiome ATAC + Gene Expression workflow, which has the same gene expression assay scheme as Single Cell 3' Gene Expression v3.1 (dual index). In this document, the libraries generated using either of these two protocols or the other Single Cell 3' protocols, are collectively referred to as Single Cell 3' Gene Expression libraries derived using the Single Cell 3' assay.

The data was processed with Cell Ranger/Cell Ranger ARC and the resulting BAM files were analyzed with custom analysis scripts to investigate properties of intronic and antisense UMIs. See Table 1 for more information regarding the datasets analyzed.

Results

Analysis of single cell or single nuclei gene expression data derived from various sample types showed the presence of intronic and antisense UMIs as summarized in Table 1.

Overall, a large fraction of UMIs in the gene expression data mapped to intronic regions even though significant variability in the intronic fraction was observed between multiple datasets. Compared to single cells, single nuclei data had a larger fraction of intronic UMIs relative to sense exonic UMIs for samples with the same cell type and cell/nuclei load. However, most of the difference in intronic UMI fraction was the result of antisense intronic UMIs. Gene expression data derived using the Single Cell 5' Gene Expression assay showed a lower fraction of antisense and intronic UMIs compared to Single Cell 3' Gene Expression data, likely due to different mechanisms of antisense and intronic read generation.

Figures 3-8 present data supporting each of the potential mechanisms or combination of mechanisms that can generate intronic and/or antisense reads in Chromium Single Cell Gene Expression assays.

Genomic DNA priming

The mechanisms in Figure 2 propose that intronic and antisense reads are of transcript-based origin, either directly or indirectly. However, it is also possible that intronic and antisense reads could result from barcoded oligos priming on genomic DNA. The mechanism for this would be similar to the poly-A priming mechanism for intronic reads (Figure 2A), except that the poly-A tract would occur on either strand of genomic DNA rather than mRNA. Any reads generated in this fashion would introduce noise if they were included in gene expression analysis. Genomic DNA priming would not be expected to favor gene regions especially strongly, but the median rate of reads mapping to intergenic regions was only 3.8% across all datasets studied compared to 25.2% for intronic reads. Moreover, Figure 3 suggests a log-linear relationship between intronic versus exonic UMIs (Figure 3A) and antisense versus sense UMIs (Figure 3B) in the gene expression data, which indicates that intronic and antisense UMIs are observed as a function of transcript abundance. Based on this evidence, it appears that genomic DNA priming based mechanisms are not significant contributors to the generation of intronic and antisense reads.

Internal poly-A priming

A high fraction of sense intronic UMIs (~50-70%) with a downstream poly-A tract was observed in Single Cell 3' Gene Expression data across the studied datasets, compared to less than 30% for the Single Cell 5' Assay (Figure 4A). The position of the observed poly-A tracts is represented by a sharp peak at ~200 bp in the plots shown in Figure 4B for an example 3' Gene Expression dataset. This data implies internal poly-A priming (Figure 2A) as a potential mechanism for generation of intronic reads for 3' Gene Expression.

TSO priming

The fraction of gene-aligned UMIs in the indicated regions (intron/exon/sense/antisense) having TSO sequence at the beginning of the read is shown in data derived from single cells and nuclei for both Single Cell 3' and 5' Gene Expression assays in Figure 5. In the Single Cell 3' Gene Expression data, the antisense reads have a slightly increased occurrence of TSO sequence compared to sense reads, but overall there is significant variation in TSO presence amongst the datasets. Nuclei datasets had a similar median rate of TSO sequence to cell datasets but greater variance, possibly a result of higher RNA fragmentation in some nuclei experiments.

The proposed mechanism for antisense read generation in the Single Cell 5' assay is TSO priming, which results in a molecule with one TSO and one barcode/UMI on each side as shown in Figure 2B. Any such molecule short enough to

Table 1. Chromium Single Cell Gene Expression data derived from various sample types using the specified 10x Genomics assays.
Each sample along with corresponding cell/nuclei numbers and fraction genic UMI distribution is listed below. The denominator of each fraction is the total UMIs that align to gene regions, sense, or antisense.

	10x Genomics Assays	Cells/ Nuclei	Sample Type	# Cells/ Nuclei Recovered (Targeted)	Fraction UMIs				
					Sense Exonic	Sense Intronic	Antisense Exonic	Antisense Intronic	Mixed* (sense/ antisense)
Single Cell 3' Gene Expression Data	Single Cell Multiome ATAC + Gene Expression	Nuclei	Mouse 3T3 + Human GM12878 (1:1)	5977 (5000)	0.704	0.206	0.026	0.063	0.002
			Human PBMC	2716 (2000)	0.508	0.258	0.040	0.183	0.011
			Mouse E18	4881 (5000)	0.405	0.365	0.041	0.178	0.010
	Single Cell 3' Gene Expression v3.1	Cells	Mouse 3T3 + Human 293T (1:1)	1233 (1000)	0.705	0.223	0.014	0.057	0.001
			Mouse 3T3 + Human 293T (1:1)	6078 (5000)	0.696	0.222	0.016	0.066	0.001
			Mouse 3T3 + Human 293T (1:1)	11512 (10000)	0.693	0.215	0.018	0.073	0.001
			Human PBMC	5184 (5000)	0.575	0.275	0.020	0.123	0.006
			Mouse E18 Brain	6516 (5000)	0.622	0.241	0.027	0.106	0.004
	Single Cell 3' Gene Expression v3.1 (Dual Index)	Cells	Mouse 3T3 + Human 293T (1:1)	1317 (1000)	0.791	0.151	0.012	0.046	0.000
			HumanJurkat + Human Raji (1:1)	1121 (1000)	0.707	0.214	0.014	0.064	0.000
			Human 293T	1634 (1000)	0.727	0.198	0.013	0.062	0.000
			Human Breast Cancer	9432 (10000)	0.739	0.178	0.017	0.066	0.000
			Human Jurkat	999 (1000)	0.499	0.342	0.022	0.137	0.000
			Human PBMCs (dataset analysis in Fig. 4)	1041 (1000)	0.548	0.309	0.021	0.117	0.004
			Human PBMCs	6944 (5000)	0.606	0.273	0.019	0.099	0.002
			Human Raji	1174 (1000)	0.638	0.244	0.017	0.100	0.000
			Mouse 3T3	1383 (1000)	0.863	0.098	0.011	0.028	0.000
			Mouse E18 Brain	1349 (1000)	0.545	0.310	0.026	0.118	0.001
		Nuclei	Mouse 3T3 + Human 293T (1:1)	1412 (1000)	0.705	0.181	0.021	0.092	0.001
			Human Jurkat + Human Raji (1:1)	1153 (1000)	0.595	0.234	0.027	0.143	0.002
			Human 293T	1515 (1000)	0.576	0.261	0.025	0.137	0.001
			Human Breast Cancer	11411 (10000)	0.640	0.216	0.025	0.118	0.001
			Human Jurkat	1448 (1000)	0.489	0.275	0.030	0.204	0.002
			Human PBMCs	958 (1000)	0.396	0.472	0.021	0.108	0.004
			Human PBMCs	4900 (5000)	0.334	0.576	0.016	0.072	0.003
			Human Raji	1072 (1000)	0.522	0.252	0.029	0.194	0.002
			Mouse 3T3	1438 (1000)	0.815	0.122	0.017	0.045	0.000
			Mouse E18 Brain (dataset analysis in Figs. 3 & 7)	1624 (1000)	0.316	0.322	0.046	0.314	0.002
Single Cell 5' Gene Expression Data	Single Cell 5' Gene Expression v1.1	Cells	Mouse 3T3 + Human 293T (1:1)	8778 (10000)	0.914	0.041	0.025	0.020	0.000
			Human B cells	1334 (1000)	0.870	0.089	0.015	0.024	0.002
			Human Lung	6013 (10000)	0.846	0.073	0.044	0.036	0.002
			Human PBMC	4250 (5000)	0.858	0.084	0.024	0.030	0.005
			Human T cells	1168 (1000)	0.898	0.059	0.017	0.023	0.002
			Mouse Splenocytes	10963 (10000)	0.883	0.074	0.016	0.027	0.000
	Single Cell 5' Gene Expression v2.0	Cells	Human B cells	1148 (1000)	0.824	0.104	0.034	0.037	0.001
			Human Lung	6541 (10000)	0.826	0.081	0.054	0.038	0.001
			Human Melanoma	865 (1000)	0.852	0.093	0.024	0.030	0.001
			Human Melanoma	932 (1000)	0.795	0.117	0.040	0.046	0.002
			Human PBMC	4024 (5000)	0.811	0.090	0.057	0.039	0.002
			Human T cells	1142 (1000)	0.869	0.065	0.033	0.032	0.001
			Mouse Splenocytes	8363 (10000)	0.816	0.083	0.064	0.036	0.001

*UMIs that have at least one sense read and at least one antisense read

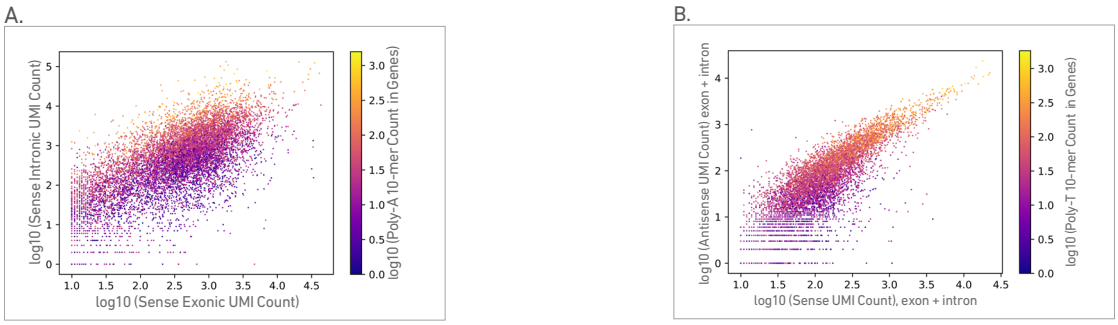

Figure 3. Relationship between sense, intronic, and antisense UMI counts for genes and poly-A tract frequency in a sample of 1,624 mouse brain nuclei (Table 1 red highlight) processed using the Chromium Single Cell 3' v3.1 – Dual Index assay. A. Relationship between sense exonic UMI counts and sense intronic UMI counts, colored by gene poly-A tract count. **B.** Relationship between sense UMI counts and antisense UMI counts, colored by gene poly-T tract count (poly-A tract according to read orientation).

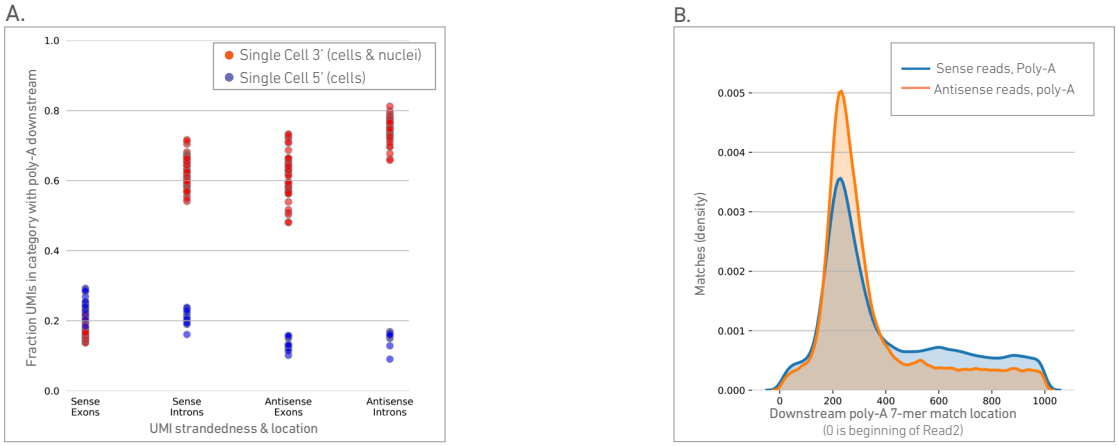

Figure 4. Frequency and position of poly-A tracts in gene expression data. A. Frequency of poly-A tracts (7 nucleotides or longer) on the reference genome within insert-size range (100-500 bp from read start) of gene-aligned UMIs is shown. Each point represents a different dataset and measures the fraction of UMIs having poly-A downstream as a fraction of UMIs with a given strandedness (sense, antisense) and location (exon, intron). UMIs with both sense and antisense annotations are not shown. **B.** Reads from a PBMC dataset produced using Single Cell Gene Expression v3.1 – Dual Index (Table 1 cyan highlight) were subsampled at a rate of 0.01 and all match positions of downstream poly-A 7mers were identified for sense and antisense reads separately. The density plots show peaks of poly-A occurrence in a similar position for both sense and antisense reads

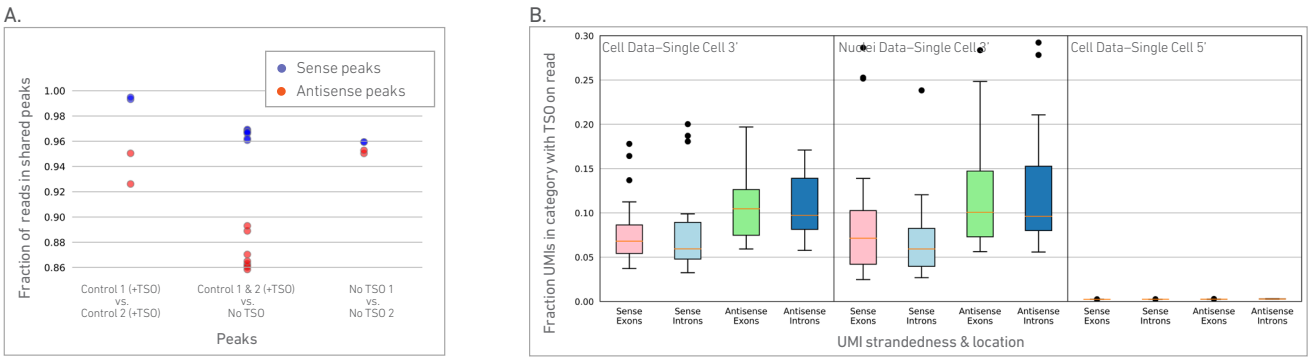

Figure 5. Impact of TSO on gene expression data. A. Concordance of antisense peaks in absence or presence of TSO. Data derived from two Single Cell 3' Gene Expression v3.0 experiments, performed with a modified protocol without TSO (No TSO), were compared to two control experiments with TSO added (~2,500 human Jurkat cells/experiment). Sense and antisense read peaks (high read density regions) were called for each sample using a read-coverage-based Hidden Markov Model and the peak set similarities between indicated pairs were determined. "Shared" peaks were defined as peaks where ≥50% of either peak is covered. The similarity of the peak sets was measured as the fraction of reads assigned to shared peaks (excluding singleton reads) as different peaks may contain vastly different read numbers. **B. TSO sequence present on reads derived from cells/nuclei processed using either Single Cell 3' or 5' assay.** The fraction of gene-aligned UMIs containing TSO sequence in different regions/orientations was calculated per-dataset as a fraction of gene-aligned UMIs in that region/orientation. The presence of TSO sequence was defined as the TSO sequence matching at the beginning of a read with an edit distance of ≤2. For 3' data, only 20 bp of TSO were required to match (first and last 5 bp were excluded), and for 5' data, the entire 13 bp TSO sequence was required to match.

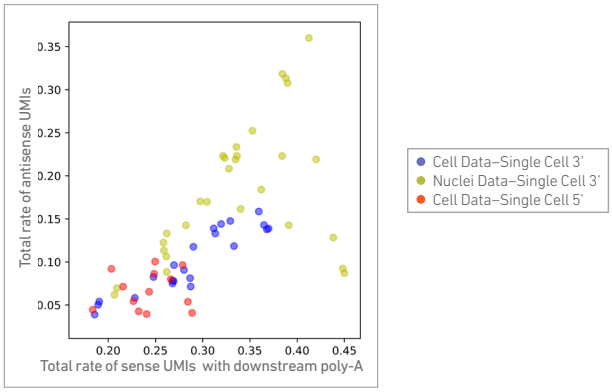

Figure 6. Relationship of sense UMIs with downstream poly-A and the occurrence of antisense UMIs in Single Cell 3' and 5' Gene Expression data (cells or nuclei). Fraction of all UMIs that are antisense as a function of the fraction of sense UMIs having downstream poly-A.

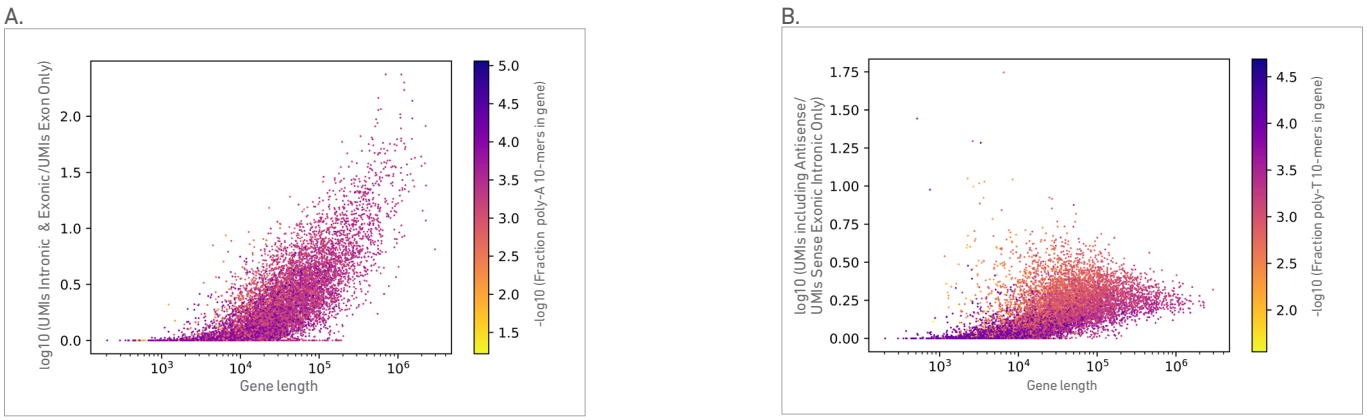

Figure 7. Gene length and poly-A/T biasof UMI gain in reads when analysis included intronic and antisense UMIs. A. Mouse brain nuclei sample (1,624 nuclei, Table 1 red highlight) was processed using the Chromium Single Cell 3' v3.1 (Dual Index) assay. The data were analyzed with introns included (sense exons+introns counted) and in normal sense-exon-only mode. The y-axis is the log10 fold change of UMIs when introns were included versus normal mode, as a function of gene length. Each gene is colored according to the fraction of 10-mers in the gene that are poly-A tracts. **B.** The same sample was processed including sense as well as antisense UMIs (exons+introns counted). The y-axis is the log10 fold change of UMIs when antisense UMIs were included versus sense exons+introns only, as a function of gene length. Each gene is colored according to the fraction of 10-mers in the gene that are poly-T tracts. In both figures **A** and **B**, genes with less than 10 total exonic UMIs are excluded and genes with no poly-A (or poly-T) were assigned a frequency of 10^{-4} for visualization.

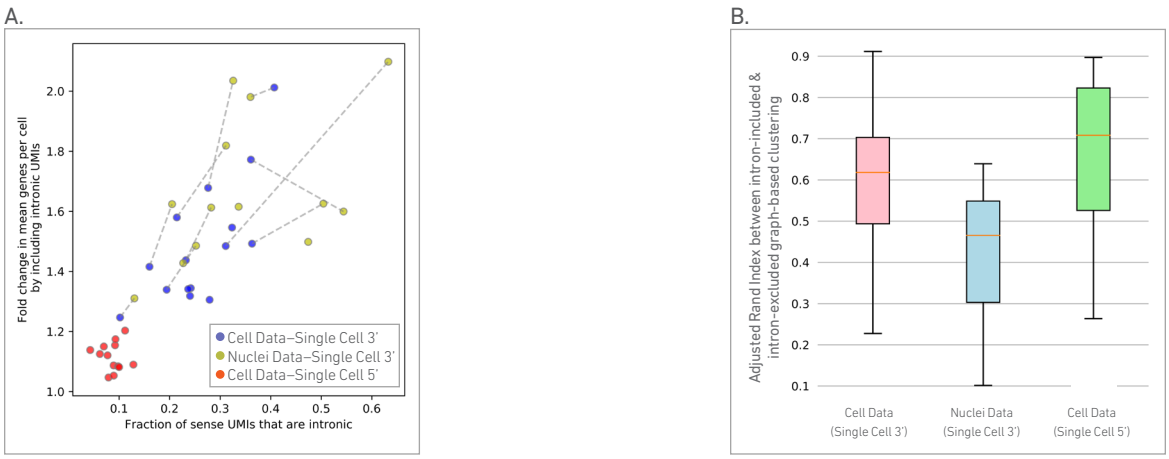

Figure 8. Increase in detected genes-per-cell by including intronic UMIs and similarity of graph-based clustering results including and excluding intronic UMIs. A. The fold change in mean genes-per-cell per indicated dataset by including (sense) intronic UMIs in the analysis as a function of the fraction of sense UMIs that are intronic. Dotted lines connect paired cell-nuclei experiments performed using the same assay, cell load, and sample type. **B.** The similarity between graph-based clustering results per dataset with and without (sense) intronic UMIs included in the analysis measured using the adjusted rand index (ARI). The maximum value is an ARI of 1.0, which means that intron-included analysis and intron-excluded analysis returned the exact same barcode clusters.

be sequenced without fragmentation could manifest in the data as an antisense UMI with the TSO sequence from the sense side of the cDNA in the read. This is not observed in Figure 5B. In Single Cell 5' Gene Expression data, almost no antisense UMIs with TSO sequence at the beginning of a read were observed. However, it is possible that none of the fragments are short enough for the TSO to be observed.

To test the effect of TSO on the location of antisense reads, a modified protocol that excludes TSO was compared to control experiments that use TSO. It is important to note that this protocol was performed only for the purpose of understanding antisense read generation, and resulted in a 57% reduction in median genes per cell, an 81% reduction in median UMIs per cell compared to the control, as well as a ~4.4 fold increase in antisense reads. It was found that presence or absence of TSO does not significantly affect the location of antisense peaks (dense clusters of reads) in Single Cell 3' data as shown in Figure 5A. When TSO is removed using the modified protocol, the positions of antisense read peaks highly correlate (~85%, weighted by number of reads) with control experiments containing TSO, implying that TSO priming is not a significant contributor to antisense reads in this assay.

Poly(dT) primer strand invasion and first-strand cDNA priming

Data presented in Figure 4A demonstrates that ~65-80% of antisense intronic reads produced by Single Cell 3' Gene Expression have a downstream poly-A tract (which is a poly-T tract according to the gene's orientation), compared to less than 20% for 5' Gene Expression. Similar to sense intronic reads, the antisense reads in 3' Gene Expression have a peak of poly-A occurrence ~200 bp downstream (Figure 4B). Higher prevalence of poly-A tracts is related to higher levels of antisense UMIs for genes across a wide range of sense UMI expression levels (Figure 3B). These observations are consistent with either a mechanism of poly(dT) primer strand invasion (Figure 2C) or a mechanism of first strand cDNA priming (Figure 2D). Both mechanisms would be evidenced by poly-A sites downstream of antisense reads and could produce similar cDNA molecules with amplifiable sense and antisense 10x Barcodes and UMIs. The sense barcode on these molecules itself may be internally primed. Indeed, the fraction of sense UMIs that have a downstream poly-A, potentially internally primed by poly(dT), and the total rate of antisense UMIs were correlated in the Single Cell 3' Gene Expression data, with higher fractions of both in nuclei (Figure 6).

Sense-antisense fusion

Sense-antisense fusion events may be observed as clipped portions of reads that match upstream reverse-complemented sequence. For this note, these events were quantified by reverse-complementing clipped ends of reads and aligning them to the 1,000 bp region upstream of the read, requiring a 90% sequence match. For most datasets, the events were detected on 0.5% or less of gene-aligned UMIs (data not shown). However, sense-antisense fusion sequences were only detected within a 21 bp window on the read. Clipped fusion sequences were required to be at least 10 bp (to reduce false positives) and less than 32 bp (due to alignment cutoffs in STAR aligner). In particular these numbers do not include sense-antisense fusions occurring past the insert on the cDNA template, which are challenging to detect. With the available data, there is limited evidence to support sense-antisense fusions as a primary mechanism for antisense read generation.

Summary of key findings

Based on these results, the high occurrence of poly-A tracts observed downstream of sense intronic reads can be clearly attributed to priming on an internal poly-A on the mRNA (Figure 2A). However, for antisense reads, the high occurrence of downstream poly-A tracts is not sufficient to understand the underlying mechanisms. Poly(dT) priming strand invasion (Figure 2C) and first strand cDNA poly-A priming (Figure 2D) are both potential mechanisms. Whether the poly-A tracts are being strand invaded by poly(dT) primers during first-strand cDNA synthesis or are serving as priming sites on the synthesized first-strand cDNA is not completely understood. It is possible that the poly-A tracts are priming sites and endogenous RNAse H action is degrading the RNA in the RNA-cDNA hybrid molecules to expose the cDNA strand. No such cDNA strand exposure would be needed if the strand is being invaded by poly(dT) primers. TSO priming followed by strand invasion (Figure 2B) may also contribute to antisense reads, but the high similarity between antisense read peaks in a TSO-free experiment compared to TSO-containing control implies that TSO is not critical for the mechanism.

The high correlation between sense UMI counts and antisense UMI counts per-gene (slope ~1 for a suboptimal antisense sample, Figure 3B) implicates poly-A priming followed by strand invasion (2C) as the potential mechanism. If true, intronic reads and antisense reads are related because poly-dT strand invasion is a premature template switch and

leaves the rest of the mRNA exposed for internal priming. Note that this could also occur with TSO priming followed by strand invasion, which would also leave the rest of the mRNA open to internal priming. However it is not clear if the observed correlation is just a result of both types of reads being a function of gene expression. For certain samples with high antisense rates, positions of intronic UMI counts and antisense UMI counts on genes both increase from 3' to 5' on the mRNA (data not shown).

The effect of intronic and antisense reads on gene expression data

Sense intronic UMIs correlate with sense exonic UMIs and have a bias for poly-A-rich genes while antisense UMIs correlate with sense UMI counts and have a bias for poly-T-rich (which is poly-A according to the orientation of the read) genes as shown in Single Cell 3' v3.1 Gene Expression data derived from mouse neuronal nuclei (Figure 3). The relative increase in UMI counts when including intronic UMIs has a gene length bias, possibly because longer genes have more poly-A tracts. There is also a bias for genes with higher frequency of poly-A tracts taking into account the correction for length (Figure 7A). Similarly, the increase in UMI counts when including antisense UMIs (intronic and exonic, compared to only intronic and exonic sense UMIs) has a gene length bias as well as a strong bias for genes with higher frequency of poly-T tracts (Figure 7B). This implies that antisense UMIs introduce a further gene length bias even relative to results with sense intronic UMIs counted, which already have a bias. The difference is not likely caused by differences in the rate of poly-A sites compared to poly-T sites in introns, as those rates are similar ($6.3*10^{-4}$ vs $8.5*10^{-4}$, as a fraction of 7-mers, respectively). The observed gene length and poly-A biases of intronic reads imply that transcripts from long genes are being primed at multiple sites and are being assigned multiple UMIs as a result. Similarly, the gene length and poly-T biases of antisense reads could imply that first strand cDNAs (multiple present per transcript) are being primed at multiple sites.

The primary reason to include intronic UMIs in gene expression analysis is to improve sensitivity, such as the mean genes-per-cell detected. Figure 8A demonstrates that the fold improvement in mean genes-per-cell correlates with the fraction of intronic sense UMIs: 5' Single Cell Gene Expression experiments have the lowest fraction of intronic UMIs (< 0.15) and the lowest fold change in genes-per-cell (< 1.25), 3' Single Cell (and nuclei) Gene Expression experiments have higher fraction of intronic UMIs (> 0.15) and larger fold change in genes-per-cell (> 1.2). Although the distributions are highly intermixed between cell and nuclei experiments, there are distinct differences in paired cell-nuclei experiments. For the same assay, cell type, and cell load, in 8 out of 10 paired cell-nuclei experiments, the nuclei data has greater intronic fraction and fold change genes-per-cell compared to the data derived from cells. The gain in sensitivity includes the detection of genes that have intronic UMIs but 0 exonic UMIs in any cell (data not shown). As a result of this large sensitivity improvement for single nuclei experiments, intronic UMIs are counted by default using `cellranger-arc count` with the nuclei-based Single Cell Multiome ATAC + Gene Expression product. Intronic UMIs may be excluded by passing the `--gex-exclude-introns` flag to the `cellranger-arc count` command.

Because intronic reads are more frequent in genes with higher poly-A content, it follows that using intron mode would have an effect on downstream analyses such as cell clustering. Indeed, the composition of cell clusters determined using graph-based clustering differ when including intronic UMIs compared to if they are excluded. Consistent with relative amounts of intronic UMI content, the changes in cluster composition (measured by the adjusted Rand index) appear to be more pronounced for nuclei samples than for cell samples, and less pronounced for Single Cell 5' Gene Expression samples than 3' Gene Expression samples (Figure 8B).

Conclusions

This Technical Note demonstrates that intronic and antisense UMIs correlate with sense UMI signal, implying that neither phenomenon is the result of priming from genomic DNA. The Note provides evidence that internal priming of poly-A tracts on the RNA results in sense intronic reads in Single Cell Gene Expression data. Counting these intronic reads alongside exonic reads can yield significant increases in assay sensitivity, especially for nuclei samples. Moreover, intronic UMI inclusion leads to the detection of genes that have no exonic UMIs in any cell. On the other hand, this internal priming appears to result in the assignment of multiple UMIs to single transcripts. This causes a gene length bias for intronic UMIs that should be taken into account when considering including intronic UMIs in Single Cell Gene Expression analysis. Four potential mechanisms to explain antisense reads in Single Cell 3' Gene Expression data were tested. The data supports two of the proposed mechanisms most strongly,

even though it is challenging to discern between the two mechanisms with the available data. Antisense reads appear to be generated either by internal priming on the first-strand cDNA or by an intermolecular template-switch onto the poly(dT) primer at poly-T sites. As with intronic UMIs, the decision to count antisense UMIs should be made after careful consideration of the potential biases that it may introduce in data analysis.

References

Ding, J. et al., Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nature biotechnology,* pp.1-10 (2020).

La Manno, et al., RNA velocity of single cells. *Nature,* 560(7719), pp.494-498 (2018).

Perocchi, F. et al., Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. Nucleic acids research, 35(19), p.e128 (2007).

Mereu, E. et al., Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. Nature Biotechnology, pp.1-9 (2020).

Peng, S. et al., November. Probing glioblastoma and its microenvironment using single-nucleus and single-cell sequencing. In IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 2757-2762). IEEE (2019).

Kreimann, K. et al., Ischemia reperfusion injury triggers CXCL13 release and B-cell recruitment after allogenic kidney transplantation. Frontiers in Immunology, 11 (2020).

Houseley, J. and Tollervey, D. Apparent non-canonical trans-splicing is generated by reverse transcriptase in vitro. PloS one, 5(8), p.e12271 (2010).

Contact:
support@10xgenomics.com
10x Genomics
6230 Stoneridge Mall Road
Pleasanton, CA 94588 USA