Interpreting Single Cell Gene Expression Data With and Without Intronic Reads

Introduction

In whole transcriptome single cell data from Chromium Single Cell 3' and 5' products, reads primarily map to exonic regions derived from mature spliced mRNA transcripts. However, reads may also map entirely or partially to intronic regions (Figure 1). This Technical Note evaluates the effects of counting intronic reads for Single Cell Gene Expression cellular data.

The "Interpreting Intronic and Antisense Reads in 10x Genomics Single Cell Gene Expression Data" (Document CG000376) Technical Note investigated mechanisms for intronic read generation. Intron sequences can be captured if the 10x Genomics gel bead poly(dT) oligos prime to internal poly-A tracts in the mRNA, rather than to the 3' poly-A tails. This mechanism was supported by analysis of UMI frequency and position in poly-A tracts for Single Cell 3' and 5' Gene Expression datasets, and was observed by other studies (Ding et al., La Manno et al.). Intronic reads accounted for 20-40% of UMI counts across several cellular and nuclei gene expression datasets, with a higher fraction for nuclei datasets (see Document CG000376 Table 1). Furthermore, analysis sensitivity improved when intronic reads were included compared to exon-only analysis, especially for nuclei samples, by increasing the amount of usable data for downstream analyses (e.g., more UMI counts, genes per cell, and mapped reads).

While it is already common to include intronic reads for nuclei data analysis (Peng et al., Kreimann et al.), this Technical Note further investigates the impact of using intronic reads for cellular samples. The following data metrics and analyses were compared: 1) sensitivity across a variety of sample types and species for Single Cell 3' and 5' Gene Expression datasets and 2) cell type classification and differential gene expression for a Single Cell 3' Gene Expression 10k peripheral blood mononuclear cell (PBMC) dataset.

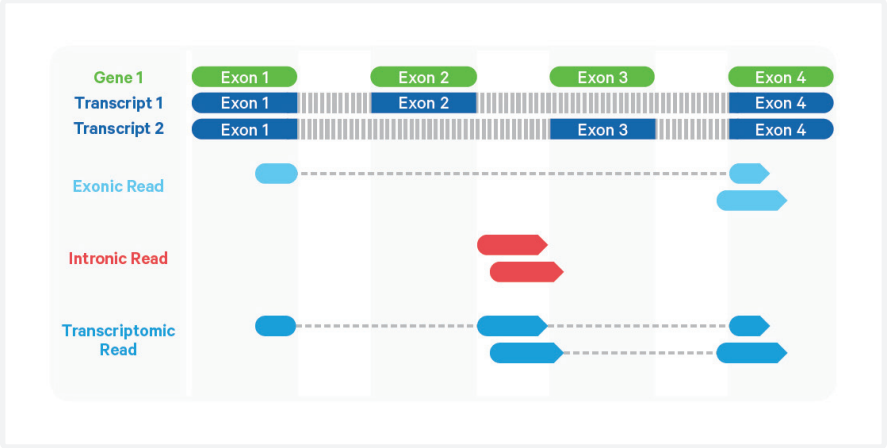

Figure 1. Genes include both exonic and intronic segments. Reads map to these transcripts in a strand-specific manner (for Single Cell 3' Gene Expression, sense reads face the end of the gene where a poly-A tail is appended to the transcript). Single Cell Gene Expression data primarily includes sense reads that map to exons, but may also include sense intronic reads. **As of version 7.0, Cell Ranger defaults to counting both exonic and intronic reads in the sense orientation (transcriptomic reads) that are confidently (uniquely) mapped to an annotated transcript by default.**

Methods

Chromium Single Cell 3' (v3.1, dual index) and Chromium Single Cell 5' v2 libraries were generated from single cell suspensions derived from several cell lines (A375, A549, SKOV3, U20S, Barnyard), peripheral blood mononuclear cells (PBMCs from human, mouse, rat, and rhesus monkey), and dissociated primary tissues. The libraries were sequenced as described in their respective user guides.

The data were processed with Cell Ranger 6.1.2 in two ways: 1) with intronic reads ("intron-mode") and 2) without intronic reads ("exon-only") using the include-introns option. The outputs were assessed to compare several data metrics across datasets. The Single Cell 3' Gene Expression 10k PBMC outputs from Cell Ranger with and without introns were further analyzed separately with a 3rd party tool, Seurat v4.0.2 (Hao et al.), using the PBMC reference dataset and procedure described here for cell type classification. Differential gene expression analysis results from Cell Ranger were visualized in Loupe Browser 6.0.0 using the cell type classifications defined by Seurat as clusters. Differential gene expression was computed for each cell type and for intron-mode and exon-only datasets separately.

Results

Usable Data and Sensitivity Increases with Intronic Reads

Cell Ranger outputs several metrics summarizing data retention throughout the pipeline. Prior to Cell Ranger 7.0, reads mapping to introns were excluded from downstream analysis by default. By including reads mapping to introns, more reads contribute to differential gene expression analysis and less data is wasted. Based on summary statistics from Cell Ranger, the fraction of usable data for downstream analysis increased when intron sequences were included in Single Cell 3' Gene Expression analysis regardless of sample type or species (Figure 2). The Cell Ranger metric, "Reads Mapped Confidently to Transcriptome", indicates the subset of reads deemed to have come from mRNA and are used for UMI counting. For a read to be considered transcriptomic, it should map uniquely to the genome, only map to one gene, and have splice junctions consistent with annotated transcripts (Figure 1). The increase in mapped transcriptomic reads correlates directly with the fraction of reads mapping to introns in a given sample.

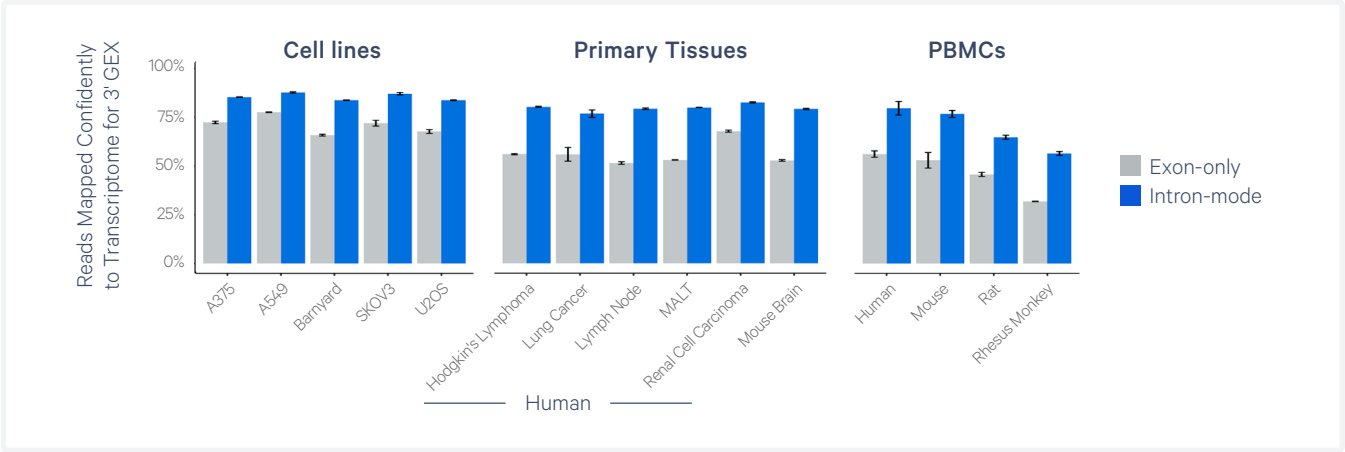

Figure 2. Plot of the "Reads Mapped Confidently to Transcriptome" metric from Cell Ranger averaged across technical replicates with standard deviation for several sample types. Single Cell 3' Gene Expression (GEX) results indicate an increase in mapped transcriptomic reads for all examined sample types and species when introns are included (blue) vs. exon-only (gray).

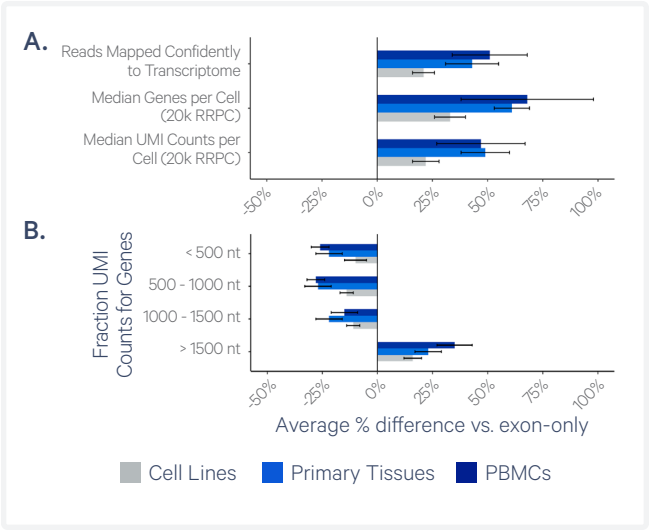

Figure 3. An increase in A) mapped reads and sensitivity metrics (RRPC = raw reads per cell) and B) the fraction of UMI counts on genes > 1,500 nucleotides (nt) was observed for all sample types with intron-mode. This is shown by the average percent difference in metrics for technical replicates (with standard deviation) with intron-mode vs. exon-only analysis of Single Cell 3' Gene Expression data for cell lines (gray), dissociated primary tissues (blue), and PBMCs (dark blue).

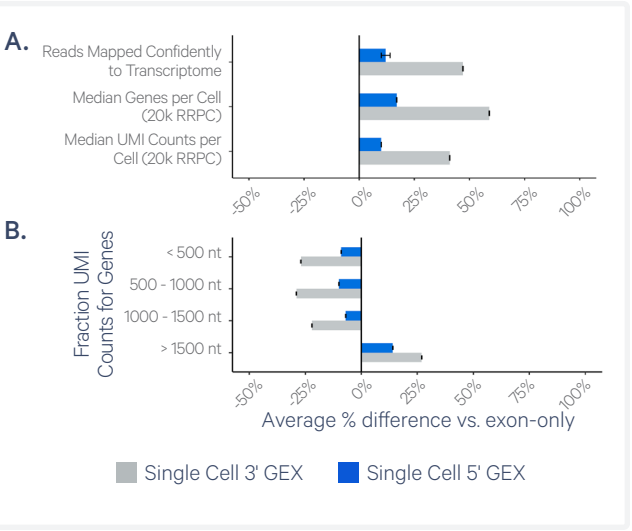

Figure 4. The trends in increased A) mapped reads and sensitivity metrics (RRPC = raw reads per cell) and B) fraction of UMI counts on genes > 1,500 nucleotides (nt) were similar for Single Cell 3' (gray) and 5' (blue) Gene Expression (GEX) PBMC datasets. This is shown by the average percent difference in metrics for technical replicates (with standard deviation) with intron-mode vs. exon-only analysis for both datasets.

Furthermore, there was an increase in sensitivity measured as "Median Genes per Cell (20k raw reads per cell (RRPC))" and "Median UMI Counts per Cell (20k RRPC)" when the same data was analyzed with introns across all tested sample types (Figure 3a). Note that sensitivity measurements were depth-normalized to 20k RRPC for fair comparison across datasets.

As observed in the previous Technical Note (Document CG000376), there is a gene-length bias in intron-mode: the fraction of UMI counts was higher for genes > 1,500 nucleotides (nt) and lower for all categories of shorter genes (Figure 3b). This length bias may result from the presence of more poly-A tracts in longer genes.

Single Cell 5' Gene Expression data showed similar increases in usable data as well as the length bias, however the differences were smaller as these datasets contain a lower fraction of intronic reads compared to Single Cell 3' Gene Expression data (Figure 4).

Analysis of Single Cell 3' Gene Expression 10k PBMCs

A Single Cell 3' Gene Expression 10k PBMC dataset was selected for further analysis since PBMC samples showed the highest differences in metrics compared to other sample types.

Metric	Exon-only	Intron-mode
Estimated Number of Cells	11,996	11,984
Mean Reads per Cell	41,379	41,421
Fraction Reads in Cells	96.0%	96.4%
Reads Mapped Confidently to Transcriptome	52.7%	77.7%
Median Genes per Cell	2,049	3,268 (+59%)
Median UMI Counts per Cell	6,896	10,186 (+48%)
Total Genes Detected	25,589	30,351 (+19%)

Table 1. A comparison of general cell statistics from the Cell Ranger web summary with and without introns (see Cell Ranger Cell Statistic Definition table). There was no significant change in cell calling for this dataset, however there was a large increase in sensitivity when introns were counted.

The estimated number of cells and proportion of reads in cells was similar for the exon-only and intron-mode analyses. However, 34.3% of the reads mapped to intronic regions. Therefore, by including introns, a significant amount of usable data was added for gene expression analysis relative to the exon-only dataset (Table 1). Note that cell calling differences may vary more between analysis modes for complex sample types.

PBMC Cell Type Classification is Similar

Next, the effect of including introns was examined for cell type classification. The exon-only dataset contained a total of 11,996 cell-associated barcodes, while the intron-mode dataset contained 11,984 cell-associated barcodes (Table 1).

Of these, nearly all (11,921, 99.37%) barcodes were shared between the two datasets. The Seurat classification assigns each cell to one of eight types (Level 1) and one of 31 subtypes (Level 2). The eight cell type categories include B cells, CD4+ T cells, CD8+ T cells, Dendritic Cells (DC), Monocytes, Natural Killer

Cell type	Exon-only	Intron-mode
B	9.61%	10.05%
CD4+ T	24.40%	24.49%
CD8+ T	15.87%	16.23%
Dendritic (DC)	2.14%	2.16%
Monocytes	39.80%	39.23%
Natural Killer (NK)	4.12%	4.43%
Other T	2.06%	1.89%
Other	1.99%	1.42%

Table 2. The percentage of cells assigned to the eight cell types was similar for exon-only and intron-mode analyses.

Cell Ranger Cell Statistic Definitions	
Estimated Number of Cells	The number of barcodes identified by the cell-calling algorithm as containing a cell.
Mean Reads per Cell	The total number of sequenced read pairs divided by the number of cell-associated barcodes.
Fraction Reads in Cells	The fraction of valid-barcode, valid-UMI, confidently-mapped-to-transcriptome reads with cell-associated barcodes.
Reads Mapped Confidently to Transcriptome	Fraction of reads that mapped to a unique gene in the transcriptome. The read must be consistent with annotated splice junctions. These reads are considered for UMI counting.
Median Genes per Cell	Median number of read pairs sequenced from the cells assigned to this sample (at least one UMI count detected).
Median UMI Counts per Cell	Median number of UMIs obtained from the cells assigned to this sample.
Total Genes Detected	The number of genes with at least one UMI count in any cell.

Cell type	Number of cells changed	Most likely change
B	1 (0.09%)	NK (100%)
CD4+ T	71 (2.44%)	CD8+ T (83.10%)
CD8+ T	58 (3.05%)	CD4+ T (60.34%)
Dendritic (DC)	5 (1.95%)	B (60.00%)
Monocytes	67 (1.41%)	B (58.21%)
Natural Killer (NK)	3 (0.61%)	B (66.67%)
Other T	35 (14.20%)	CD8+ T (51.43%) / NK (45.71%)
Other	85 (39.91%)	CD4+ T (48.23%) / CD8+ T (27.06%)

Table 3. Change in Level 1 classification by cell type between exon-only and intron-mode. The first column is the cell type for exon-only mode cells. The second column shows the percentage of cells with changes in classification when introns were included. The third column shows the cell types that cells were most frequently classified as in intron-mode, which were predominantly changes between T cell categories.

(NK) cells, Other T (mature T cells), and Other, which includes progenitors and rare populations expressing erythroid or platelet lineage markers (Hao et al.).

The cell type classifications were nearly the same with and without introns: 11,596 (97.27%) of the cells had the same cell type classification and 11,195 (93.91%) had the same subtype. All the major PBMC cell types were observed in both analyses (Table 2). The accuracy of cell type classification analysis depends on cell type and sample complexity. 325 cells were assigned different cell types, with the majority of these classifications switching between T cell or the Other categories (Table 3).

Top Differentially Expressed Genes are Similar

1,174 genes had more than one count per cell in the intron-mode dataset, compared to 445 genes for exon-only. Of these, 425 genes were present in both datasets. Differential expression was largely unchanged for most cell types, but improved for others (e.g., reduction in median log p-value for B cells). For each cell type, the top 50 up-regulated genes and top 50 down-regulated genes were generated separately for intron-mode and exon-only datasets, and then compared to identify shared genes. The top 50 up-regulated genes were nearly identical, with 49 of 50 differentially expressed genes shared between intron-mode and

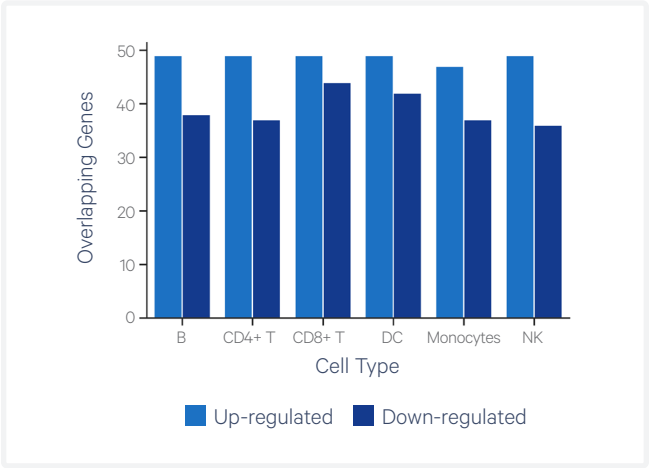

Figure 5. For cell type clusters, the top 50 up-regulated (blue) genes are nearly identical between intron-mode and exon-only analysis, with slightly lower overlap for the top 50 down-regulated (dark blue).

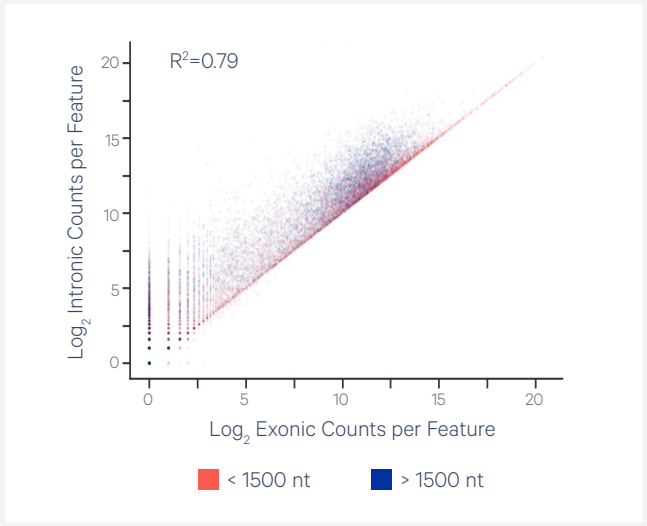

Figure 6. Plot of the aggregate counts across all cells for each feature present in the filtered feature-barcode matrix for protein-coding transcripts captured in exon-only and intron-mode. Each point in the plot represents an individual feature. Points along the diagonal represent features with no change in UMI counts between intron-mode and exon-only mode, while points above the diagonal represent features with higher expression in intron-mode. Points with > 1 intronic count but no exonic counts are likely due to the presence of genomic DNA in the intron-mode dataset. Point color represents transcript sequence length based on the longest isoform (red < 1,500 nucleotides (nt), blue ≥ 1,500 nt).

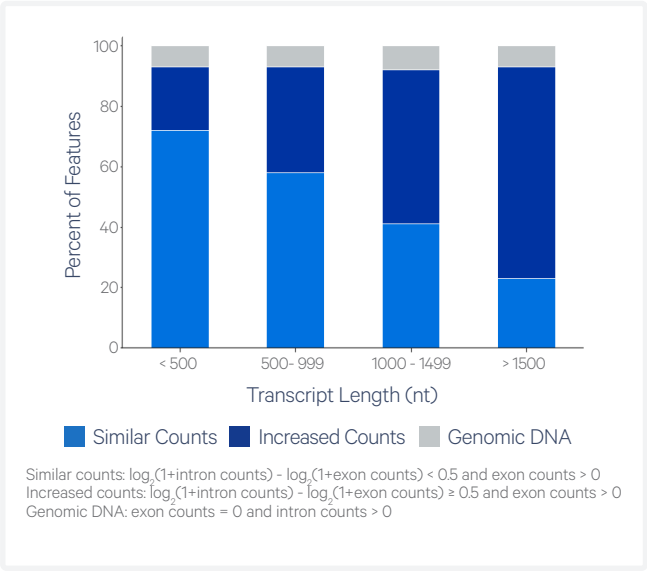

Figure 7. Stacked barplot of feature categories by transcript length, showing the percentage of features for which exon-only and intron-mode counts are similar (blue), there are more intron than exon counts (dark blue), and there are intron counts but no exon counts (gray). As expected, the dark blue fraction increased dramatically with longer sequences > 1,500 nt. The gray portion shows the percentage of transcripts that most likely represent genomic DNA.

exon-only datasets for all cell types except monocytes (47 of 50 genes shared) (Figure 5). 35-40 of the top 50 down-regulated genes overlapped.

As observed in the Single Cell 3' and 5' Gene Expression datasets discussed earlier, there was a length bias towards higher fractions of UMI counts on transcripts > 1,500 nt for intron-mode datasets. This trend was observed in the 10k PBMC dataset as well, where transcript lengths > 1,500 nt increased for intronic vs. exonic counts (Figures 6 and 7).

Genomic DNA (gDNA) transcripts are likely present in intron-mode data based on the observation of high intronic counts per feature for features with no exonic counts among transcripts annotated as "protein coding" (Figure 6 and 7). gDNA may artificially inflate the counts of lowly expressed genes, and consequently, down-regulated gene lists may be less reliable in intron-mode compared to exon-only. However, the fraction of gDNA was a small percentage of total UMIs (0.08%) and was a consistent 7-8% of features regardless of transcript length for this dataset (Figure 7). The gDNA UMI

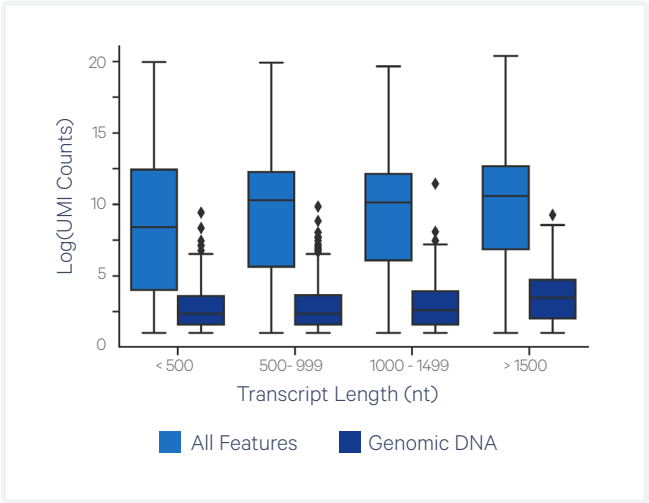

Figure 8. UMI counts per feature in genomic DNA features (dark blue) were lower than those for all features (blue). Features arising from gDNA are defined as in the previous stacked barplot (Figure 7). "All Features" represents all features included in the feature list, excluding those with no intron or exon counts. UMI counts were aggregated across cells on a per feature basis for all protein-coding features.

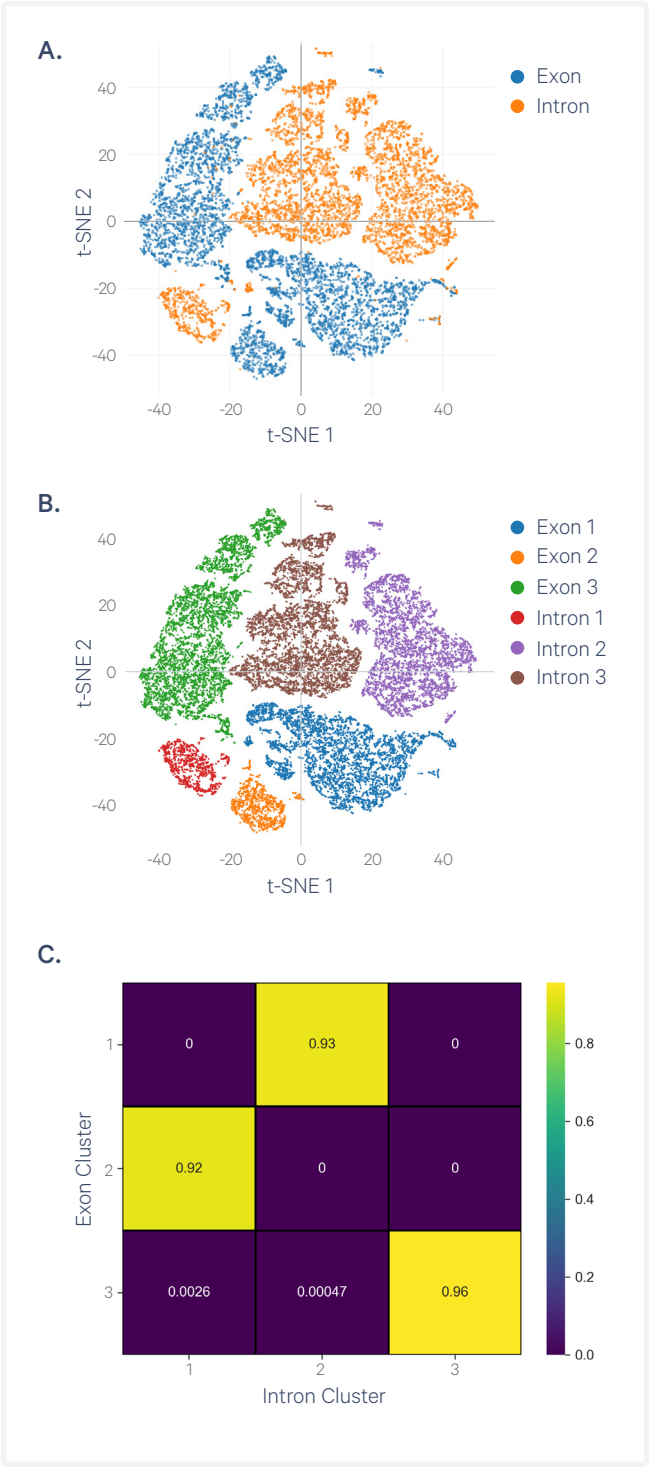

Figure 9. Clusters are shown by A) intron-mode (orange) vs. exon-only (blue) dataset and B) six manually annotated intron-mode and exon-only clusters from Loupe Browser. The clusters are almost entirely distinct. C) The heatmap shows pairwise Jaccard similarity between the manually annotated clusters. The barcode overlap between clusters is nearly 1:1 for each intron-mode and exon-only pair of clusters (yellow squares).

counts were significantly lower than non-gDNA features across transcript lengths, and represent ~0.1% of the overall dataset (Figure 8).

Exon-only and Intron-mode Datasets Cannot be Aggregated

The datasets for these comparisons were analyzed separately with the `cellranger count` pipeline. An aggregated analysis of exon-only with intron-mode data using `cellranger aggr` is not supported, as such analyses result in artificial non-overlapping clusters (Figure 9a). However, the intron-mode clusters appear to be shifted versions of exon-only clusters based on barcode similarity between manually annotated clusters (Figure 9c), suggesting that cluster quality is not dependent on whether introns are included.

Conclusions

This Technical Note demonstrates how including intronic reads in Cell Ranger analysis results in more usable data and higher sensitivity for both Single Cell 3' and 5' Gene Expression datasets for multiple sample types. Secondary analysis of the Single Cell 3' Gene Expression 10k PBMC dataset with and without introns suggests the impact of the gene-length bias and presence of genomic DNA are low. Furthermore, for this particular sample type, there were no significant differences in biological interpretation (e.g., cell type classification and differential gene expression analyses) between the exon-only and intron-mode analysis modes.

References

1. Ding, J. et al., Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nature biotechnology*, pp.1-10 (2020).
2. La Manno, et al., RNA velocity of single cells. *Nature*, 560(7719), pp.494-498 (2018).
3. Peng, S. et al., 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).
4. Kreimann, K. et al., Ischemia reperfusion injury triggers CXCL13 release and B-cell recruitment after allogenic kidney transplantation. *Frontiers in Immunology*, 11 (2020).
5. Hao, Y. et al., Integrated analysis of multimodal single-cell data. *Cell*, 184 (2021).

Datasets

- https://www.10xgenomics.com/resources/datasets/10k-human-pbmcs-3-v3-1-chromium-x-with-intronic-reads-3-1-high
- https://www.10xgenomics.com/resources/datasets/10k-human-pbmcs-3-v3-1-chromium-x-without-introns-3-1-high

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