TECHNICAL NOTE

Interpreting Cell Ranger ARC Web Summary Files for Single Cell Multiome ATAC + Gene Expression Assay

Introduction

The web summary file in the output folder of Cell Ranger ARC analysis software is the initial point of reference for determining sample performance in the Chromium Single Cell Multiome ATAC + Gene Expression assay. This Technical Note presents an overview of how to interpret the web summary file, including a selection of metrics used for QC and characteristic plots generated using the Single Cell Multiome ATAC + Gene Expression workflow.

Interpreting Web Summary File Metrics

Representative summary files for Chromium Single Cell Multiome ATAC + Gene Expression libraries and other Cell Ranger ARC output files are available for download on the 10x Genomics Support website. The top of the web summary file displays key metrics (Figure 1). Green text indicates that the key metrics are in the expected range while red/yellow text indicates errors/warnings. Underneath the key metrics, there are three analysis tabs that display results for joint analysis metrics and plots as well as ATAC and Gene Expression specific metrics. Descriptions of the metrics can also be found by clicking the icon ⓘ next to the section header.

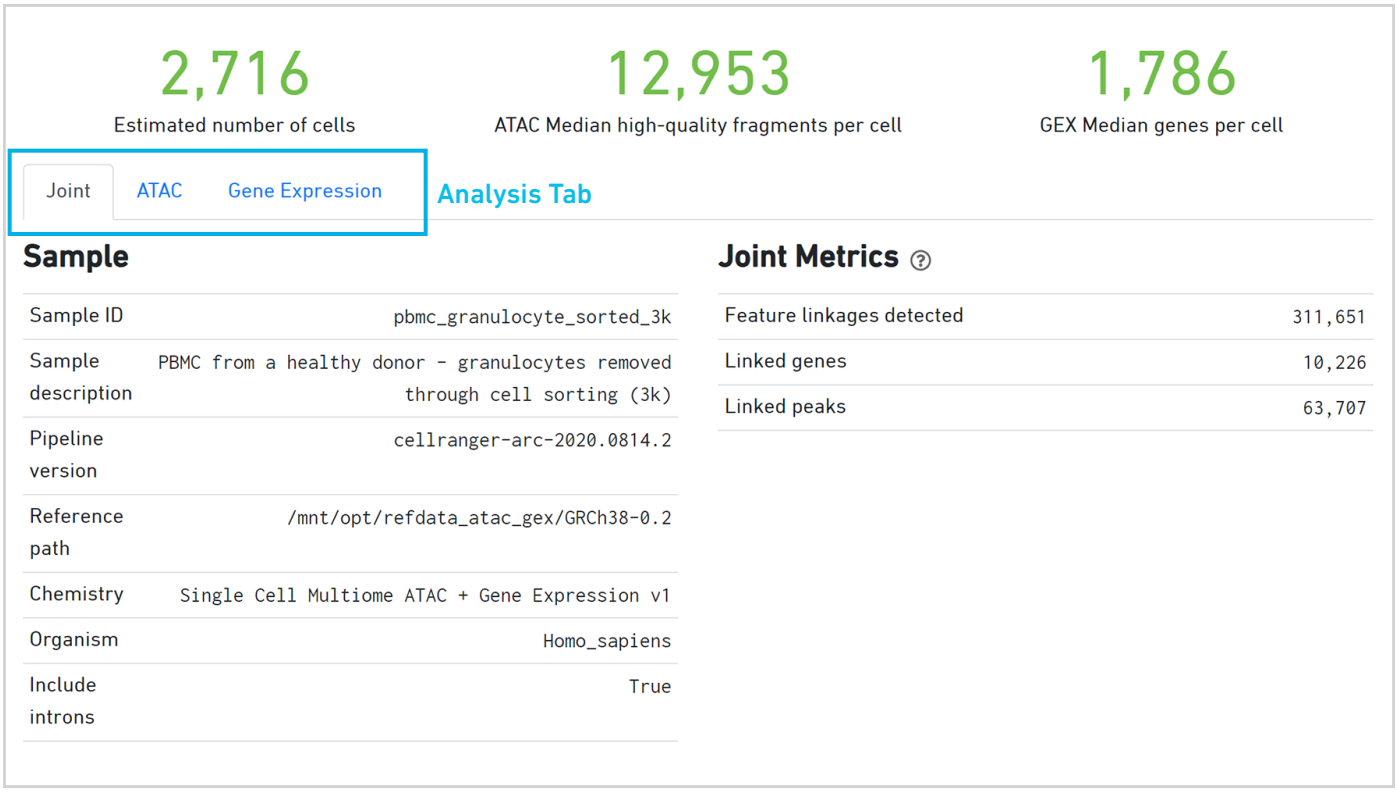


Figure 1. A representative web summary file (top section) for a PBMCs sample targeting 3,000 nuclei.

Joint Analysis Tab

The joint analysis tab (Figure 1) contains sample information on the parameters used to run the analysis such as pipeline version, reference path and organism. Cell Ranger ARC maps intronic reads as a default which is specified as "TRUE" in the "Include introns" argument. This table will expand to show any custom pipeline parameters set by the user such as force-cells, and custom peak.bed files.

Table 1. Metrics in the Cell Ranger ARC ATAC + Gene Expression web summary file Joint analysis tab.

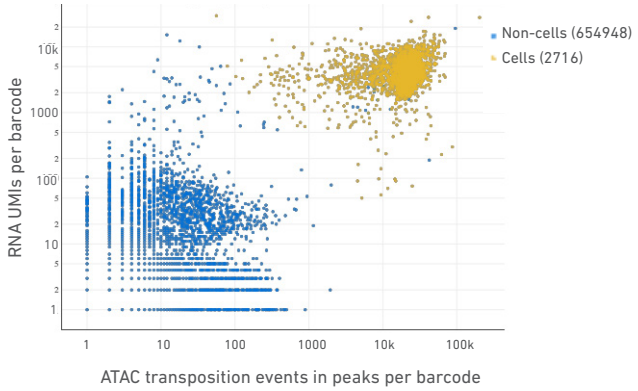
Metrics	Definition	Expected Value	Notes
Joint Metrics			
Feature linkages detected	Total number of gene-peak and peak-peak linkages detected	>100	Total number of feature linkages are dependent on sample type and sequencing depth. A low number of feature links detected may be caused by a low number of nuclei recovered, low sequencing depth, poor peak calling, or a sample that is relatively homogeneous.
Linked genes	Total number of genes that are linked to peaks	Dependent on sample type and sequencing depth	Total number of linked genes is dependent on sample type. A low number of linked genes may be caused by a low number of nuclei recovered, low sequencing depth, poor peak calling, or a sample that is relatively homogeneous.
Linked peaks	Total number of peaks that are linked to genes or other peaks	Dependent on sample type and sequencing depth	Total number of linked peaks is dependent on sample type. A low number of linked peaks may be caused by a low number of nuclei recovered, low sequencing depth, poor peak calling, or a sample that is relatively homogeneous.

Interpreting the Plots in the Joint Analysis Tab

Table 2. Plots in the Cell Ranger ARC ATAC + Gene Expression web summary file Joint analysis tab.

Plot & Interpretation
Cross-sensitivity Plot: Each point in the cross-sensitivity plot represents a barcode. Every cell is shown as a yellow point. The non-cell barcodes are downsampled and shown in blue. The position on the x-axis reflects the number of ATAC transposition events in peaks associated with that barcode, while the position on the y-axis corresponds to the number of gene expression UMI counts.
Example

Ideal Sample: Cell barcodes cluster at the upper right hand corner. Cell-associated barcodes contain both a large number of ATAC transposition events in peaks and a large number of RNA UMIs. A clear separation between cells and non-cells is observed.



Compromised Sample: Cell-associated barcodes show low ATAC transposition events in peaks or low RNA UMIs. No clear separation of Cells and Non-cells is observed. Lack of separation can point to emulsion failures or poor sample preparation.

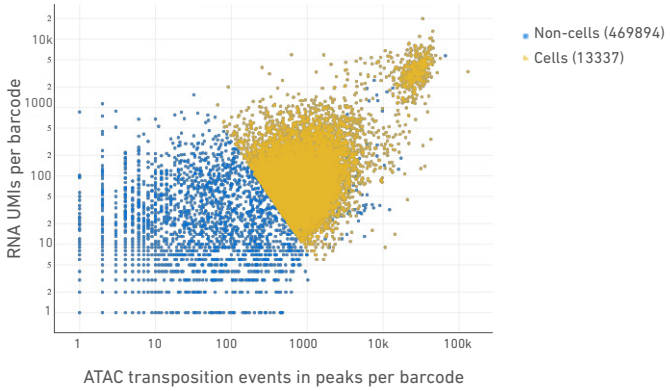
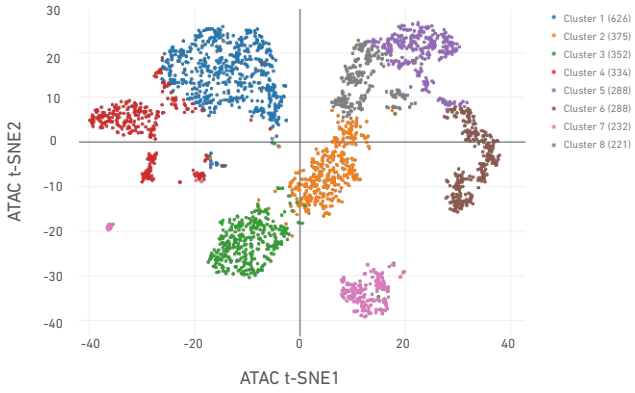


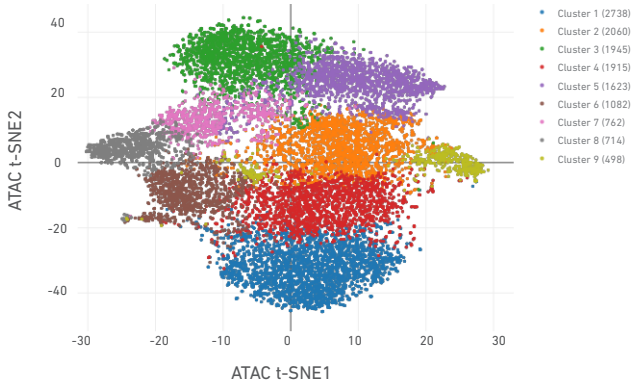
Table 2 contd. Plots in the Cell Ranger ARC ATAC + Gene Expression web summary file Joint analysis tab.

Plot & Interpretation
Cell Clustering Scatter Plots: The Cell Clustering (colored by cluster) plot shows the cell-associated barcodes in a 2-D t-SNE projection, with colors showing an automated graph clustering analysis which groups together cells with similar peak profiles. The ATAC Clusters plot shows cell-associated barcodes clustered by ATAC data, and the Gene Expression Clusters plot shows cell-associated barcodes clustered by Gene Expression data.
Example

Ideal Sample: Structured clusters with good separation (for a sample with expected heterogeneous cell populations).



Compromised Sample: Lack of cluster structure, one large cluster or no separation (for a sample with expected heterogeneous cell populations) may indicate sample quality issue or loss of single cell behavior.



ATAC Analysis Tab

Table 3. Metrics in the Cell Ranger ARC ATAC + Gene Expression web summary file ATAC analysis tab

Metrics	Definition	Expected Value	Notes
Sequencing Metrics			
Sequenced read pairs	Total number of sequenced read pairs assigned to the Multiome ATAC library	User defined	Suggested sequencing depth is 25,000 read pairs per cell. Sequencing depth may need to be adjusted based on the sample complexity and experimental goals. Low ATAC sequencing depth negatively impacts the quality of peak calling, clustering, differential analysis, and feature linkage.
Valid Barcodes	Fraction of read pairs with barcodes that match the whitelist after error correction	>85%	Low valid barcodes may indicate a problem in the library preparation workflow, e.g., poor sequencing quality, or cycle failures in the Illumina i7 read. Application performance may be affected.
Percent duplicates	Fraction of high-quality read pairs that are deemed to be PCR duplicates. A high-quality read-pair is one with mapq >30 that is not chimeric and does not map to a non-nuclear contig. This is the fraction of high-quality fragments with a valid barcode that align to the same genomic position as another read pair in the library	Variable	This metric is a measure of sequencing saturation and is a function of library complexity and sequencing depth.
Targeting Metrics			
Number of peaks	Total number of peaks on primary contigs either detected by the pipeline or input by the user	Variable	Total number of peaks is dependent on sample complexity and sequencing depth. Very low peak numbers that cover a large fraction of the genome may indicate issues in peak calling.
Fraction of genome in peaks	Fraction of bases in primary contigs that are defined as peaks	<75%	Sample type dependent. A large fraction of genome called as peaks may indicate incorrect peak calling, loss of chromatin structure or a population of cells with highly accessible DNA (e.g., activated granulocytes, dead or dying cells, unsupported organisms).
TSS enrichment score	The TSS profile is the summed number of cut sites per base in a window of 2,000 bases around all the annotated TSSs, normalized by the minimum signal in the window	>5	TSS regions (known to be open) are expected to have lots of cut sites from accessible fragments. Low score may indicate poor sample preparation, poor sample quality, a population of cells with highly accessible DNA (e.g., activated granulocytes, dead or dying cells, unsupported organisms), or poor quality reference genome annotation.
Fraction of high quality fragments overlapping peaks	Fraction of high quality fragments in cell barcodes that overlap called peaks	>25%	Low fraction of fragments overlapping peaks can indicate a problem during the transposition step, a population of cells with highly accessible DNA (e.g., activated granulocytes, dead or dying cells, unsupported organisms), or unexpected behavior in the peak calling algorithm resulting in very few peaks called.

Table 3 contd. Metrics in the Cell Ranger ARC ATAC + Gene Expression web summary file ATAC analysis tab

Metrics	Definition	Expected Value	Notes
Cell Metrics			
Estimated number of cells	Number of barcodes associated with cell-containing partitions	500-10,000	± 20% expected value is acceptable. Higher or lower values outside of this range may indicate incorrect quantification of the nuclei suspension, improper handling of nuclei, excessive background RNA and DNA, or unexpected behavior in the cell calling algorithm. Warning if <300 or >12,000 cells called.
Mean raw read pairs per cell	Total number of read pairs divided by the number of barcodes	User defined, expected >5,000 read pairs per cell	This metric is a function of the number of cells and sequencing depth. Suggested sequencing depth is 25,000 read pairs per cell. Low ATAC sequencing depth negatively impacts the quality of peak calling, clustering, differential analysis, and feature linkage. At very low sequencing depth, < 5,000 raw read pairs per cell, identification of cell barcodes may be unreliable.
Fraction of high quality fragments in cell	Fraction of high quality fragments with a valid barcode that are associated with cell-containing partitions. High quality fragments are defined as read pairs with a valid barcode that map to the nuclear genome with mapq > 30, are not chimeric and not duplicate	>40%	A lower than expected value may indicate that many of the ATAC fragments were not assigned to cell-associated barcodes. This could be caused by high levels of ambient ATAC fragments caused by a poor nuclei preparation, or nuclei with highly accessible DNA as seen for instance in activated granulocytes.
Fraction of transposition events in peaks in cells	Fraction of transposition events that are associated with cell-containing partitions and fall within peaks	>25%	A lower than expected value could suggest a problem during library preparation causing excessive background transposition, low sequencing depth, or unexpected behavior in the peak calling algorithm resulting in very few peaks detected.
Median high quality fragments per cell	The median number of high-quality fragments per cell barcode	>100	A lower than expected value may indicate low sequencing depth, the wrong genome reference, or low library complexity that could be due to a problem during the transposition step or a problem in the library preparation workflow. Low fragment counts negatively impact clustering, differential analysis, and feature linkage detection.
Mapping Metrics			
Confidently mapped read pairs	Fraction of sequenced read pairs with mapping quality >30	>80%	A lower than expected value may indicate use of a wrong reference genome or a poor quality genome assembly. Application performance may be affected.
Non-nuclear read pairs	Fraction of sequenced read pairs that have a valid barcode and map to non-nuclear genome contigs, including mitochondria,with mapping quality > 30	<10%	Sample preparation dependent. Expected range of <10% when using recommended Demonstrated Protocol for nuclear extraction. A large number of non-nuclear read pairs may indicate a need for sample preparation optimization.

Interpreting the Plots in the ATAC Analysis Tab

Table 4. Plots in the Cell Ranger ARC ATAC + Gene Expression web summary file ATAC analysis tab

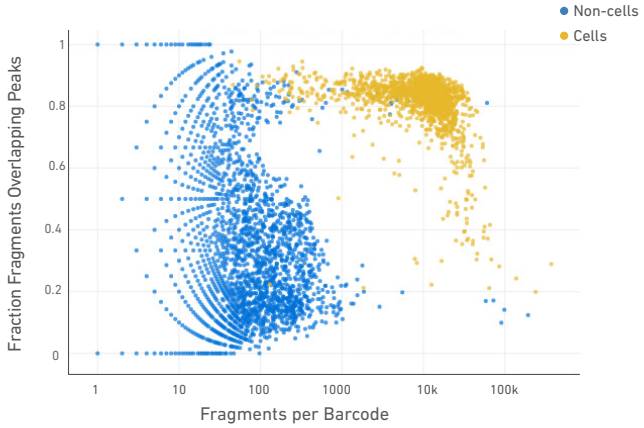
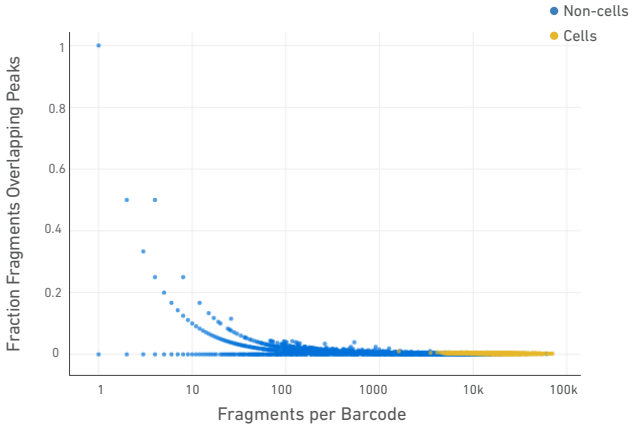
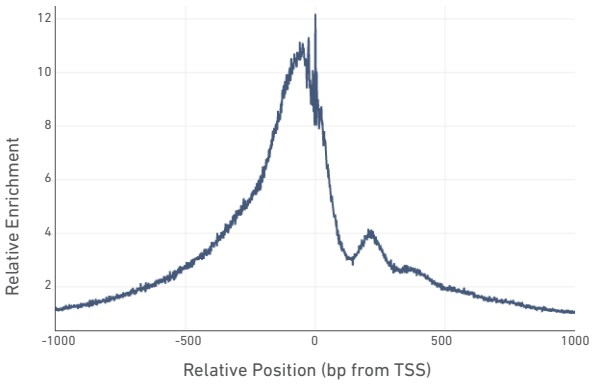
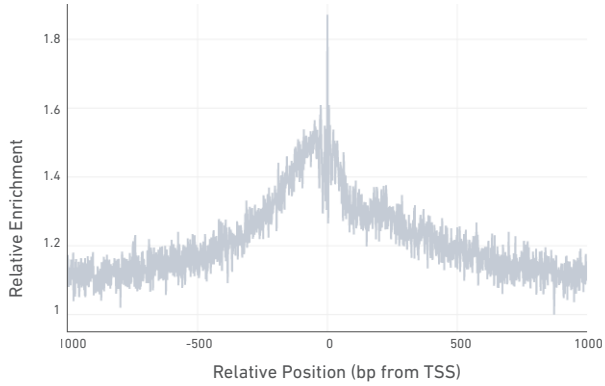
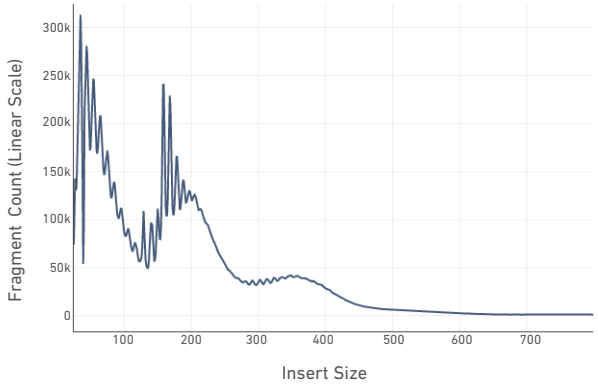
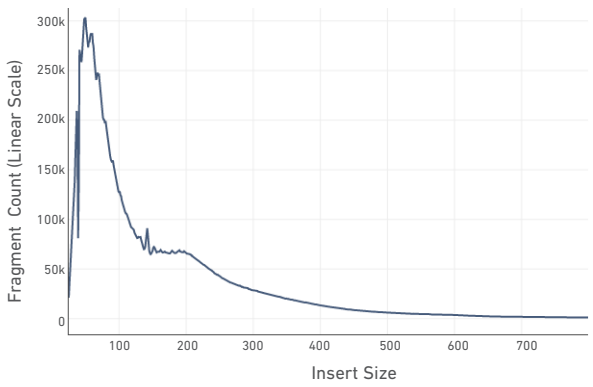
Plot & Interpretation
Single Cell Peak Targeting Plot: A scatter plot displaying the number of fragments per barcode and the percent of fragments overlapping peaks.
Example
Ideal Sample: Cell-associated barcodes are expected to have a large number of fragments per barcode and a high percentage of fragments overlapping peaks (upper right corner). Non-cell associated barcodes are expected to have a small number of fragments per barcode and a low percentage of fragments overlapping peaks (lower left corner). An ideal sample should show good separation of cells and non-cells at the opposite ends. 
Compromised Sample: Cell-associated barcodes have a low fraction of the barcode fragments overlapping peaks. Concentration of cell-associated and non-cell associated barcodes tightly concentrated in the same location may indicate issues with cell calling or sample preparation. 
Plot & Interpretation
Transcription Start Site (TSS) Plot: The Transcription Start Site (TSS) profile, which is computed as the number of cut sites per base, of all the barcodes irrespective of cell versus non-cell assignment in a window of 2,000 bases around the full set of annotated TSSs is displayed in the Transcription Start Site plot. The y-axis scale is normalized by the minimum signal in the window.
Example
Ideal Sample: Large enrichment around TSS, as these regions are known to have a high degree of chromatin accessibility compared to the flanking regions. 
Compromised Sample: Low enrichment (<4) around TSS sites may indicate improper lysis or loss of chromatin structure. 

Table 4 contd. Plots in the Cell Ranger ARC ATAC + Gene Expression web summary file ATAC analysis tab

Plot & Interpretation	
Insert Size Distribution Plot: Insert size distribution of transposase accessible fragments sequenced is displayed in the Insert Size Distribution plot.	
Example	
Ideal Sample: A periodicity of ~150 bp corresponds to the number of nucleosomes, the transposase accessible fragments span (nucleosome free, mononucleosome, and dinucleosome fragments). Sawtooth pattern in fragments with insert size <200 bp corresponds to the helical pitch of DNA (~10.5 bp).	**Compromised Sample:** Absence of periodicity and sawtooth pattern may indicate loss of chromatin structure due to low sample quality.
	

Gene Expression Analysis Tab

Table 5. Metrics in the Cell Ranger ARC ATAC + Gene Expression web summary file Gene Expression analysis tab

Metrics	Definition	Expected Value	Notes
Sequencing Metrics			
Sequenced read pairs	Total number of sequenced read pairs assigned to the Multiome Gene Expression library	User defined	Suggested sequencing depth of 20,000 read pairs per cell. Sequencing depth may need to be adjusted based on the sample complexity and experimental goals. Very low sequencing depth can negatively impact clustering, differential analysis and feature linkage.
Valid barcodes	Fraction of read pairs with barcodes that match the whitelist after error correction	>80%	Low valid barcodes may indicate poor sequencing quality. Application performance may be affected.
Reads with TSO	Fraction of reads with an alignment score ≥20 for the template switch oligo (TSO) sequence	<25%	Sequencing depth dependent. A high fraction of reads containing TSO can indicate sample quality or library preparation issues.
Percent duplicates	Fraction of high quality read pairs that are deemed to be PCR duplicates. A high-quality read-pair is one with mapq >30 that is not chimeric and does not map to a non-nuclear contig. This is the fraction of high-quality fragments with a valid barcode that aligns to the same genomic position as another read pair in the library	Variable	This metric is a measure of sequencing saturation and is a function of library complexity and sequencing depth.
Mapping Metrics			
Reads mapped to genome	Fraction of reads that are mapped to the genome	>80%	Dependent on the quality of genome annotation. Lower than expected values can indicate the use of an incorrect reference genome or a poor quality genome assembly. Application performance may be affected.
Reads mapped confidently to intergenic regions	Fraction of reads that mapped uniquely to an intergenic region of the genome.	<30%	May vary based on sample type and genome annotation.
Reads mapped confidently to intronic regions	Fraction of reads that mapped uniquely to an intronic region of the genome	Variable	Fraction of reads mapping to introns. The default analysis mode includes these reads in the analysis.
Reads mapped confidently to exonic regions	Fraction of reads that mapped uniquely to an exonic region of the genome	Variable	Value complimentary to reads mapped confidently to intronic reads.
Reads mapped confidently to transcriptome	Fraction of sequenced reads that map to a unique gene in the transcriptome. In the default mode the transcriptome includes intronic alignments	>50%	A lower than expected value can indicate use of the wrong reference transcriptome, a reference transcriptome with overlapping genes, poor library quality, poor sequencing quality, or reads shorter than the recommended minimum. Application performance may be affected.
Reads mapped antisense to gene	Fraction of reads that map to the transcriptome, but on the opposite strand of one or more overlapping annotated genes. These reads are consistent with annotated splice junctions and do not align in the sense orientation with any overlapping genes	<30%	Antisense reads tend to map to introns. These values are higher than typical whole cell samples due to the inclusion of intronic reads for analysis of nucleic data.

Table 5 contd. Metrics in the Cell Ranger ARC ATAC + Gene Expression web summary file Gene Expression analysis tab

Metrics	Definition	Expected Value	Notes
Cell Metrics			
Estimated number of cells	Number of barcodes associated with cell-containing partitions	500-10,000	± 20% expected value is acceptable. Higher or lower values outside of this range may indicate inaccurate nuclei count, nuclei lysis or failures during GEM generation. Warning if <100 or >12,000 cells called.
Mean raw read pairs per cell	Total number of read pairs divided by the number of cell barcodes	User defined, >5,000 read pairs per cell	This metric is a function of the number of cell and sequencing depth. Low Gene Expression sequencing depth negatively impacts the quality of clustering, differential analysis and feature linkage. At very low sequencing depth, < 2,000 raw read pairs per cell, identification of cell barcodes may be unreliable.
Fraction of transcriptomic reads in cells	Fraction of transcriptomic reads with a valid barcode that are associated with cell-containing partitions. Transcriptomic reads are defined as reads that map to a unique gene, including intronic alignments (default mode)	>60%	A lower than expected value may indicate that many of the reads were not assigned to cell-associated barcodes. This is generally indicative of poor sample preparation resulting in high levels of ambient RNA. It could also indicate a problem in the cell calling algorithm that could be caused by high RNA or DNA background, exclusion of a large number of barcodes from cell calling due to low targeting, or due to a population of nuclei with low RNA content. Application performance may be affected.
Median UMI counts per cell	The median number of UMI counts per cell barcode	>100	This metric is dependent on sample type complexity and sequencing depth. A lower than expected value may indicate very low sequencing depth, poor sample quality, an error in the library preparation workflow, the wrong reference genome, or poor genome annotations. Low UMI counts negatively impact clustering, differential analysis and feature linkage detection.
Median genes per cell	The median number of genes detected per cell barcode	Dependent on cell type and sequencing depth	This metric is dependent on sample type complexity and sequencing depth. A lower than expected value may indicate very low sequencing depth, poor sample quality, an error in the library preparation workflow, the wrong reference genome, or poor genome annotations.
Total genes detected	The number of genes with at least one UMI count in any cell barcode	Dependent on cell type and sequencing depth	This metric is dependent on sample type complexity and sequencing depth. A lower than expected value may indicate very low sequencing depth, poor sample quality, an error in the library preparation workflow, the wrong reference genome, or poor genome annotations.

References

- Chromium Next GEM Single Cell Multiome ATAC + Gene Expression User Guide (Document CG000338)
- Sequencing Metrics & Base Composition of Single Cell Multiome ATAC Library (Document CG000373)

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