Single-cell ATAC QC with Cell Ranger ATAC

scATAC-seq: attacking open chromatin in single cells (2024)

Christian Arnold

10.04.2024

EMBL

10x GENOMICS Cell Ranger ATAC Sample

← Summary Tab

10k_pbmc_ATACv2_nextgem_Chromium_X - Human PBMCs

For guidance, please consult "Interpreting Cell Ranger ATAC Web Summary Files" or contact 10x Genomics Support (support@10xgenomics.com).

10,273	20,046	65.9%
Estimated number of cells	Median high-quality fragments per cell	Fraction of high-quality fragments overlapping peaks

← Key Metrics

Summary | Data Quality

Sample

Sample ID	10k_pbmc_ATACv2_nextgem_Chromium_X
Sample description	Human PBMCs
Pipeline version	cellranger-atac-2.1.0
Reference path	…ata-cellranger-arc-GRCh38-2020-A-2.0.0
Chemistry	ATAC
Organism	Homo_sapiens

Sequencing

Sequenced read pairs	466,894,746
Valid barcodes	96.3%
Q30 bases in barcode	90.1%
Q30 bases in read 1	95.4%
Q30 bases in read 2	94.4%
Q30 bases in sample index i1	91.7%

Cells

Estimated number of cells	10,273
Mean raw read pairs per cell	45,448.72
Fraction of high-quality fragments in cells	95.7%
Fraction of transposition events in peaks in cells	62.5%
Median high-quality fragments per cell	20,046

Table 1. Metrics in the ATAC web summary file.

Metrics	Definition	Expected Value	Notes
Sequencing Metrics			
Sequenced read pairs	Total number of sequenced read pairs assigned to the sample	User defined	Suggested sequencing depth of 25,000 read pairs per cell.
Valid barcodes	Fraction of read pairs with barcodes that match the whitelist with error correction	>75%	Low valid barcodes may indicate sequencing related problem or issues with library preparation.
Q30 bases in barcode	Fraction of barcode read (i2) bases with Q-score ≥30	Sequencing platform dependent (ideally >65%)	Low Q30 base percentages could indicate sequencing issue such as sub-optimal loading concentration of the library.
Q30 bases in read 1	Fraction of read 1 bases with Q-score ≥30	Sequencing platform dependent (ideally >65%)	Expected to be higher than Q30 Bases in barcode (i5 read) or Sample Index (i7 read) and is sequencing platform dependent. Low Q30 base percentages could indicate sequencing issue such as sub-optimal loading concentration of the library.
Q30 bases in read 2	Fraction of read 2 bases with Q-score ≥30	Sequencing platform dependent (ideally >65%)	Expected to be higher than Q30 bases in barcode (i5 read) or Sample Index (i7 read) and is sequencing platform dependent. Low Q30 base percentages could indicate sequencing issue such as sub-optimal loading concentration of the library.
Q30 bases in sample index i1	Fraction of sample index read (i1) bases with Q-score ≥30	Sequencing platform dependent (ideally >90%)	Low Q30 base percentages could indicate sequencing issue such as sub-optimal loading concentration of the library.

Cell Metrics			
Estimated number of cells	The total number of barcodes identified as cells	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.
Mean raw read pairs per cell	Total number of read pairs divided by the number of cell barcodes	Dependent on sequencing depth	-
Fraction of high-quality fragments in cells	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 mapping quality (map Q) ≥30, are not chimeric, and not duplicate.	>40%	-
Fraction of transposition events in peaks in cells	Fraction of transposition events that are associated with cell-containing partitions and fall within peaks Transposition events are located at both ends of all high-quality fragments. This metric measures the percentage of such events that overlap with peaks.	>15%	-
Median high-quality fragments per cell	The median number of high-quality fragments per cell barcode	Dependent on cell type & sequencing depth	-

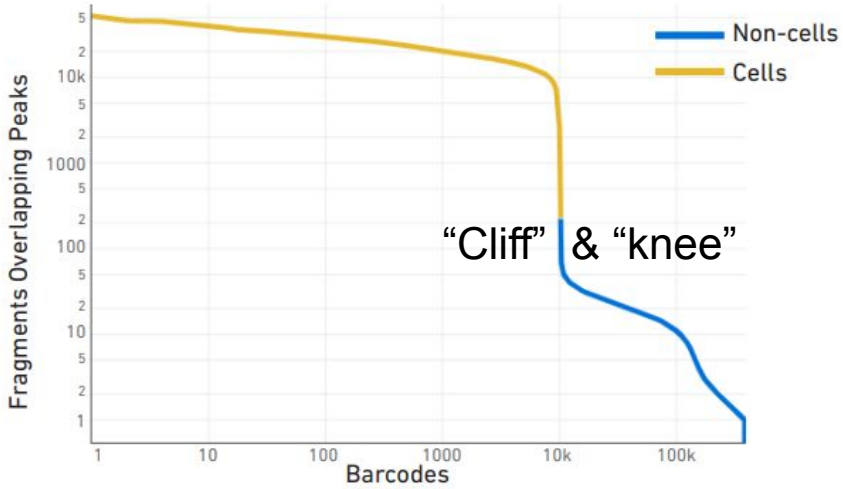
Metrics	Definition	Expected Value	Notes
Library Complexity Metric			
Percent duplicates	Fraction of high-quality read pairs that are deemed to be PCR duplicates. This is the fraction of high-quality fragments with a valid barcode that align to the same genomic position as another read pair in the library.	≥30%	A high-quality read-pair is one with mapping quality (mapQ) ≥30, that is not chimeric and maps to nuclear contigs. This metric is a measure of sequencing saturation and is a function of library complexity and sequencing depth.
Mapping Metrics			
Confidently mapped read pairs	Fraction of sequenced read pairs with mapping quality (mapQ) ≥30	>80%	-
Unmapped read pairs	Fraction of sequenced read pairs that have a valid barcode but could not be mapped to the genome	<5%	-
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 (mapQ) ≥30	<20%	-
Fragments in nucleosome-free regions	Fraction of fragments passing all filters with a size smaller than 124 basepairs	>40%	Expected to be the highest proportion as compared to mononucleosome and dinucleosome fragment.
Fragments flanking a single nucleosome	Fraction of fragments passing all filters with a size between 124 and 296 basepair	Dependent on sample type	An increased proportion of mononucleosome fragments may indicate dead/dying cells or granulocyte contamination.

Targeting Metrics			
Number of peaks	Total number of peaks on primary contigs either detected by the pipeline or input by the user	>45,000	-
Fraction of genome in peaks	Fraction of bases in primary contigs that are defined as peaks	>2% and <20%	-
TSS enrichment score	Maximum value of the transcription-start-site (TSS) profile. The TSS profile is the summed accessibility signal (defined as 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	-
Fraction of high-quality fragments overlapping TSS	Fraction of high-quality fragments in cell barcodes that overlap transcription start sites (TSS)	>15%	-
Fraction of high-quality fragments overlapping peaks	Fraction of high-quality fragments in cell barcodes that overlap called peaks	>15%	Low percentage indicates that fragments are not coming from called peaks but rather from random regions of the genome. Causes include dead cells, or very low sequencing depth.

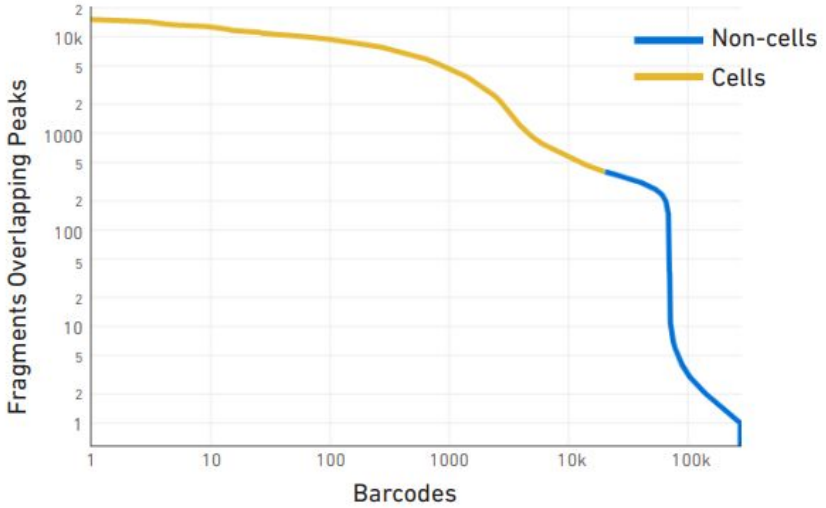
Barcode Rank Plot: The Barcode Rank (or knee plot) for fragments overlapping peaks marks the barcodes that were inferred to be associated with cells.

Example

Ideal Sample: A steep drop-off is indicative of good separation between the cell-associated barcodes and the barcodes associated with empty GEMs.



Compromised Sample: Round curve and lack of steep drop-off may indicate low sample quality or loss of single-cell behavior.



EMBL

Barcode rank plot

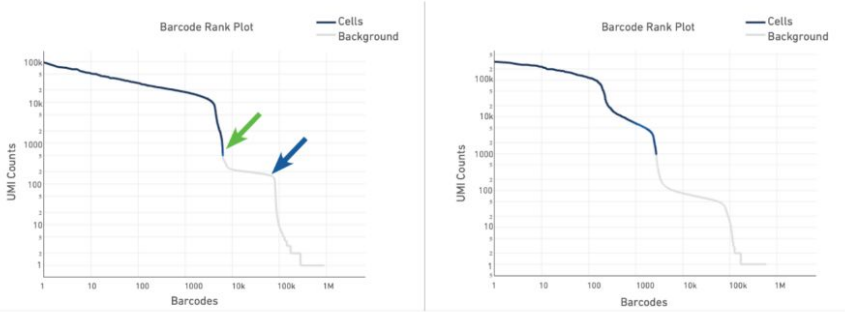
6. Barcode rank plot

The most important and informative plot in the Gene Expression Web Summary is the **Barcode Rank Plot**, which shows the distribution of UMI counts in barcodes. All detected barcodes are plotted in decreasing order of the number of UMIs associated with that particular barcode.

We can take a look at a few examples before looking into the barcode rank plot for our sample using this Technical Note: Interpreting Cell Ranger Web Summary Files for Single Cell Gene Expression Assays. Page 4 of the technical note shows a few examples of good and compromised samples. We will look at the typical sample and the heterogeneous sample here.

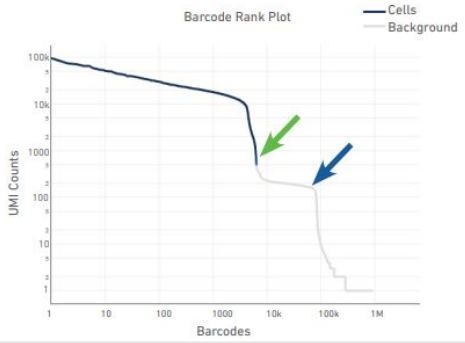
The plot for a typical sample (left in the screenshot below) has a distinctive shape, which is referred to as a "cliff and knee". The blue-to-gray transition (green arrow) is referred to as the cliff; the solid gray (blue arrow) is the knee. The steep cliff is indicative of good separation between the cell-associated barcodes and the barcodes associated with empty GEMs.

Depending on the sample type, heterogeneous populations of cells in a sample may result in a bimodal plot (right in the screenshot below). In these situations, the cell-associated barcodes will have two cliff and knee distributions. However, there should still be clear separation between the barcodes called as cells (blue) and barcodes called as background (gray).

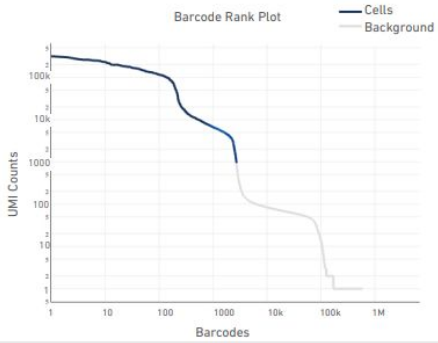


The barcode rank plot for our sample looks similar to the heterogeneous sample in the technical notes. This is consistent with what we expect from our sample, which contains PBMCs (relatively high number of expressed genes) and neutrophils (only expressing around a few hundred genes). However, the transition of blue to gray is lower than expected, which means some of the cells we included could be background (gray). This is fine for now, because we can filter those out in Loupe Browser (Filter out background GEMs).

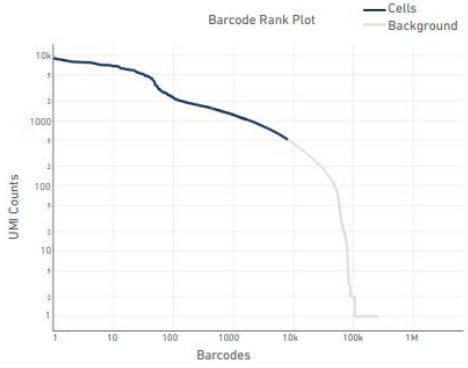
Typical Sample: A steep drop-off is indicative of good separation between the cell-associated barcodes and the barcodes associated with empty GEMs. An ideal Barcode Rank plot has a distinctive shape, which is referred to as a "cliff and knee". The blue-to-gray transition (green arrow) is referred to as the cliff; the solid gray is referred to as the knee (blue arrow).



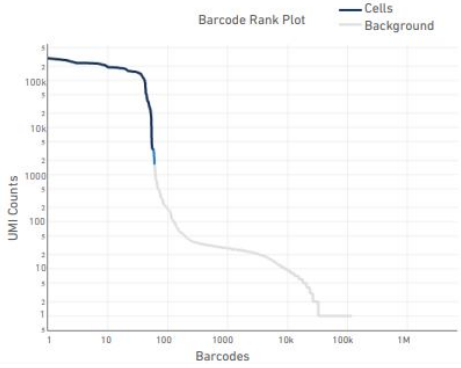
Heterogeneous Sample: Occasionally and based on sample type, there can be heterogeneous populations of cells in a sample that may result in a bimodal plot. In these situations, the cell-associated 10x Barcodes will have two "cliff and knee" distributions. However, there should still be clear separation between the barcodes called as 'cells' and barcodes called as 'background'.



Compromised Sample: Round curve and lack of steep cliff may indicate low sample quality or loss of single-cell behavior. This can be due to a wetting failure, premature cell lysis, or low cell viability.



Compromised Sample: Defined cliff and knee, but the total number of barcodes detected may be lower than expected. This can be caused by a sample clog or inaccurate cell count.

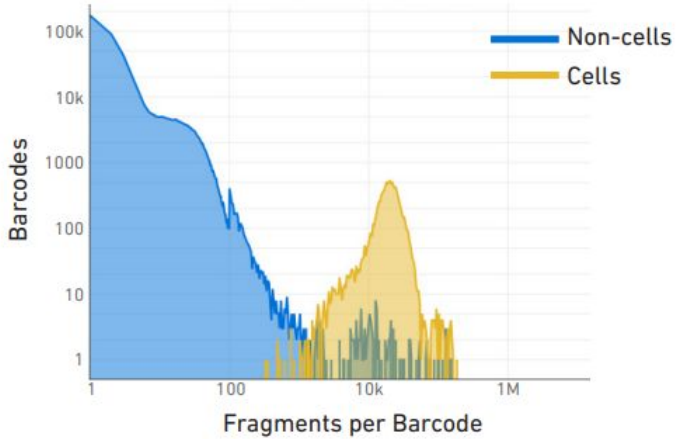


EMBL

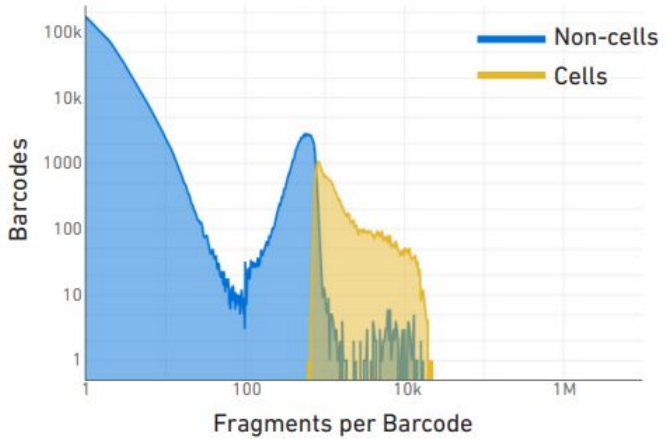
Fragment Distribution Plot: The distribution of the number of fragments per barcode for the non-cell and cell groups is displayed in the Fragment Distribution plot.

Example

Ideal Sample: A good separation between cell and non-cell groups indicate proper distinction between cells and non-cells.



Compromised Sample: Large overlap between cells and non-cells may indicate issues with sample quality.

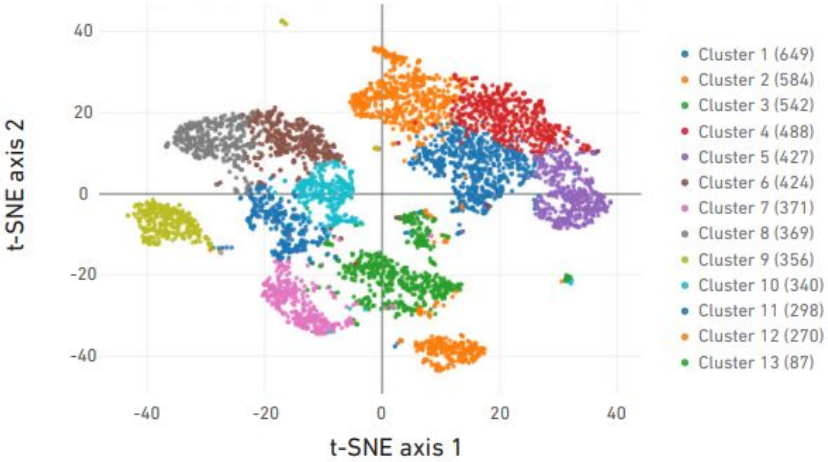


EMBL

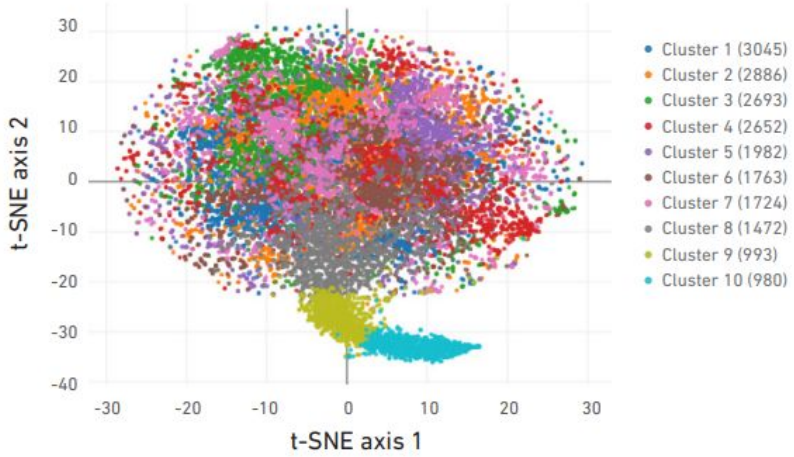
Cell Clustering Scatter Plot: The Cell Clustering (colored by cluster) plot shows the cell-associated barcodes in a 2-D t-SNE projection, with colors showing an automated graph clustering analysis which groups together cells with similar peak profiles.

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.



EMBL

tSNE and UMAP - be aware of not overinterpreting!

How to (mis)read UMAP

While UMAP offers a number of advantages over t-SNE, it's by no means a silver bullet - and reading and understanding its results requires some care. It's worth revisiting our previous work on (mis)reading t-SNE, since many of the same takeaways apply to UMAP:

1. Hyperparameters really matter

Choosing good values isn't easy, and depends on both the data and your goals (eg, how tightly packed the projection ought to be). This is where UMAP's speed is a big advantage - By running UMAP multiple times with a variety of hyperparameters, you can get a better sense of how the projection is affected by its parameters.

2. Cluster sizes in a UMAP plot mean nothing

Just as in t-SNE, the size of clusters relative to each other is essentially meaningless. This is because UMAP uses local notions of distance to construct its high-dimensional graph representation.

3. Distances between clusters might not mean anything

Likewise, the distances between clusters is likely to be meaningless. While it's true that the global positions of clusters are better preserved in UMAP, the distances between them are not meaningful. Again, this is due to using local distances when constructing the graph.

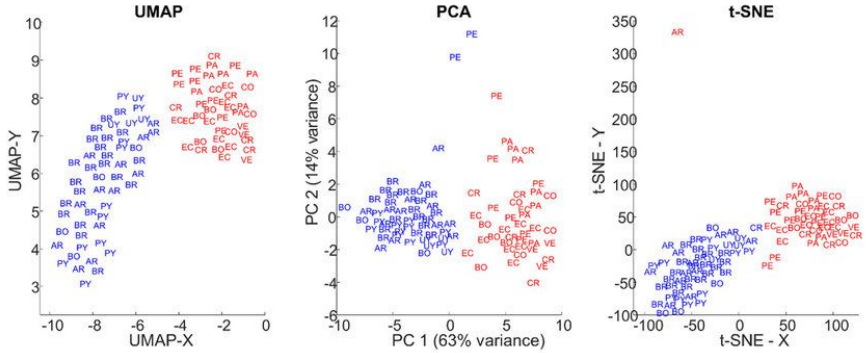
4. Random noise doesn't always look random.

Especially at low values of `n_neighbors`, spurious clustering can be observed.

5. You may need more than one plot

Since the UMAP algorithm is stochastic, different runs with the same hyperparameters can yield different results. Additionally, since the choice of hyperparameters is so important, it can be very useful to run the projection multiple times with various hyperparameters.

https://pair-code.github.io/understanding-umap/

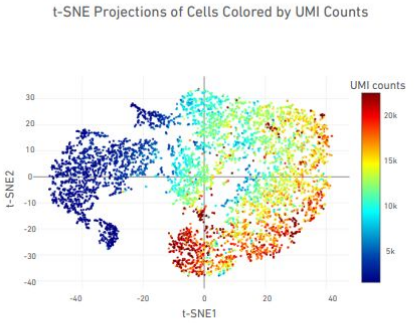


https://www.mdpi.com/2076-2615/10/8/1406

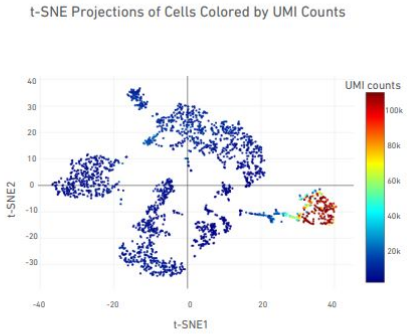
EMBL

t-SNE Projection of Cells Colored by UMI counts / clustering

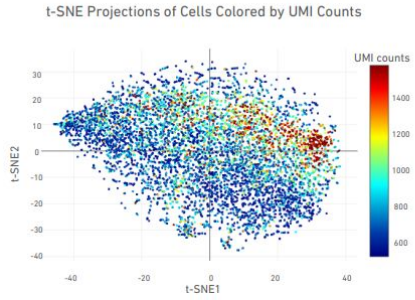
Typical Sample: Structured clusters with clear separation between high UMI and low UMI containing barcodes.



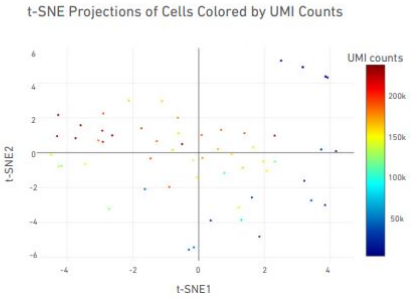
Heterogeneous Sample: Heterogeneous samples that contain high and low RNA containing cells should have visible differences in the UMIs levels between the two populations.



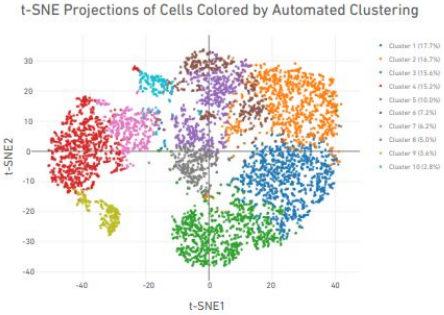
Compromised sample: Lack of cluster structure, one large cluster, or no separation (for a sample with expected heterogeneous cell populations) may indicate sample quality issues or loss of single-cell behavior.



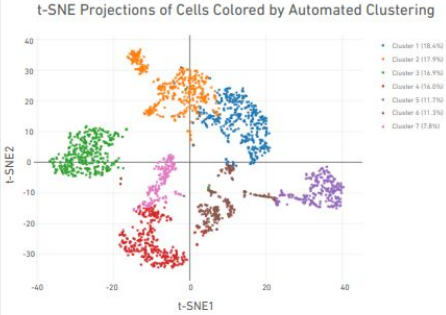
Compromised sample: Very low cell numbers may result in scattered t-SNE plots with little cluster structure.



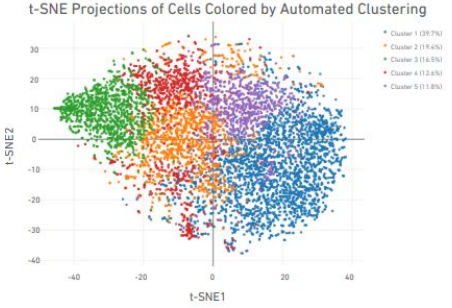
Typical Sample: Structured clusters with good separation (for a sample with expected heterogeneous cell populations).



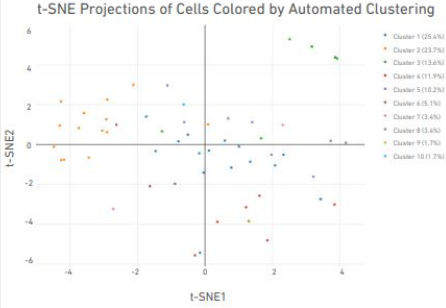
Heterogeneous sample: Structured clusters with good separation (for a sample with expected heterogeneous cell populations).



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.



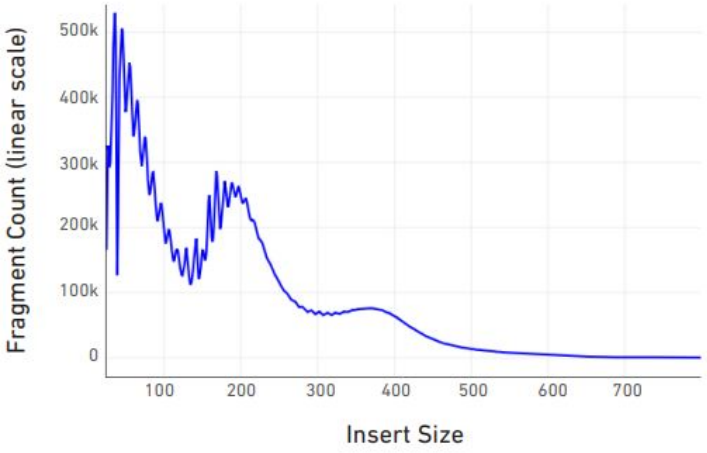
Compromised sample: Very low cell numbers may result in scattered cell t-SNE plots with little cluster structure.



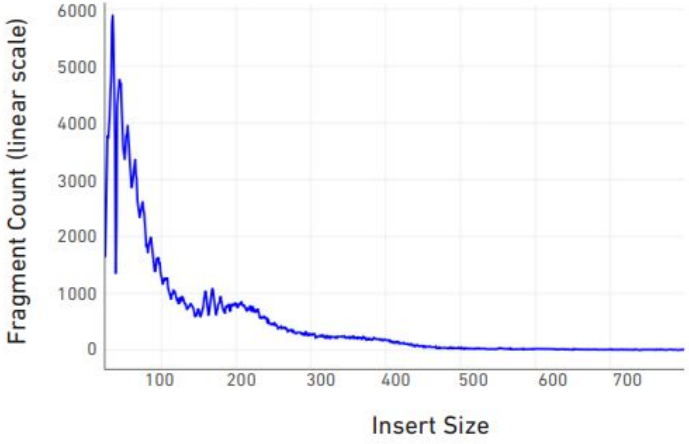
Insert Size Distribution Plot: Insert size distribution of transposase accessible fragments sequenced is displayed in the Insert Size Distribution plot.

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 may indicate loss of chromatin structure due to low sample quality.

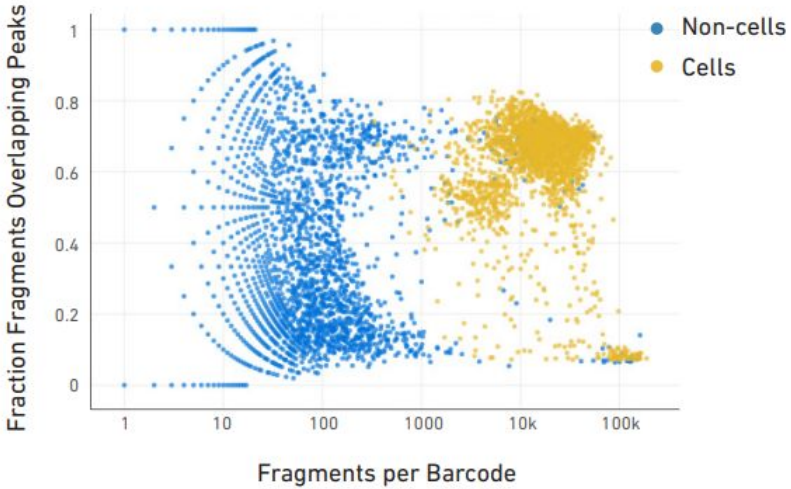


EMBL

Single Cell Targeting Plot: A scatterplot 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. Cell-associated and non-cell associated barcodes tightly concentrated in the same location may indicate issues with cell calling or sample preparation.