

Analysing 10X Single Cell RNA-Seq Data

v2022-06

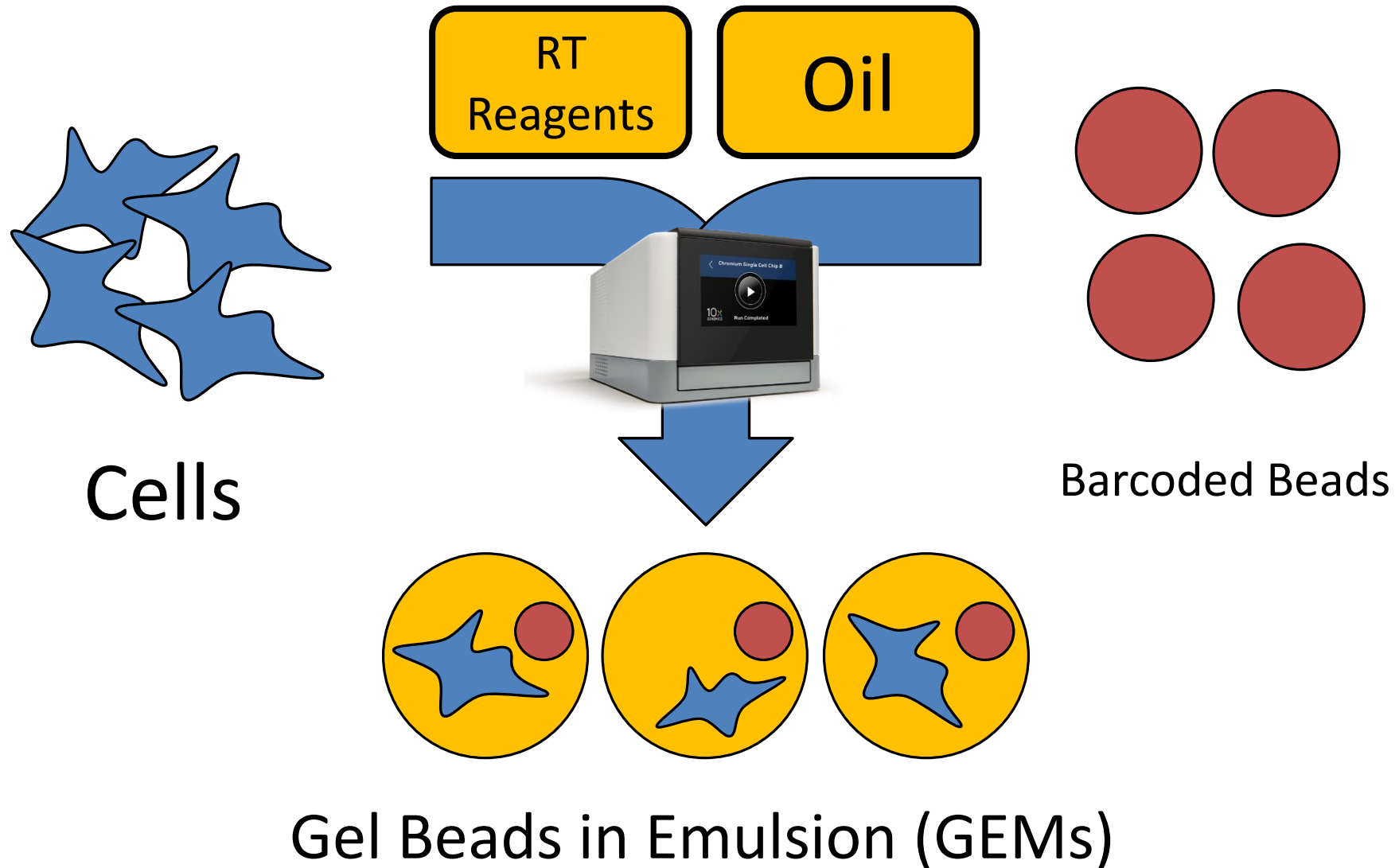
Simon Andrews

simon.andrews@babraham.ac.uk

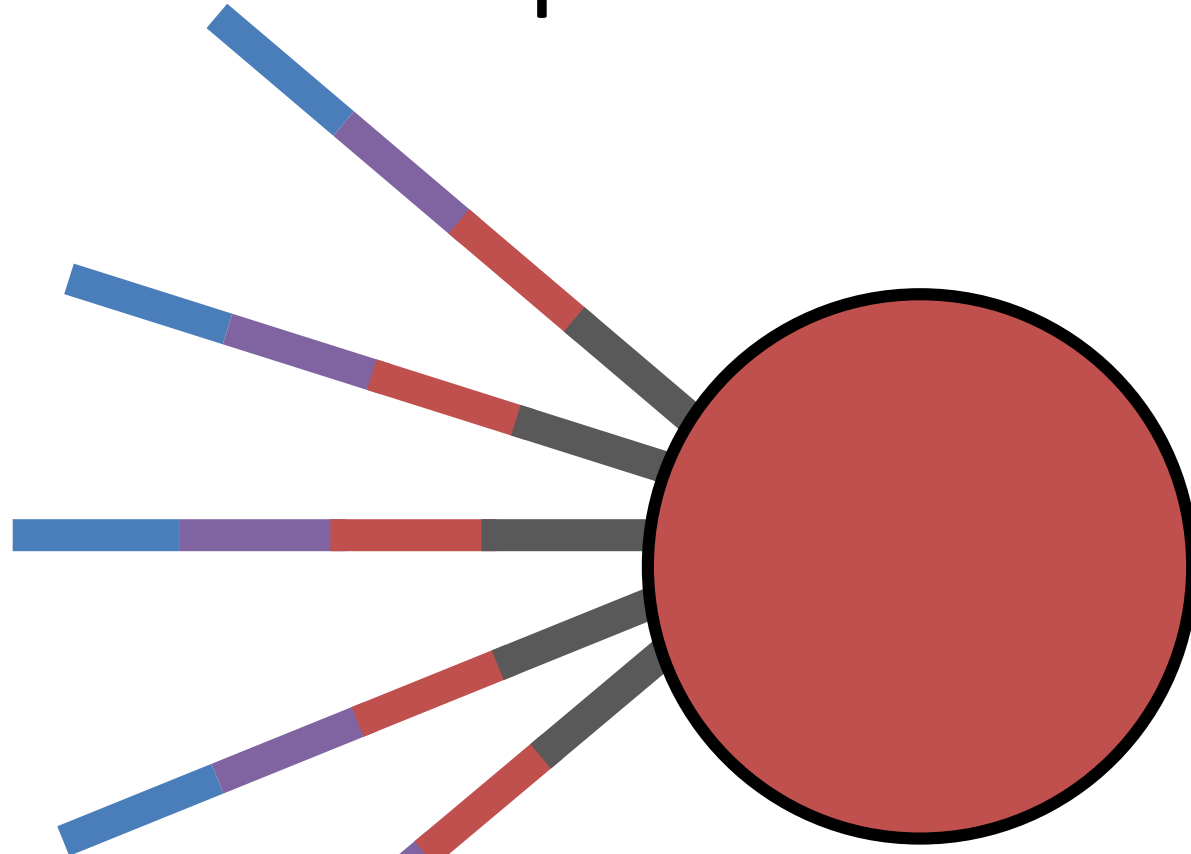
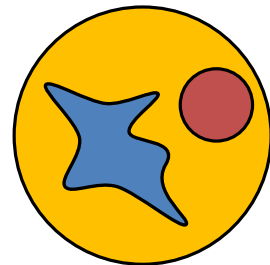
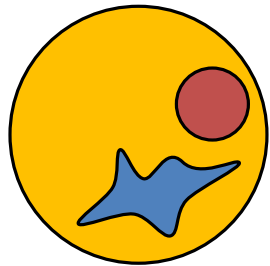
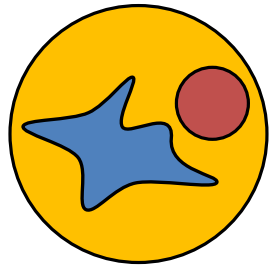
Course Outline

- How 10X single cell RNA-Seq works
- Evaluating CellRanger QC
 - [Exercise] Looking at CellRanger QC reports
- Dimensionality Reduction (PCA, tSNE, UMAP)
 - [Exercise] Using the Loupe cell browser
- R Frameworks for scRNA analysis
 - [Exercise] Analysing data in R using Seurat

How 10X RNA-Seq Works

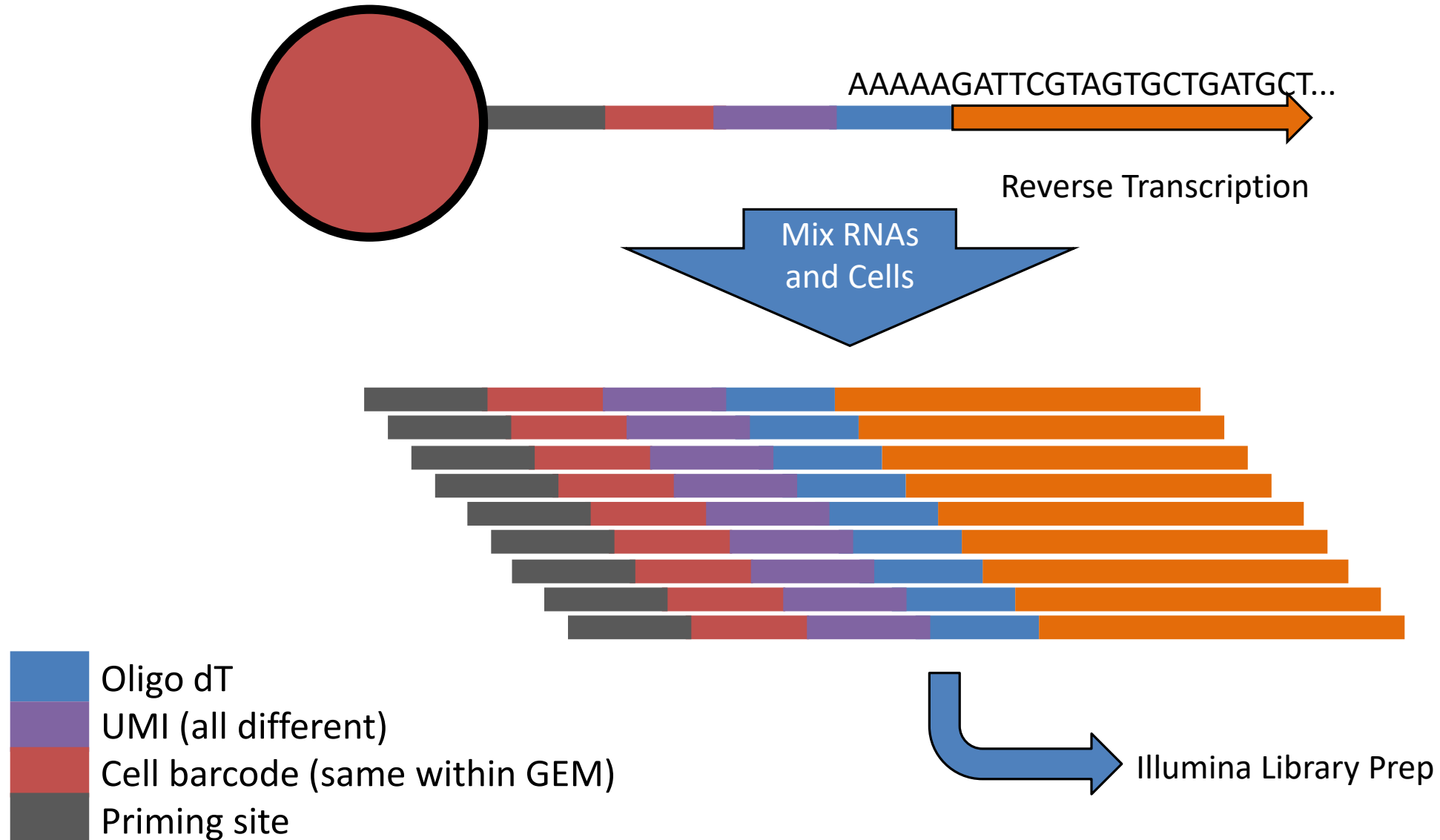


How 10X RNA-Seq Works

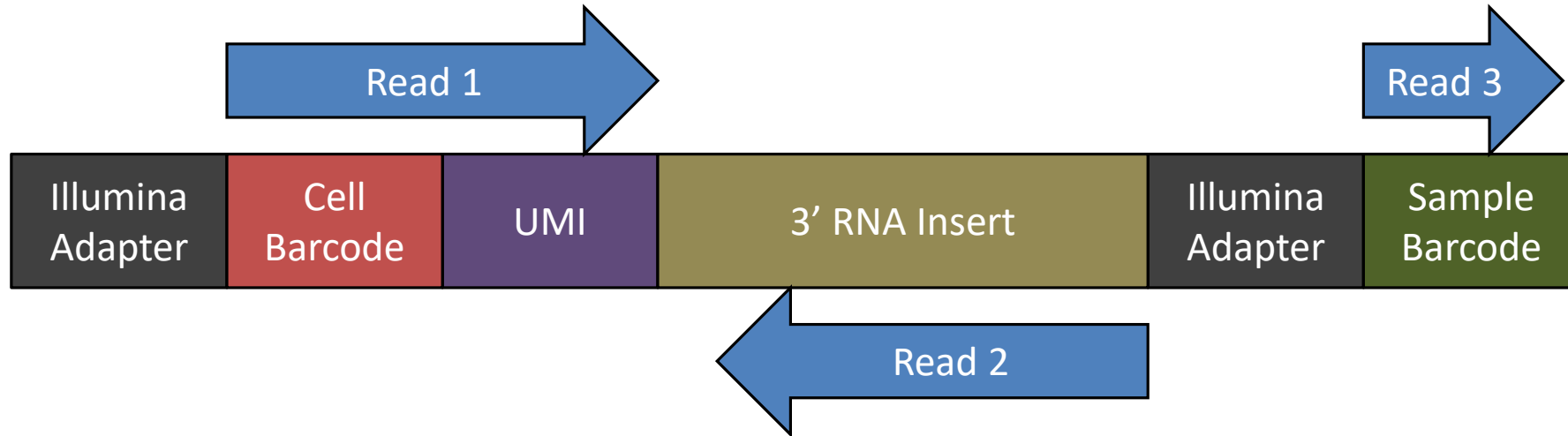





- Blue segment: Oligo dT
- Purple segment: UMI (all different)
- Red segment: Cell barcode (same within GEM)
- Grey segment: Priming site

How 10X RNA-Seq Works

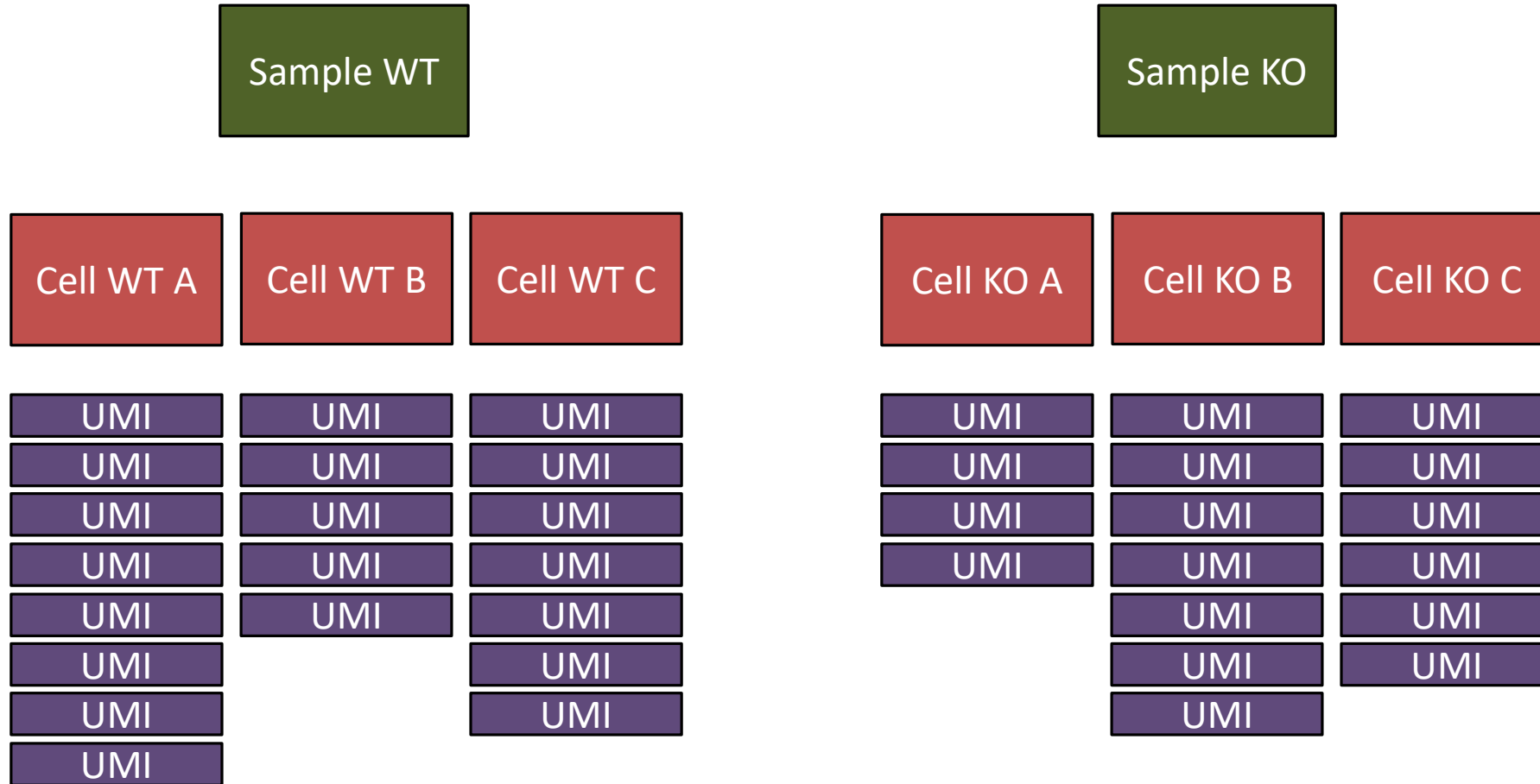


How 10X RNA-Seq Works



-  Sample level barcode – same for all cells and RNAs in a library
-  Cell level barcode (16bp) – same for all RNAs in a cell
-  UMI (10bp) – unique for one RNA in one cell

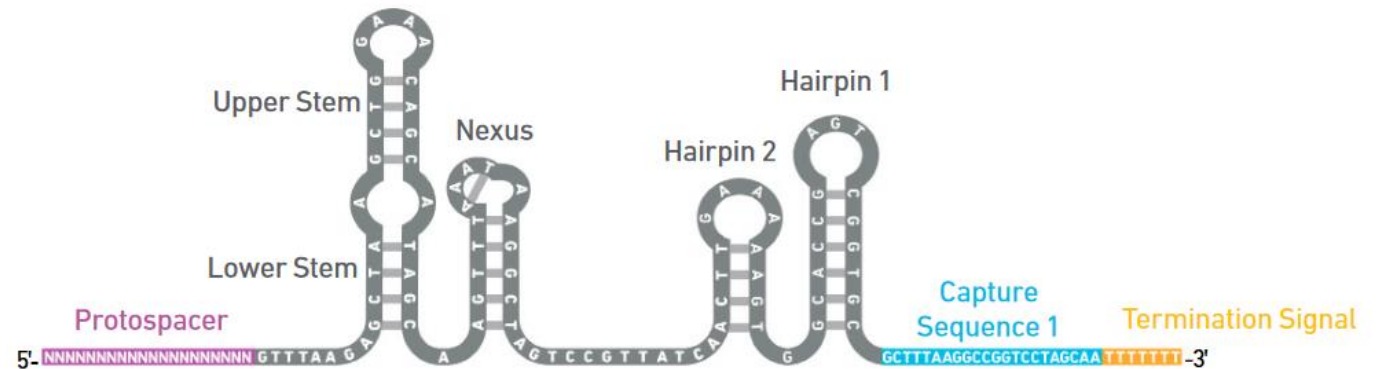
10X Produces Barcode Counts



UMIs are finally related to genes to get per-gene counts

Extension Techniques

- Variants of the basic protocol which allow for other measures
- Introduce artificial sequences which are measured alongside the normal RNAs
 - Cell Surface Markers
 - CRISPR guide RNAs



- Beads use custom captures (in addition to TTTT)
- Attach sequences to sgRNA or tag to antibodies

The 10X Software Suite

Chromium
Controller

Runs the chromium
system for creating
GEMs

Cell
Ranger

Pipeline for
mapping, filtering,
QC and quantitation
of libraries

Loupe
Browser

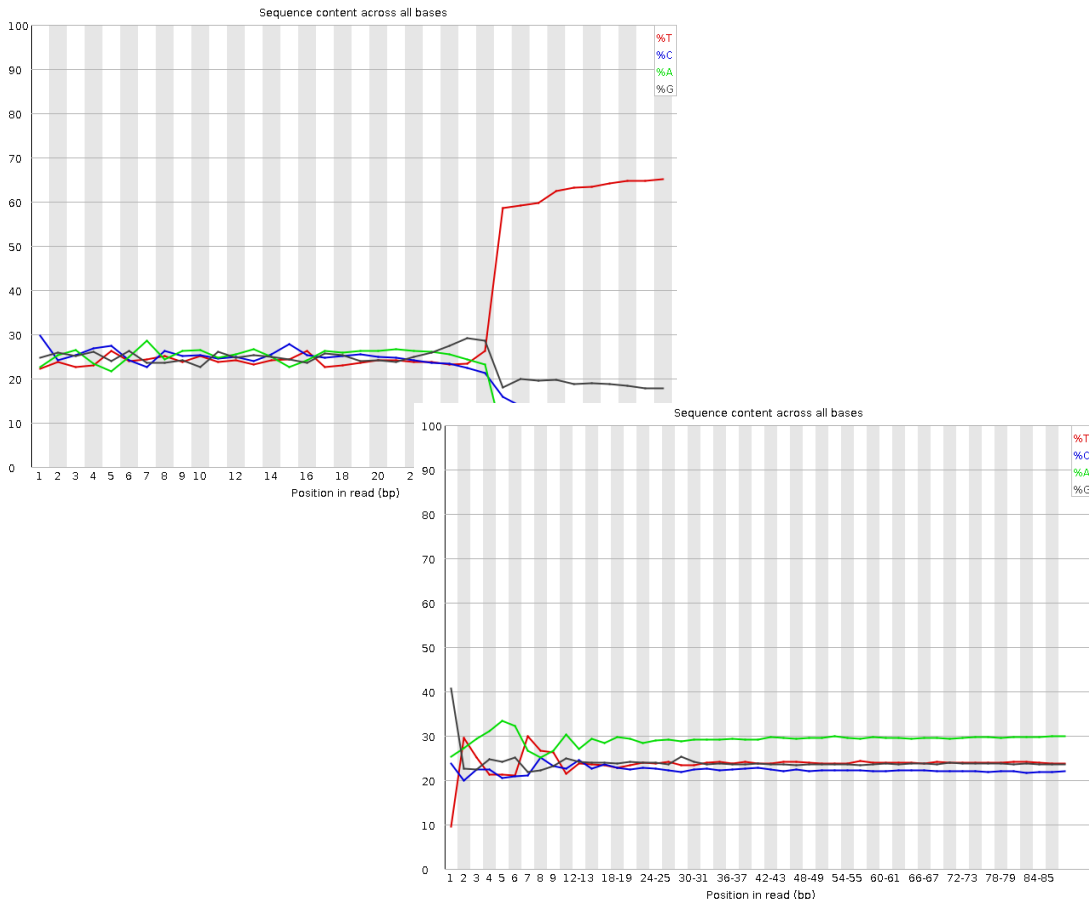
Desktop software for
visualisation and
analysis of single cell
data.

Cell Ranger

- Barcode Extraction and filtering
 - Identifies cell level barcodes
- Mapping to reference
 - Uses STAR aligner
- Generate count table
 - UMIs per gene in each cell
- Dimensionality Reduction
 - PCA and tSNE
- Clustering
 - K-means and Graph Based

CellRanger Commands

```
scrALI001_S1_L001_I1_001.fastq.gz  
scrALI001_S1_L001_R1_001.fastq.gz  
scrALI001_S1_L001_R2_001.fastq.gz
```



- I1
 - Index file. Sets of 4 barcodes per sample
- R1
 - Barcode reads
 - 16bp cell level barcode
 - 10bp UMI
- R2
 - 3' RNA-seq read

CellRanger Commands

- **CellRanger Count (quantitates a single run)**

```
$ cellranger count --id=COURSE \  
                  --transcriptome=/bi/apps/cellranger/references/GRCh38/ \  
                  --fastqs=/bi/home/andrewss/10X/ \  
                  --localcores=8 \  
                  --localmem=32
```

- **CellRanger aggr (merges multiple runs)**

```
$ cellranger aggr --id=MERGED \  
                 --csv=merge_me.csv \  
                 --normalize=mapped
```

CellRanger Aggregate CSV file

Required		Optional	
library_id	molecule_h5	sex	genotype
WT1	/data/WT1/outs/molecule_info.h5	Male	WT
WT2	/data/WT2/outs/molecule_info.h5	Female	WT
WT3	/data/WT3/outs/molecule_info.h5	Male	WT
WT4	/data/WT4/outs/molecule_info.h5	Female	WT
KO1	/data/KO1/outs/molecule_info.h5	Male	KO
KO2	/data/KO2/outs/molecule_info.h5	Female	KO
KO3	/data/KO3/outs/molecule_info.h5	Male	KO
KO4	/data/KO4/outs/molecule_info.h5	Female	KO

Output files generated

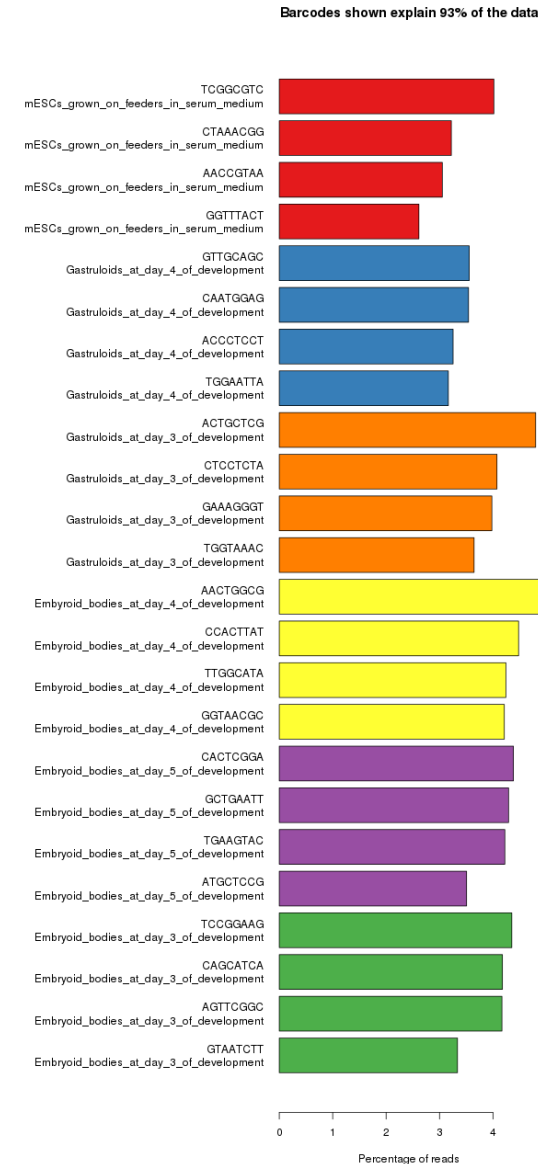
- `web_summary.html` - Web format QC report
- `filtered_feature_bc_matrix.h5` - Single file of cell counts
- `filtered_features_bc_matrix`
 - `barcodes.tsv.gz` - cell level barcodes seen in this sample
 - `features.tsv.gz` - list of quantitated features (usually Ensembl genes)
 - `matrix.mtx.gz` - (sparse) matrix of counts for cells and features
- `possorted_genome_bam.bam` - BAM file of mapped reads
- `molecule_info.h5` - Details of the cell barcodes – used for merging, can also use for analysis
- `cloupe.cloupe` - Analysis data for Loupe Cell browser

Evaluating CellRanger Output

- Look at barcode splitting report
 - Check sample level barcodes
- Look at `web_summary.html` file
 - Check number of cells
 - Check quality of data
 - Check coverage per cell
 - Check library diversity

Sample Level Barcodes

- Only present if multiple libraries mixed in a lane
- Get standard barcode split report, but with 4 barcodes used per sample
- Even coverage within and between libraries



CellRanger Reports

- HTML report – comes with each sample and aggregated group of samples
- Gives some basic metrics to judge the quality of the samples and spot any issues in the data or processing

COURSE

Alerts

The analysis detected ⓘ 1 informational notice.

Alert	Value	Detail
ⓘ Intron mode used		This data has been analyzed with intronic reads in Ranger versions. If you would not like to count intronic reads, please contact support@10xgenomics.com for an analysis rerun.

Summary

Gene Expression

5,201

Estimated Number of Cells

48,978

Mean Reads per Cell

1,660

Median Genes per Cell

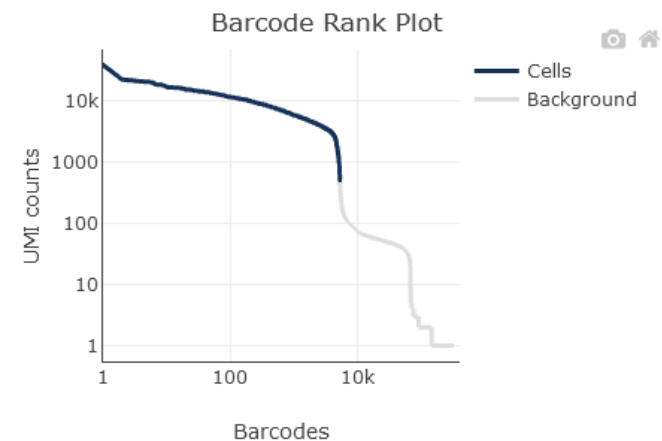
Sequencing ⓘ

Number of Reads	254,736,630
Number of Short Reads Skipped	0
Valid Barcodes	98.3%
Valid UMIs	100.0%
Sequencing Saturation	72.6%
Q30 Bases in Barcode	97.6%
Q30 Bases in RNA Read	83.3%
Q30 Bases in UMI	97.5%

Mapping ⓘ

Reads Mapped to Genome	47.7%
Reads Mapped Confidently to Genome	46.5%
Reads Mapped Confidently to Intergenic Regions	1.7%
Reads Mapped Confidently to Intronic Regions	14.2%
Reads Mapped Confidently to Exonic Regions	30.6%
Reads Mapped Confidently to Transcriptome	38.9%
Reads Mapped Antisense to Gene	5.5%

Cells ⓘ



Estimated Number of Cells	5,201
Fraction Reads in Cells	92.6%
Mean Reads per Cell	48,978
Median UMI Counts per Cell	3,901
Median Genes per Cell	1,660
Total Genes Detected	25,810


Sample

Sample ID	COURSE
Sample Description	
Chemistry	Single Cell 3' v2
Include introns	True
Reference Path	...r/references/refdata-gex-GRCh38-2020-A
Transcriptome	GRCh38-2020-A
Pipeline Version	cellranger-7.0.0

Errors and Warnings



The analysis detected some serious issues with your sequencing run. [Details »](#)

The analysis detected some issues with your sequencing run. [Details »](#)

Alert	Value	Detail
 Low Fraction Reads Confidently Mapped To Transcriptome	51.5%	Ideal > 60%. This can indicate use of the wrong reference transcriptome, poor library quality, or poor sequencing quality. Application performance may be affected.

Alerts

The analysis detected  2 errors.

Alert	Value	Detail
 Low Fraction Reads Confidently Mapped To Transcriptome	19.6%	Ideal > 30%. This 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.
 Low Fraction Reads in Cells	48.8%	Ideal > 70%. Application performance may be affected. Many of the reads were not assigned to cell-associated barcodes. This could be caused by high levels of ambient RNA or by a significant population of cells with a low RNA content, which the algorithm did not call as cells. The latter case can be addressed by inspecting the data to determine the appropriate cell count and using <code>--force-cells</code> .

How many cells do you have?

- Cell number is determined from the number of cell barcodes with 'reasonable' numbers of observations
- Need to separate signal from background – real cell associated barcodes vs noise from empty GEMs and mis-called sequences
- Changing the thresholds used can give very different predictions for cell numbers

How many cells do you have?

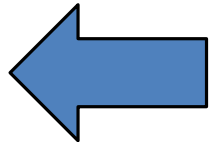
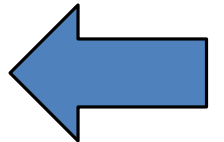
- Start by looking at the quality of the base calls in the barcodes
- Bad calls will lead to inaccurate cell assignments

5,201

Estimated Number of Cells

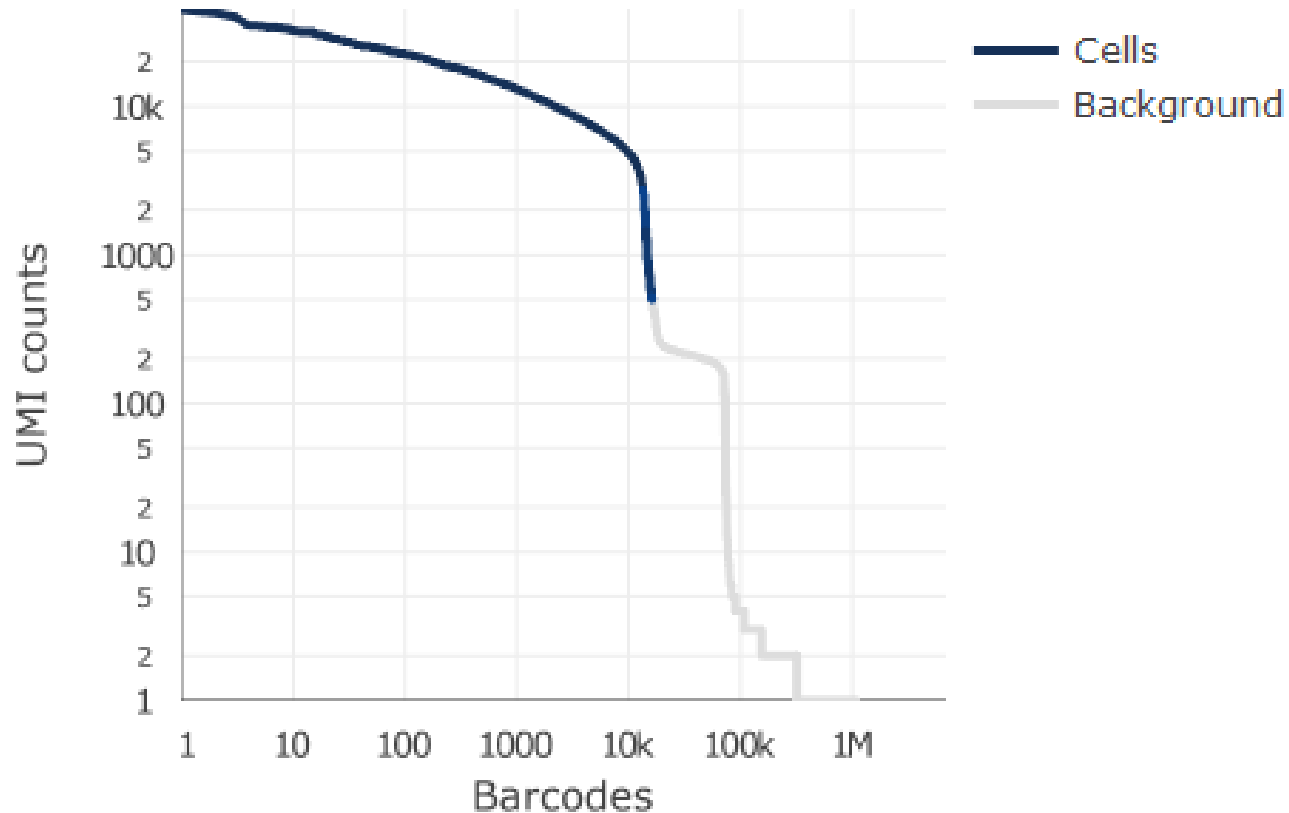
Sequencing ?

Number of Reads	254,736,630
Number of Short Reads Skipped	0
Valid Barcodes	98.3%
Valid UMIs	100.0%
Sequencing Saturation	72.6%
Q30 Bases in Barcode	97.6%
Q30 Bases in RNA Read	83.3%
Q30 Bases in UMI	97.5%



How many cells do you have

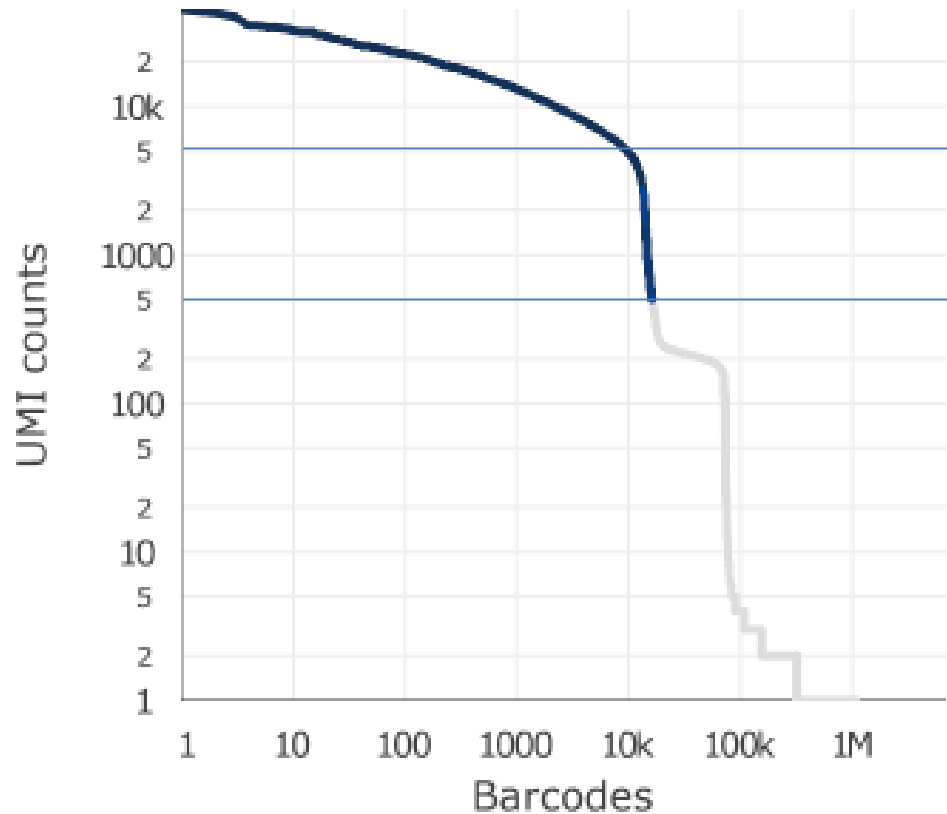
Cells



- Plot of UMIs (reads) per cell vs number of cells
- Blue region was called as valid cells
- Grey region is considered noise
- Both axes are log scale!!!

How many cells do you have

Cells



5000 reads per cell. 10k cells

500 reads per cell. 15k cells

CellRanger v3 uses a liberal cutoff to define cells. This was designed to accommodate (normally cancer) samples where cells might have wildly different amounts of RNA. It will include large numbers of cells with small numbers of UMIs. If this doesn't apply to your sample then this will over-predict valid cells.

How much data do you have per cell?

Mean Reads per Cell

11,380

Median Genes per Cell

2,174

Mapping ?

Reads Mapped to Genome	95.4%
Reads Mapped Confidently to Genome	90.2%
Reads Mapped Confidently to Intergenic Regions	3.0%
Reads Mapped Confidently to Intronic Regions	12.8%
Reads Mapped Confidently to Exonic Regions	74.4%
Reads Mapped Confidently to Transcriptome	71.9%
Reads Mapped Antisense to Gene	0.9%

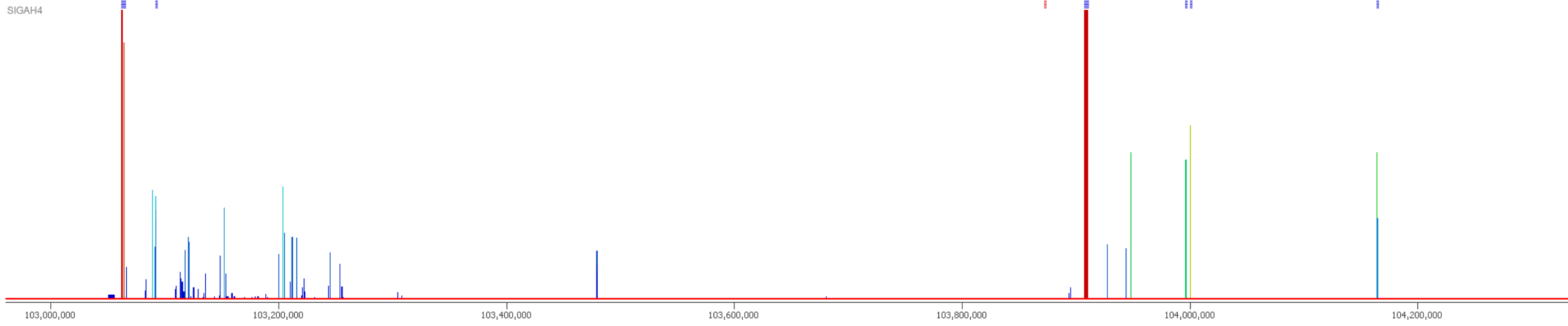
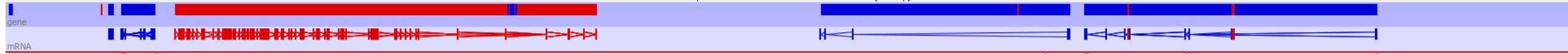
Estimated Number of Cells	15,894
Fraction Reads in Cells	88.1%
Mean Reads per Cell	11,380
Median Genes per Cell	2,174
Total Genes Detected	20,185
Median UMI Counts per Cell	5,742

- Reads should map well
- Check reads are mostly in transcripts
- Means and medians can be misleading when cells are variable
- Note difference between read and UMI

How much data do you have per cell?

- Some details about mapping
 - Reads should map to the 3' end of transcripts (oligo dT selection)
 - Reads count as exonic if 50% of them overlaps an exon
 - Multi-mapped reads which only hit one exon are considered to be uniquely mapped
 - Reads associate with genes based on overlap and direction
 - Only confident (unique) transcriptome reads are used for analysis. As of CellRanger v7 intronic reads as well as exonic are counted by default

Homo sapiens GRCh38 chr 11:102960104-104331592 (1.3 Mbp)



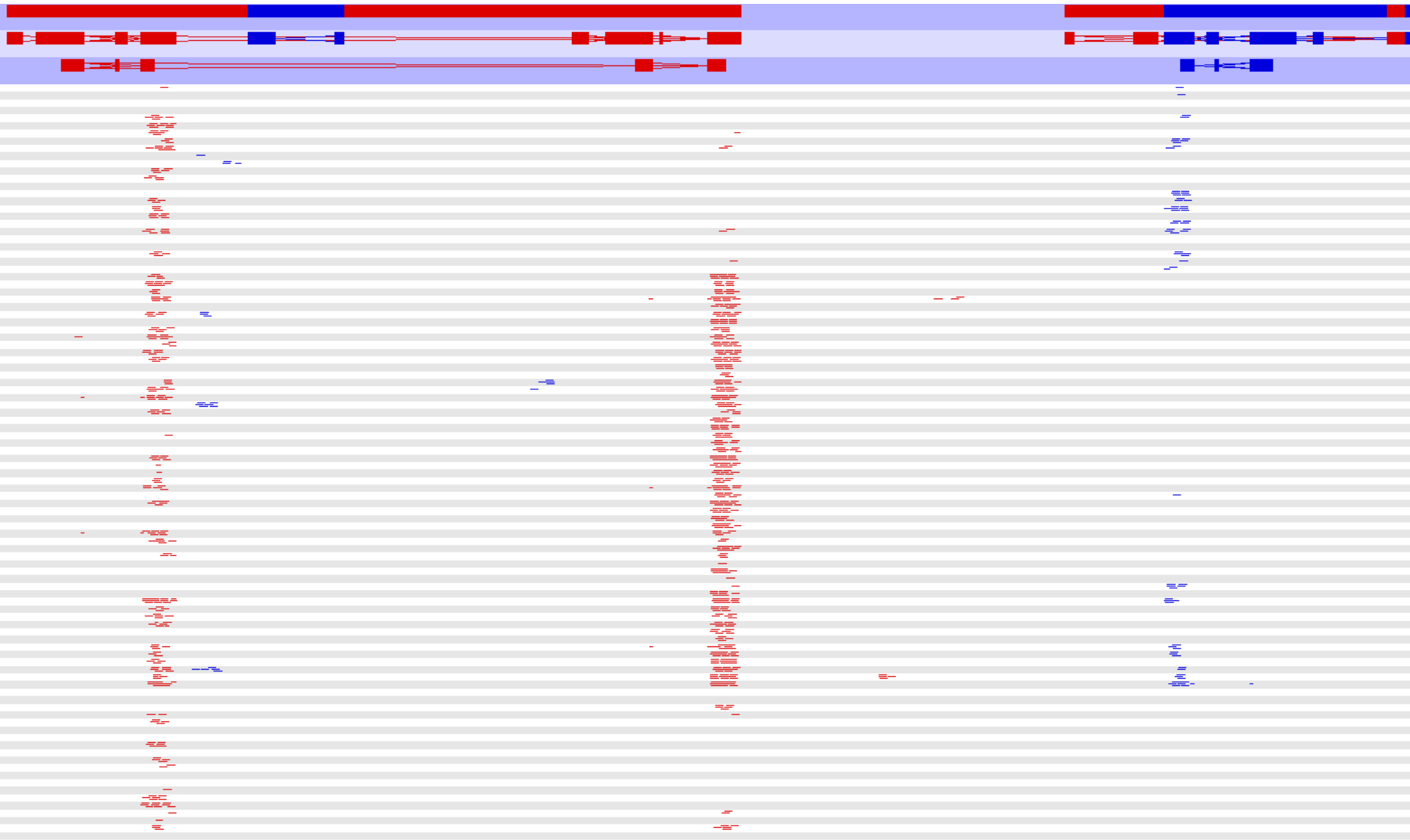
103,000,000 103,200,000 103,400,000 103,600,000 103,800,000 104,000,000 104,200,000

gene

mRNA

CDS

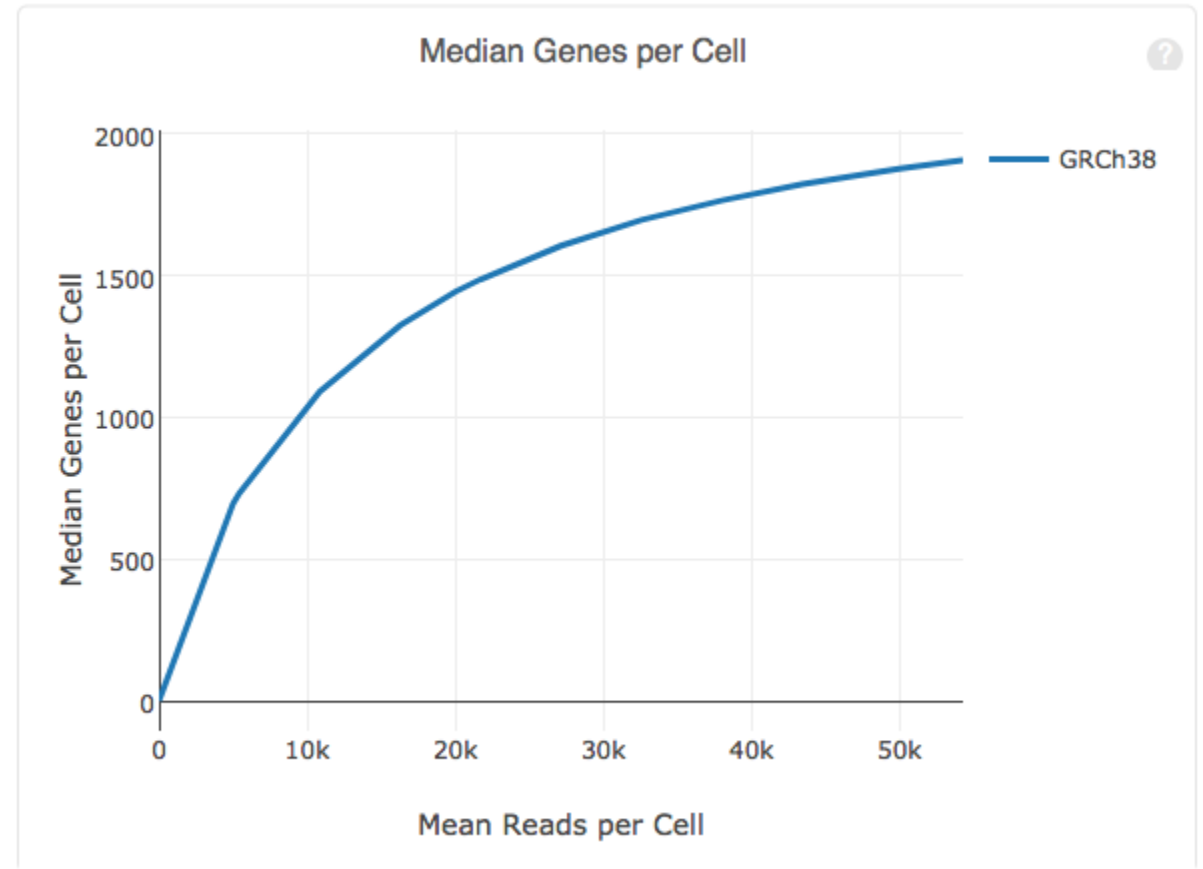
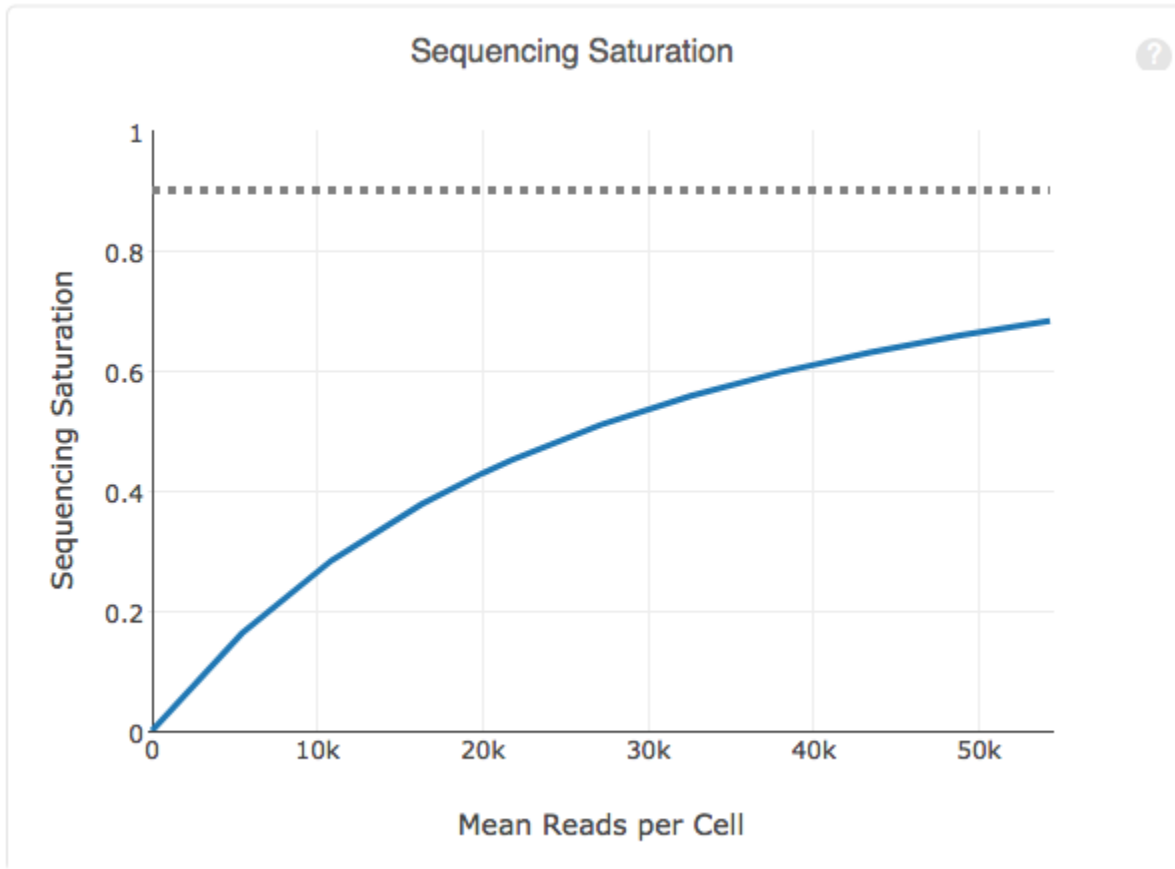
cell_02.sam
cell_06.sam
cell_07.sam
cell_08.sam
cell_09.sam
cell_10.sam
cell_11.sam
cell_12.sam
cell_13.sam
cell_14.sam
cell_15.sam
cell_16.sam
cell_17.sam
cell_18.sam
cell_19.sam
cell_20.sam
cell_21.sam
cell_22.sam
cell_23.sam
cell_24.sam
cell_25.sam
cell_26.sam
cell_27.sam
cell_28.sam
cell_29.sam
cell_30.sam
cell_31.sam
cell_32.sam
cell_33.sam
cell_34.sam
cell_35.sam
cell_36.sam
cell_37.sam
cell_38.sam
cell_39.sam
cell_40.sam
cell_41.sam
cell_42.sam
cell_43.sam
cell_44.sam
cell_45.sam
cell_46.sam
cell_47.sam
cell_48.sam
cell_49.sam
cell_50.sam
cell_51.sam
cell_52.sam
cell_53.sam
cell_54.sam
cell_55.sam
cell_56.sam
cell_57.sam
cell_58.sam
cell_59.sam
cell_60.sam
cell_61.sam
cell_62.sam
cell_63.sam
cell_64.sam
cell_65.sam
cell_66.sam
cell_67.sam
cell_68.sam
cell_69.sam
cell_70.sam
cell_71.sam
cell_72.sam
cell_73.sam
cell_74.sam
cell_75.sam
cell_76.sam
cell_77.sam
cell_78.sam
cell_79.sam
cell_80.sam
cell_81.sam
cell_82.sam
cell_83.sam
cell_84.sam
cell_85.sam
cell_86.sam
cell_87.sam
cell_88.sam
cell_89.sam
cell_90.sam
cell_91.sam
cell_92.sam
cell_93.sam
cell_94.sam
cell_95.sam
cell_96.sam
cell_97.sam
cell_98.sam
cell_99.sam
cell_100.sam
cell_101.sam
cell_102.sam
cell_103.sam
cell_104.sam
cell_105.sam
cell_106.sam
cell_107.sam
cell_108.sam
cell_109.sam
cell_110.sam
cell_111.sam
cell_112.sam
cell_113.sam
cell_114.sam
cell_115.sam
cell_116.sam
cell_117.sam
cell_118.sam
cell_119.sam
cell_120.sam
cell_121.sam
cell_122.sam
cell_123.sam
cell_124.sam
cell_125.sam
cell_126.sam
cell_127.sam
cell_128.sam
cell_129.sam
cell_130.sam
cell_131.sam
cell_132.sam
cell_133.sam
cell_134.sam
cell_135.sam
cell_136.sam
cell_137.sam
cell_138.sam
cell_139.sam
cell_140.sam
cell_141.sam
cell_142.sam
cell_143.sam
cell_144.sam
cell_145.sam
cell_146.sam
cell_147.sam
cell_148.sam
cell_149.sam
cell_150.sam



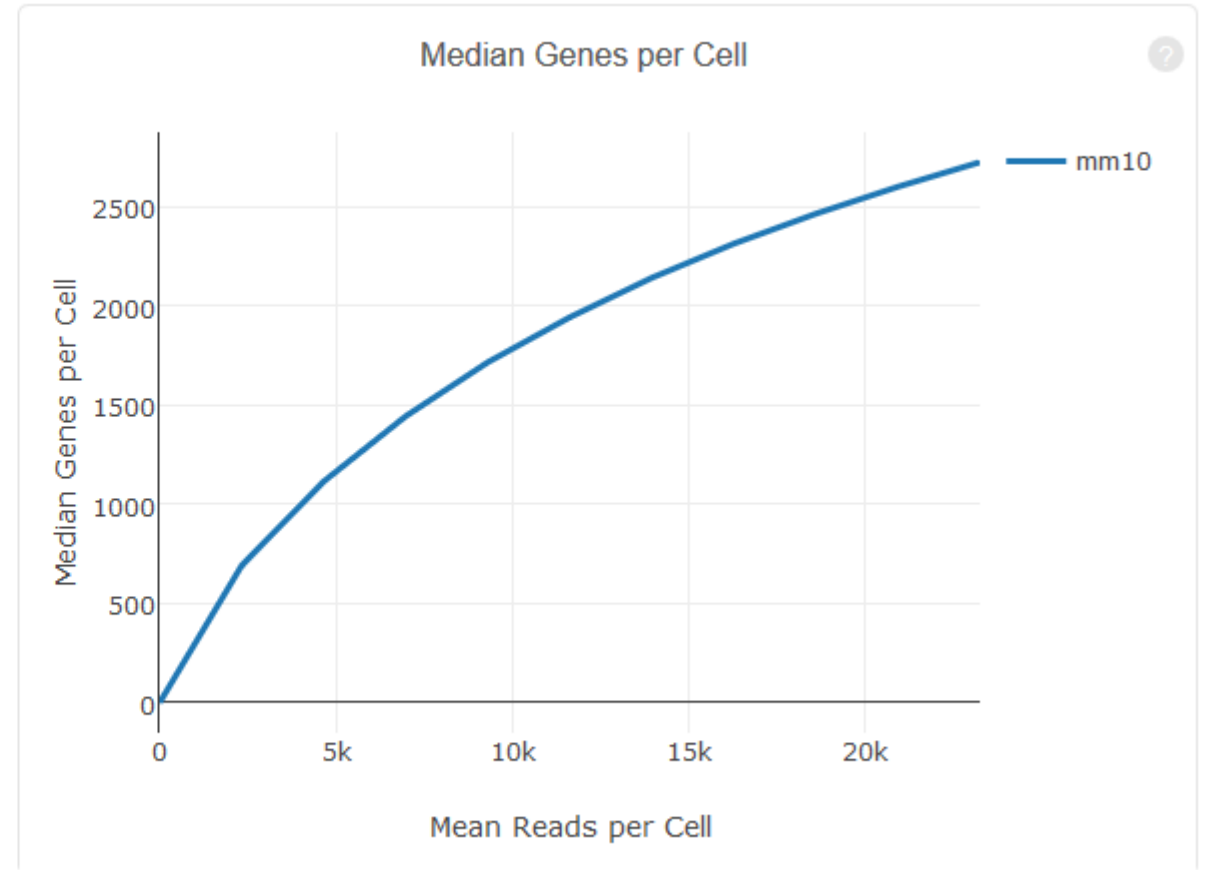
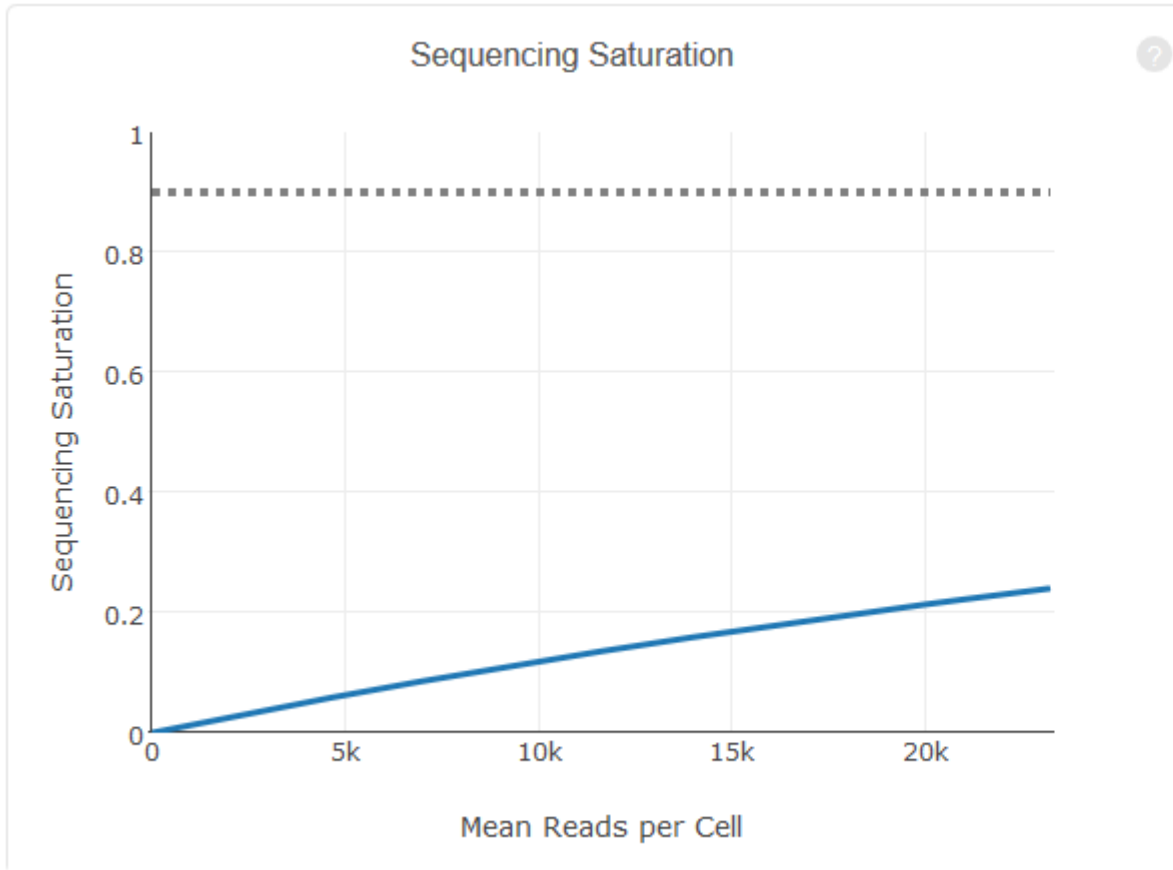
How much data do you have per cell?

- Difficult to generalise how much data to create/expect
 - Depends on cell type, genome and other factors
- In general though, sensible numbers would be:
 - Reads per cell ~10,000
 - Genes per cell 2000 – 3000
- Be aware of the difference between reads (raw) and UMIs (deduplicated) – they can be **very** different

How deeply sequenced is your library

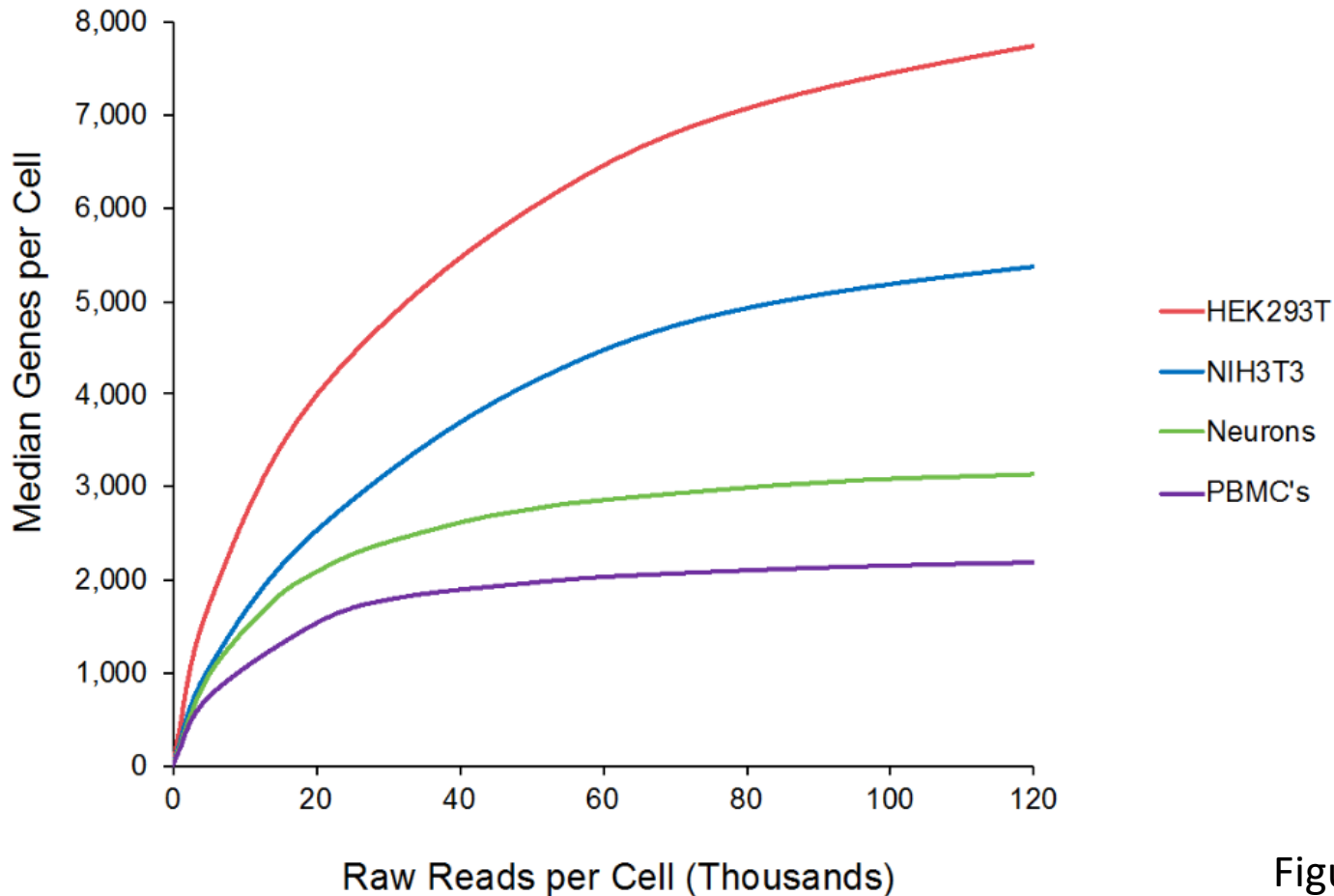


How deeply sequenced is your library

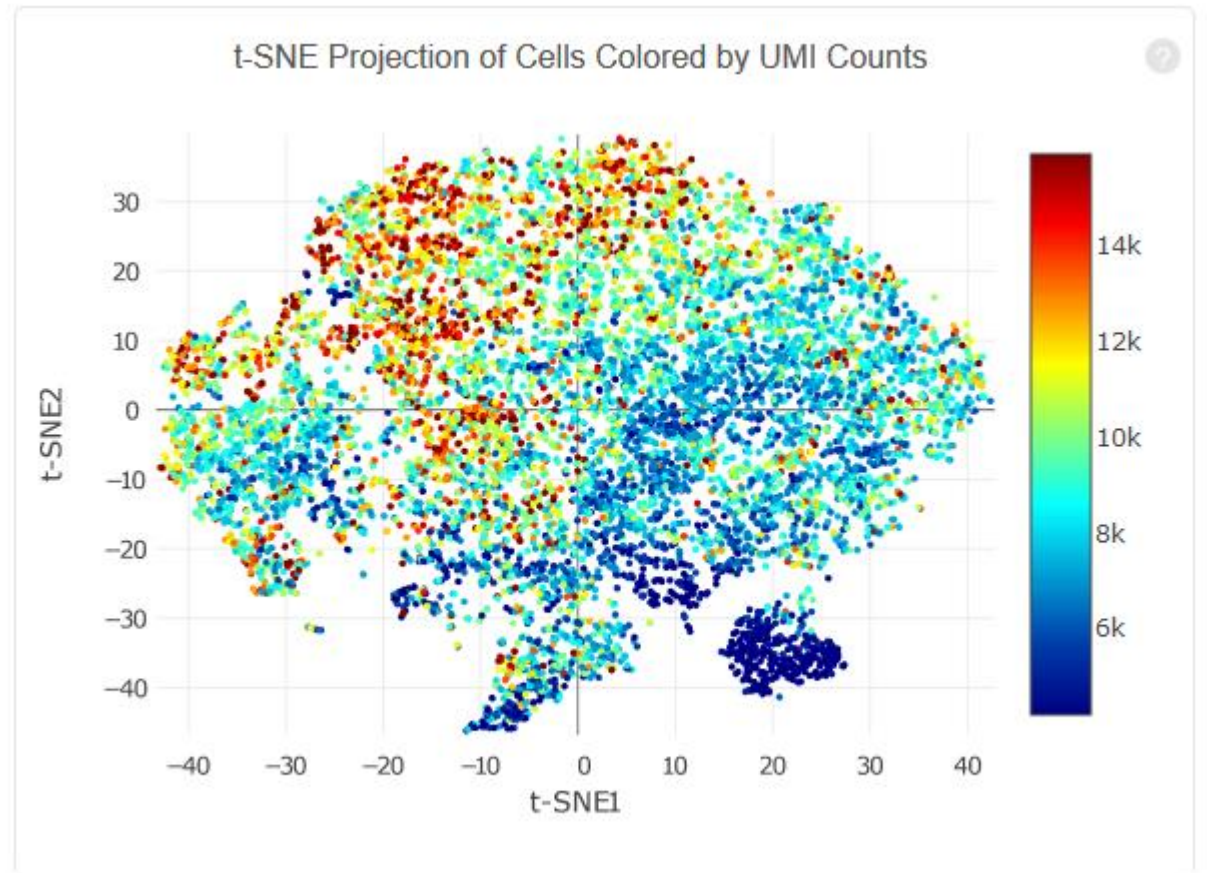
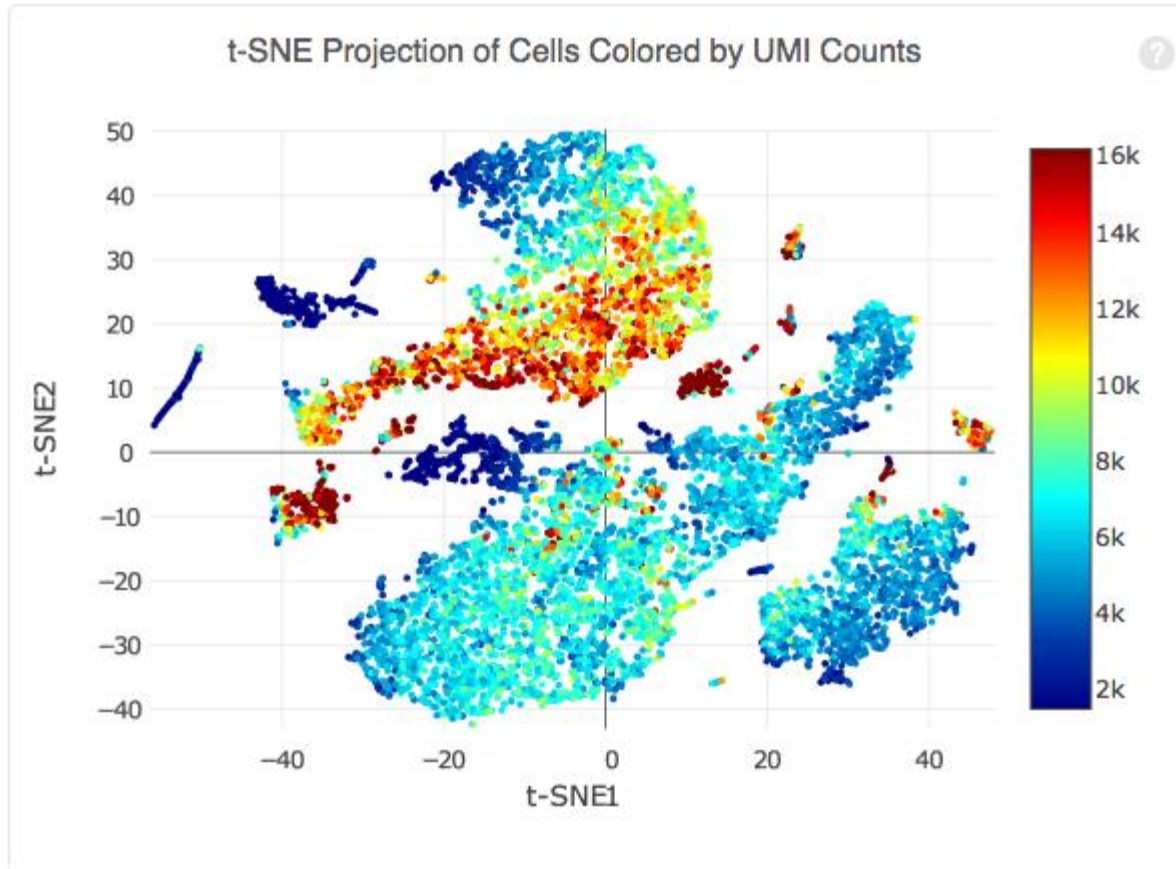


How deeply sequenced is your library

- Expected diversity varies by cell type



Is coverage variation affecting your data?




Aggregation QC

- Web Summary with added “Aggregation” section
- Aggregation ‘normalisation’ is done via sub-sampling to get even coverage
- Can be problematic if libraries are of different sizes (especially if one is really small)

Alerts

The analysis detected  1 warning.

Alert	Value	Detail
 Low Post-Normalization Read Depth	47.2%	Ideal > 50%. There may be large differences in sequencing depth across the input libraries. Application performance may be affected.

Aggregation QC

Aggregation ?

Pre-Normalization Total Number of Reads	3,430,270,725
Post-Normalization Total Number of Reads	2,502,681,800
Pre-Normalization Mean Reads per Cell	75,074
Post-Normalization Mean Reads per Cell	54,773
Fraction of Reads Kept (Influenza_day1)	100.0%
Fraction of Reads Kept (Influenza_day3)	95.2%
Fraction of Reads Kept (Influenza_day6)	72.9%
Fraction of Reads Kept (Influenza_mock)	47.2%


Pre-Normalization Total Reads per Cell (Influenza_day1)	51,029
Pre-Normalization Total Reads per Cell (Influenza_day3)	50,856
Pre-Normalization Total Reads per Cell (Influenza_day6)	84,665
Pre-Normalization Total Reads per Cell (Influenza_mock)	128,146

Exercise – Evaluating CellRanger Reports

- Look at the selection of CellRanger reports to get an idea for the metrics they provide
 - Is the quality of the data good
 - How many cells are there
 - How much data per cell is there (both UMIs and Genes)
 - Is there any separation? Is it driven by amount of data?
- The data we're going to use for the rest of the day is in "course_web_summary.html", do you see any problems which would concern us with this data at this stage?

Course Data CellRanger QC

The analysis detected some issues. [Details »](#)

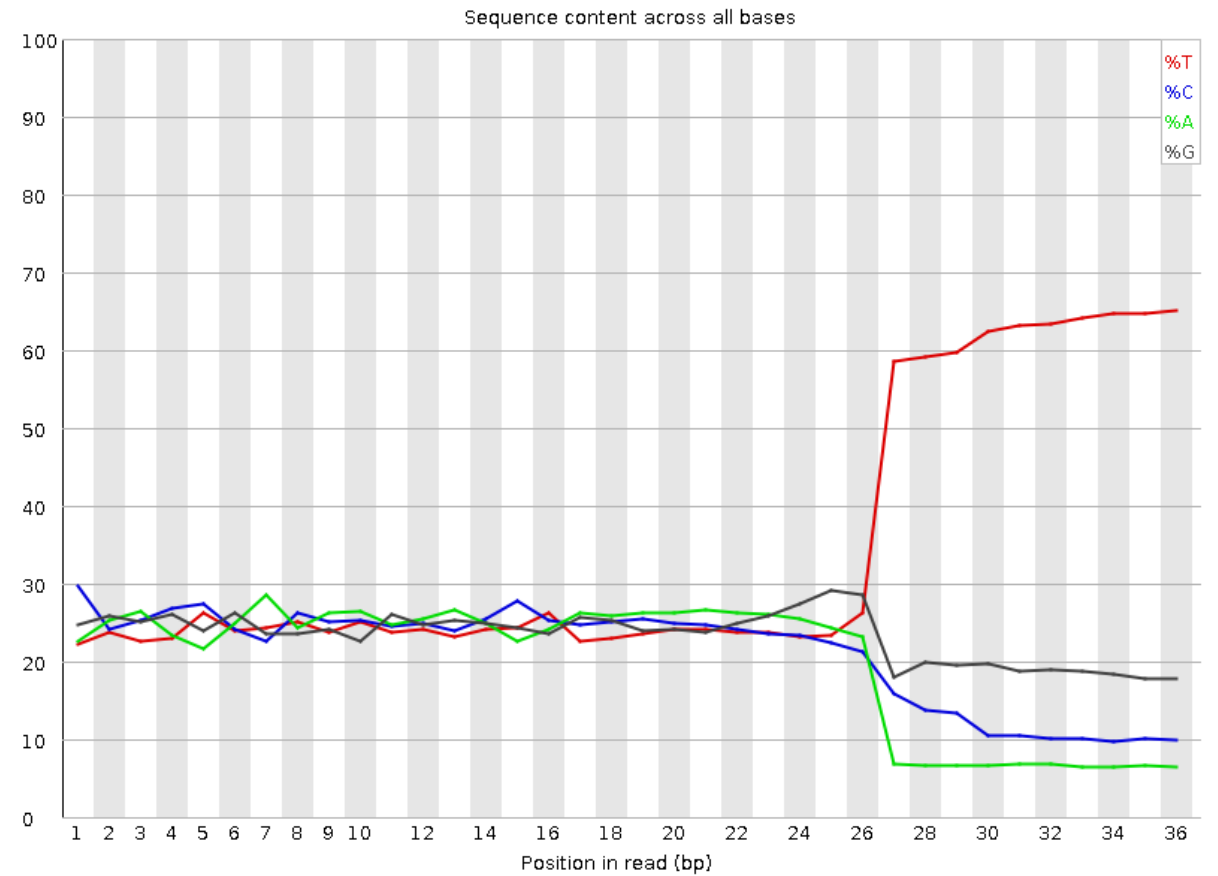
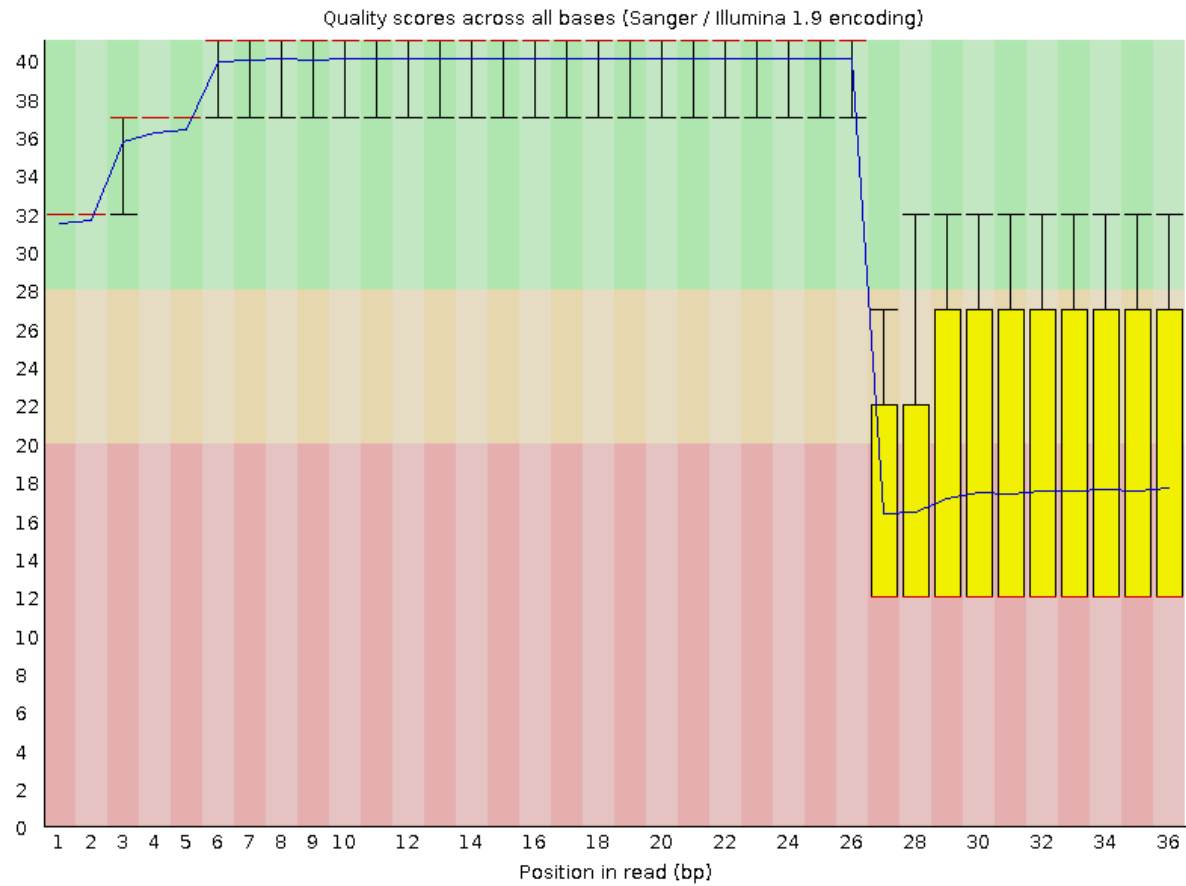
Alert	Value	Detail
 Low Fraction Reads Confidently Mapped To Transcriptome	28.2%	Ideal > 30%. This can indicate use of the wrong reference transcriptome, a reference transcriptome with overlapping genes, poor library quality, poor sequencing quality, or reads shorter than the recommended minimum. Application performance may be affected.

Mapping ?	
Reads Mapped to Genome	47.5%
Reads Mapped Confidently to Genome	46.1%
Reads Mapped Confidently to Intergenic Regions	2.0%
Reads Mapped Confidently to Intronic Regions	14.2%
Reads Mapped Confidently to Exonic Regions	29.9%
Reads Mapped Confidently to Transcriptome	28.2%
Reads Mapped Antisense to Gene	0.6%

← Actual Problem

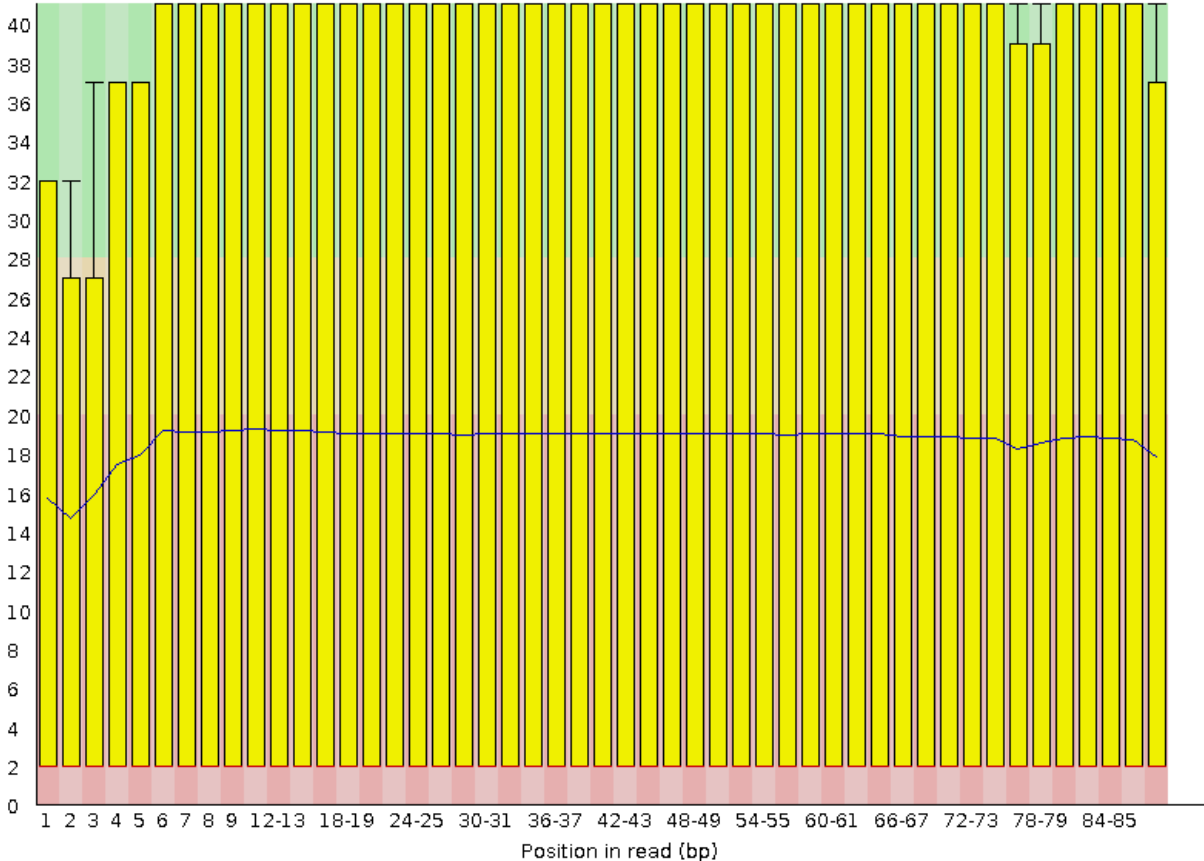
← Value Reported

Course Data QC – Read1 (Barcodes)



Course Data QC – Read2 (RNA)

Quality scores across all bases (Sanger / Illumina 1.9 encoding)



Quality per tile

