Analysing 10X Single Cell RNA-Seq Data

v2019-06

Simon Andrews

simon.andrews@babraham.ac.uk

Babraham Bioinformatics

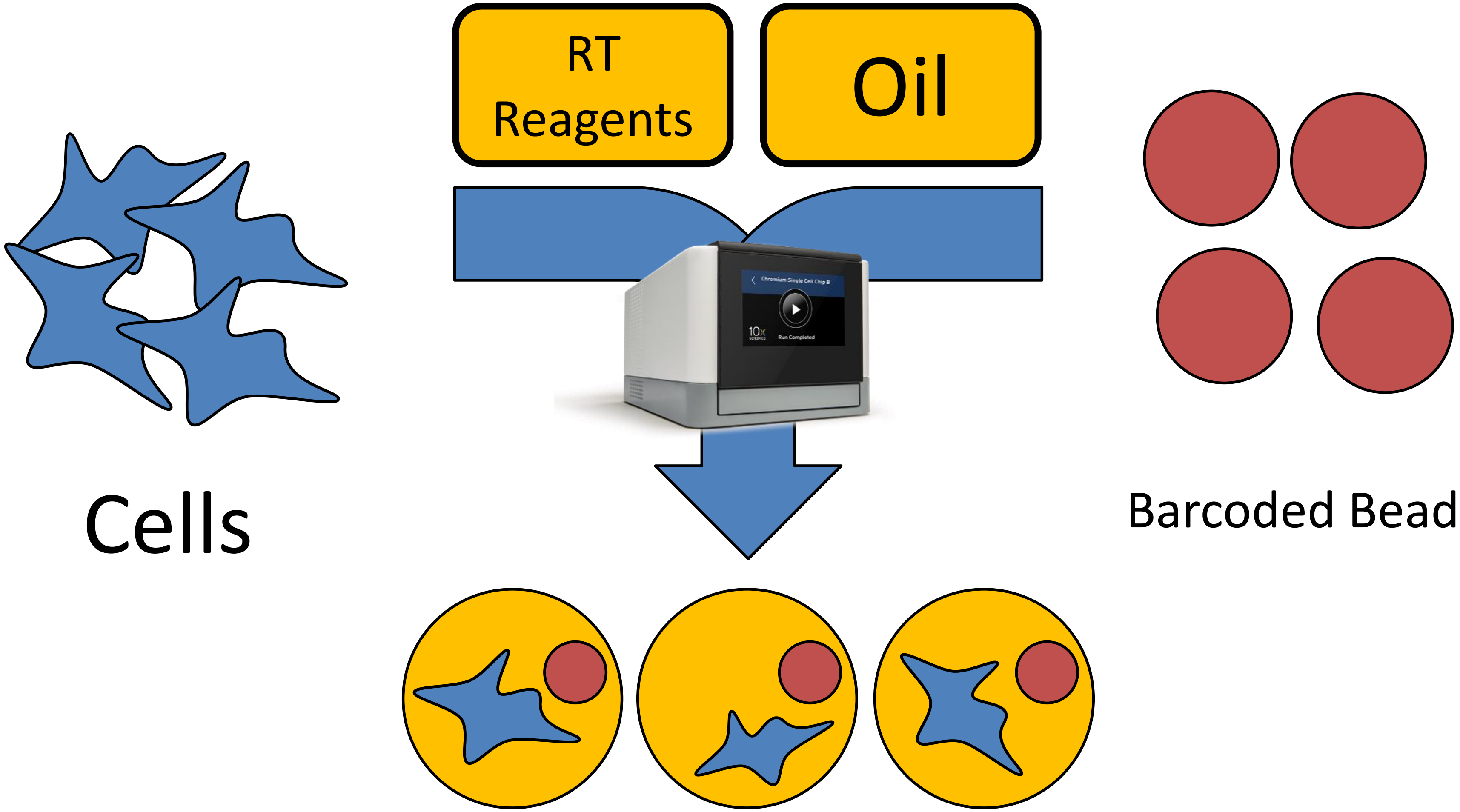
Course Outline

- How 10X single cell RNA-Seq works

- Evaluating CellRanger QC
  - [Exercise] Looking at CellRanger QC reports

- Dimensionality Reduction (PCA and tSNE)
  - [Exercise] Using the Loupe cell browser
  - [Exercise] Analysing data in R using Seurat

How 10X RNA-Seq Works



RT Reagents

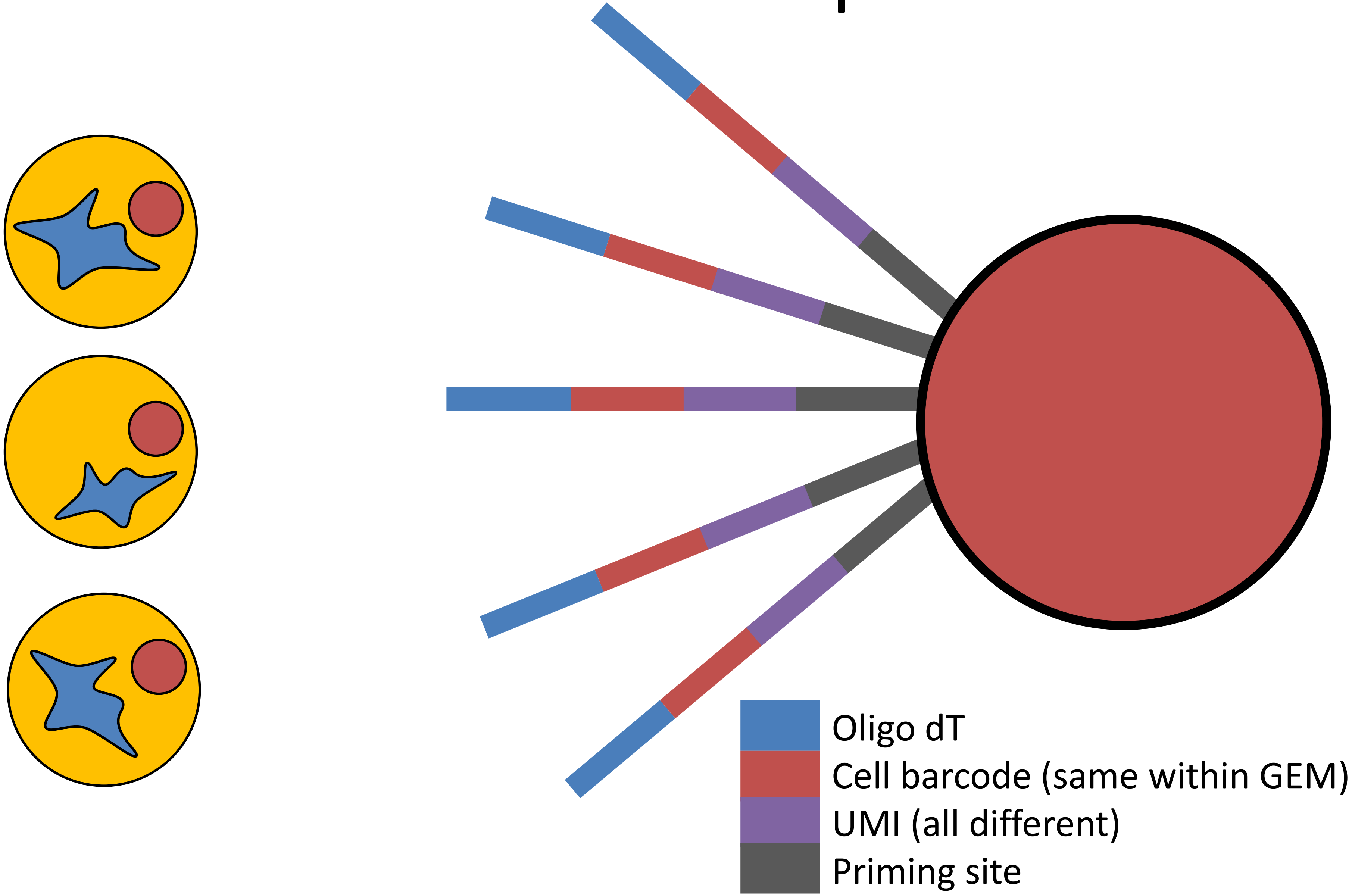
Oil

Cells

Barcoded Beads

Gel Beads in Emulsion (GEMs)

How 10X RNA-Seq Works



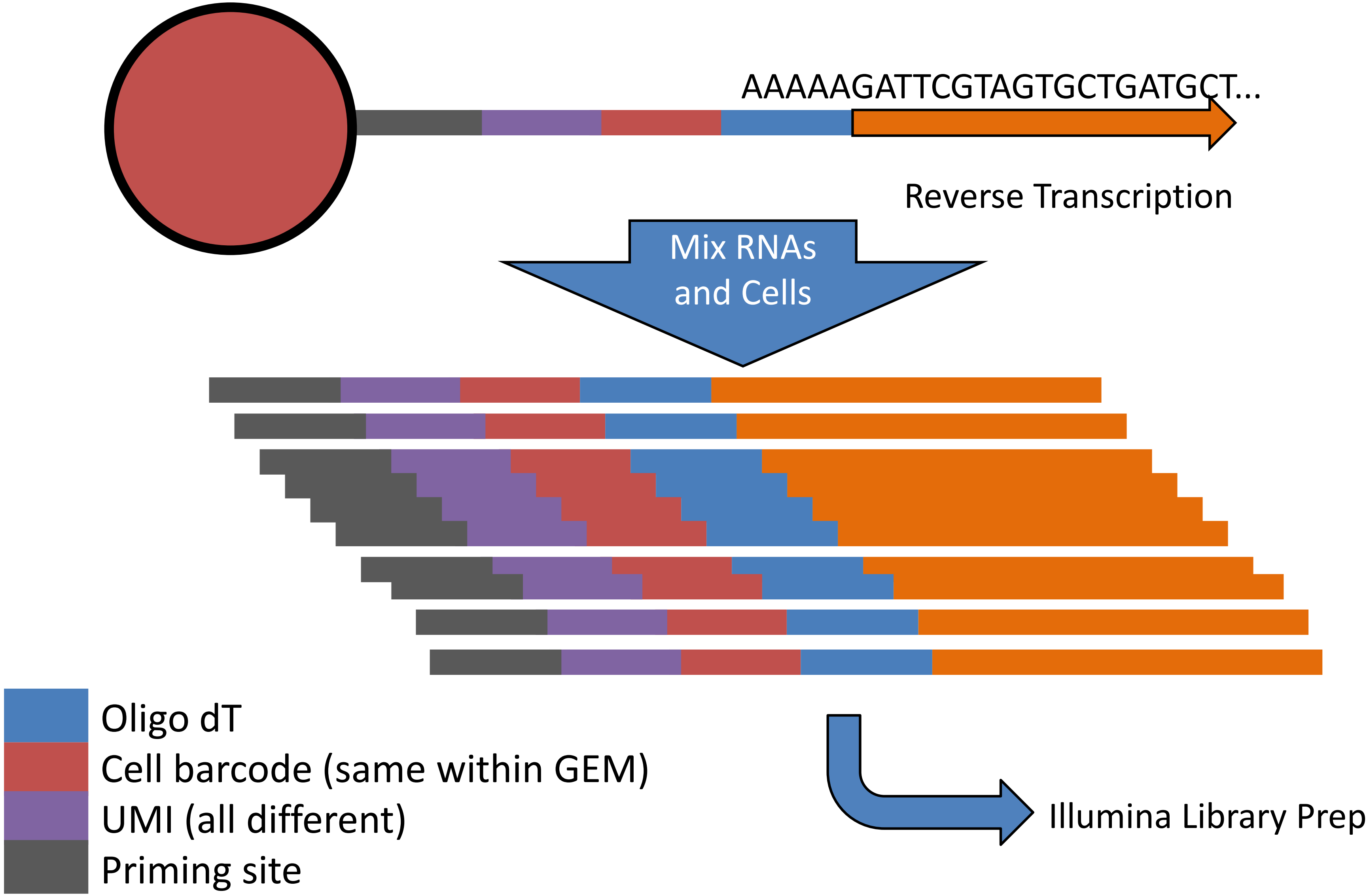
🟦 Oligo dT

🟥 Cell barcode (same within GEM)

🟪 UMI (all different)

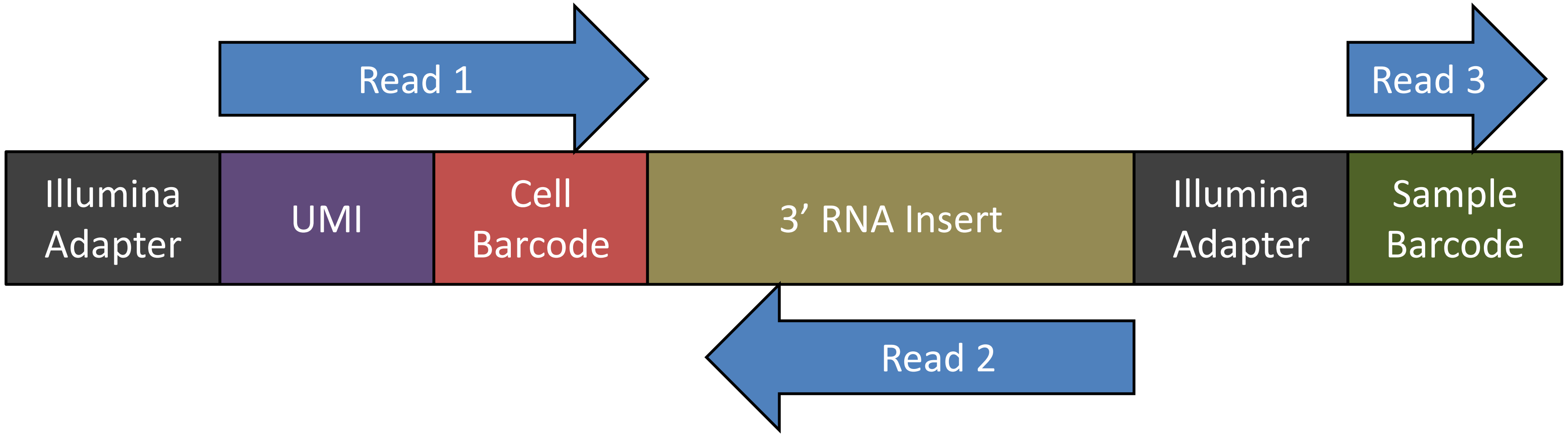
⬛ Priming site

How 10X RNA-Seq Works



🟦 Oligo dT
🟥 Cell barcode (same within GEM)
🟪 UMI (all different)
⬛ Priming site

How 10X RNA-Seq Works



Read 1

Read 3

Illumina Adapter

UMI

Cell Barcode

3’ RNA Insert

Illumina Adapter

Sample Barcode

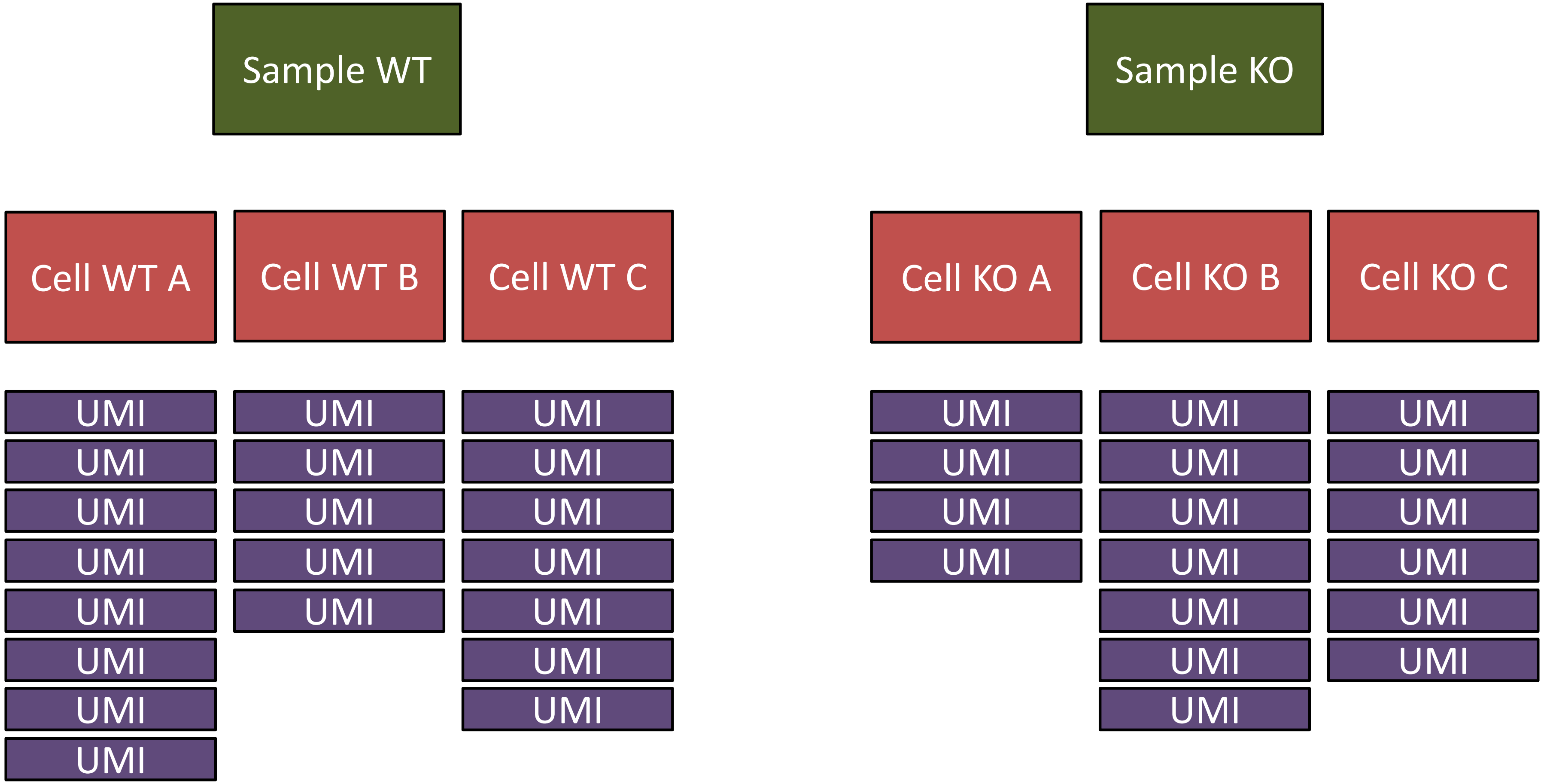
Read 2

🟩 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



Sample WT

Cell WT A

UMI
UMI
UMI
UMI
UMI
UMI
UMI
UMI

Cell WT B

UMI
UMI
UMI
UMI
UMI

Cell WT C

UMI
UMI
UMI
UMI
UMI
UMI
UMI

Sample KO

Cell KO A

UMI
UMI
UMI
UMI

Cell KO B

UMI
UMI
UMI
UMI
UMI
UMI
UMI

Cell KO C

UMI
UMI
UMI
UMI
UMI
UMI

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

The 10X Software Suite

Chromium Controller

Cell Ranger

Loupe Browser

Runs the chromium system for creating GEMs

Pipeline for mapping, filtering, QC and quantitation of libraries

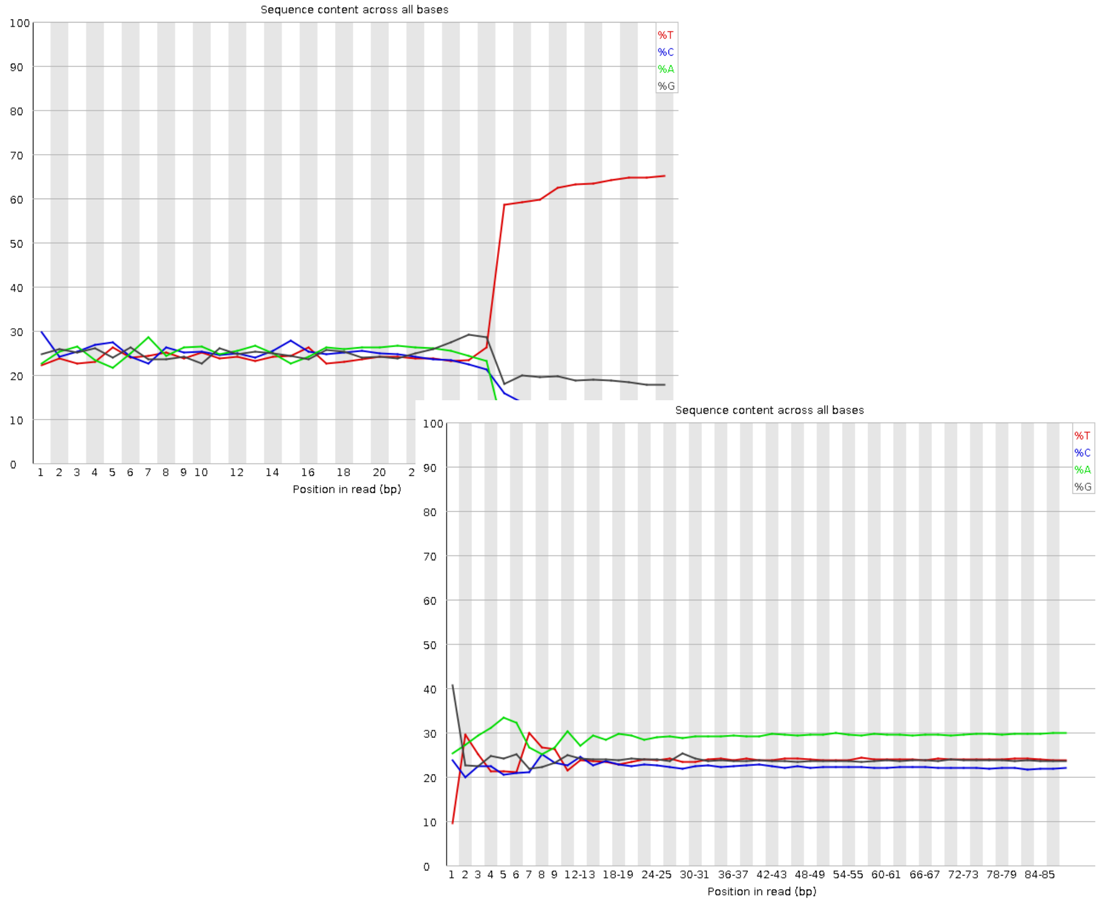
Desktop software for visualisation and analysis of single cell data.

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. All identical (or one of 4) at Babraham
- `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
```

Output files generated

- `web_summary.html` - Web format QC report

- `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

- `cloupe.cloupe` - Analysis data for Loupe Cell browser

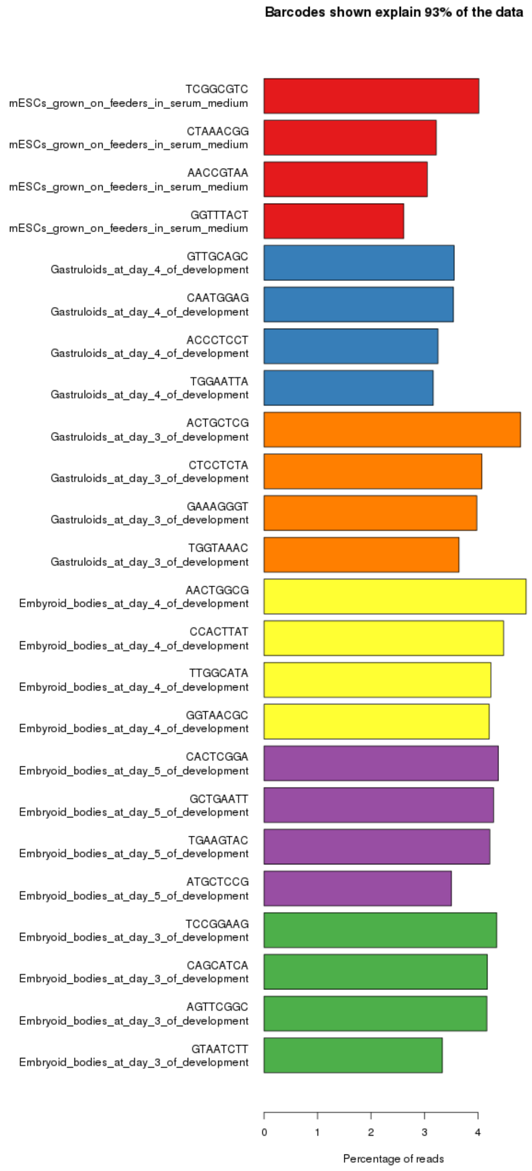
Evaluating CellRanger Output

- 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

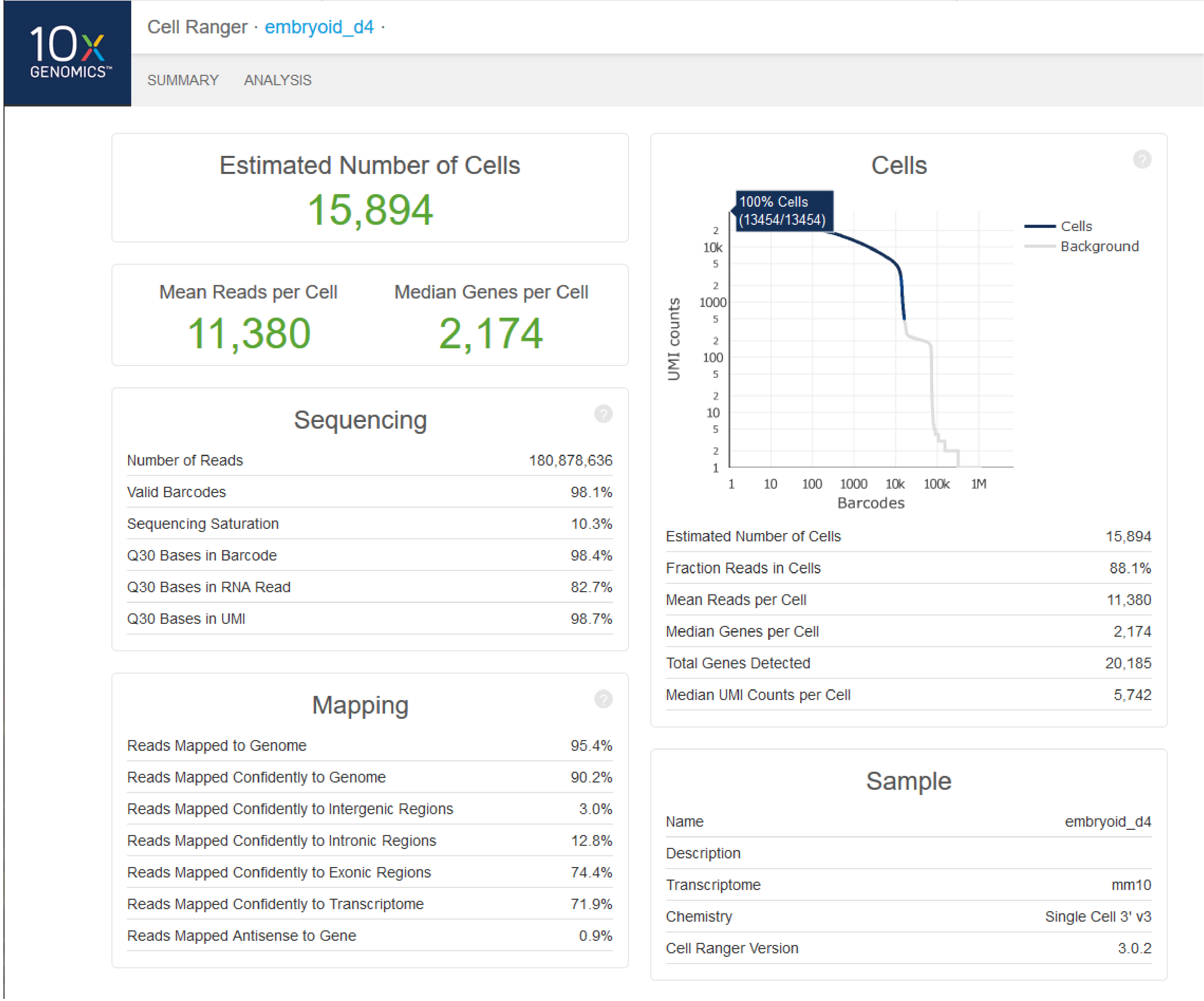
- Only present if multiple libraries mixed in a lane
- Get standard barcode split report, but with 4 barcodes used per sample
- Even coverage within and between libraries



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



Estimated Number of Cells

15,894

Mean Reads per Cell	Median Genes per Cell
11,380	2,174

Sequencing

Number of Reads	180,878,636
Valid Barcodes	98.1%
Sequencing Saturation	10.3%
Q30 Bases in Barcode	98.4%
Q30 Bases in RNA Read	82.7%
Q30 Bases in UMI	98.7%

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%

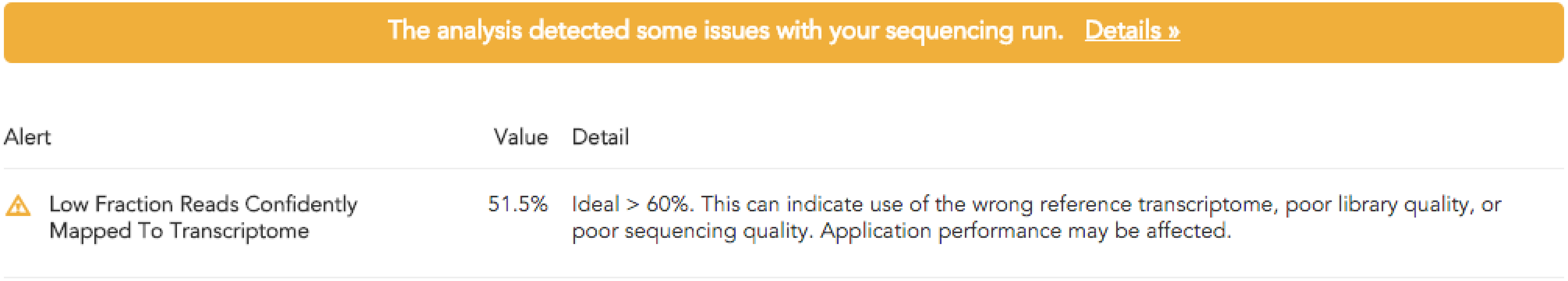
Cells

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

Sample

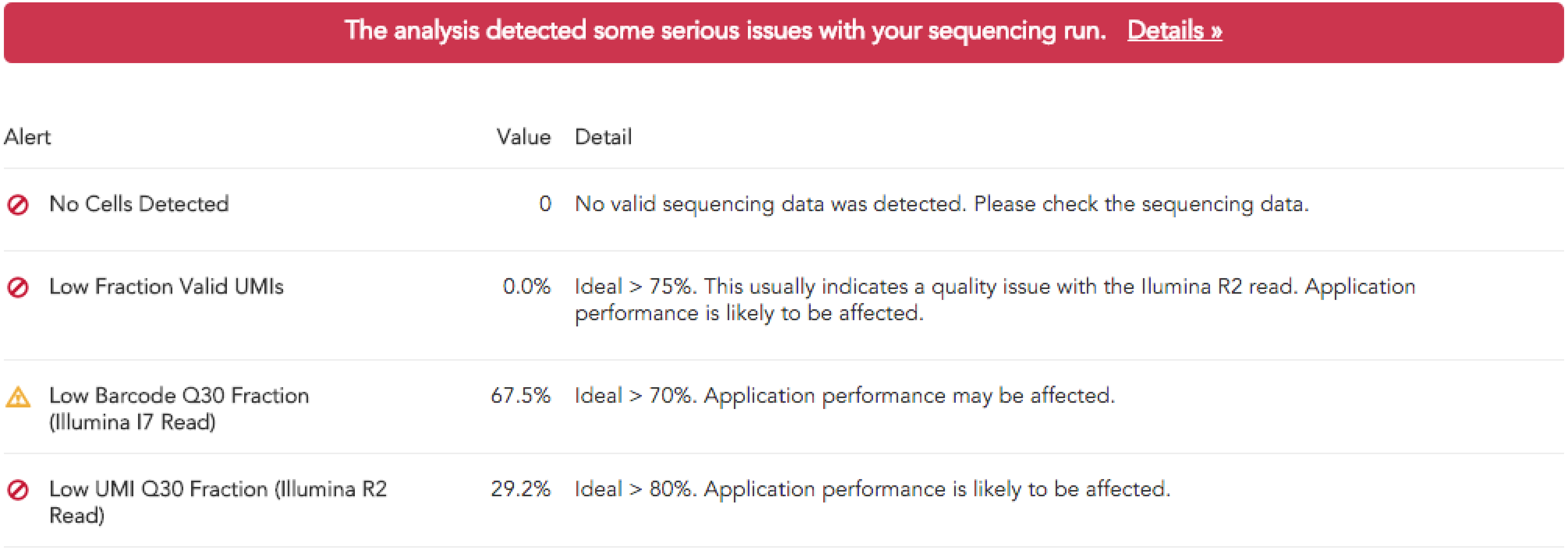
Name	embryoid_d4
Description	
Transcriptome	mm10
Chemistry	Single Cell 3' v3
Cell Ranger Version	3.0.2

Errors and Warnings



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.



The analysis detected some serious issues with your sequencing run. Details »

Alert	Value	Detail
⊘ No Cells Detected	0	No valid sequencing data was detected. Please check the sequencing data.
⊘ Low Fraction Valid UMIs	0.0%	Ideal > 75%. This usually indicates a quality issue with the Ilumina R2 read. Application performance is likely to be affected.
⚠ Low Barcode Q30 Fraction (Illumina I7 Read)	67.5%	Ideal > 70%. Application performance may be affected.
⊘ Low UMI Q30 Fraction (Illumina R2 Read)	29.2%	Ideal > 80%. Application performance is likely to be affected.

How many cells do you have?

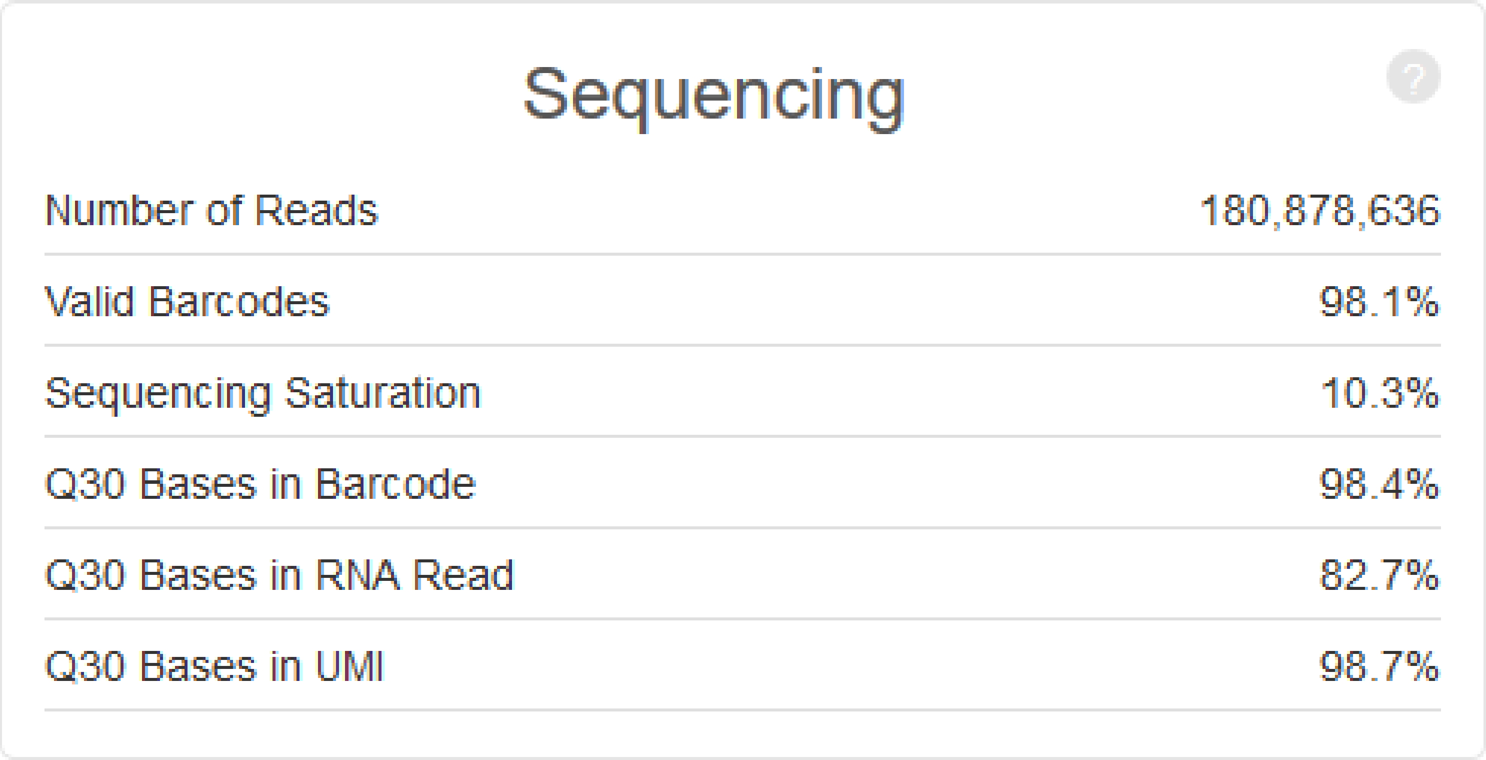
- Cell number is determined from the number of cell barcodes with ‘reasonable’ numbers of observations

- 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

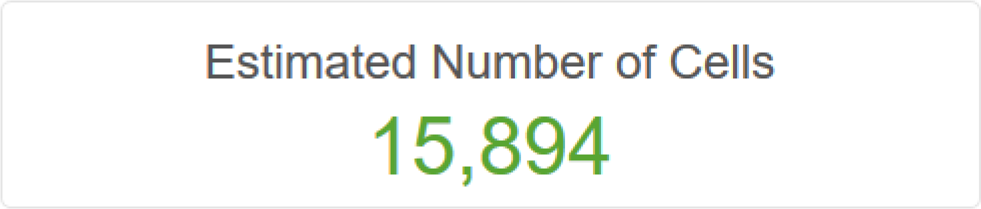


Sequencing

Number of Reads	180,878,636
Valid Barcodes	98.1%
Sequencing Saturation	10.3%
Q30 Bases in Barcode	98.4%
Q30 Bases in RNA Read	82.7%
Q30 Bases in UMI	98.7%

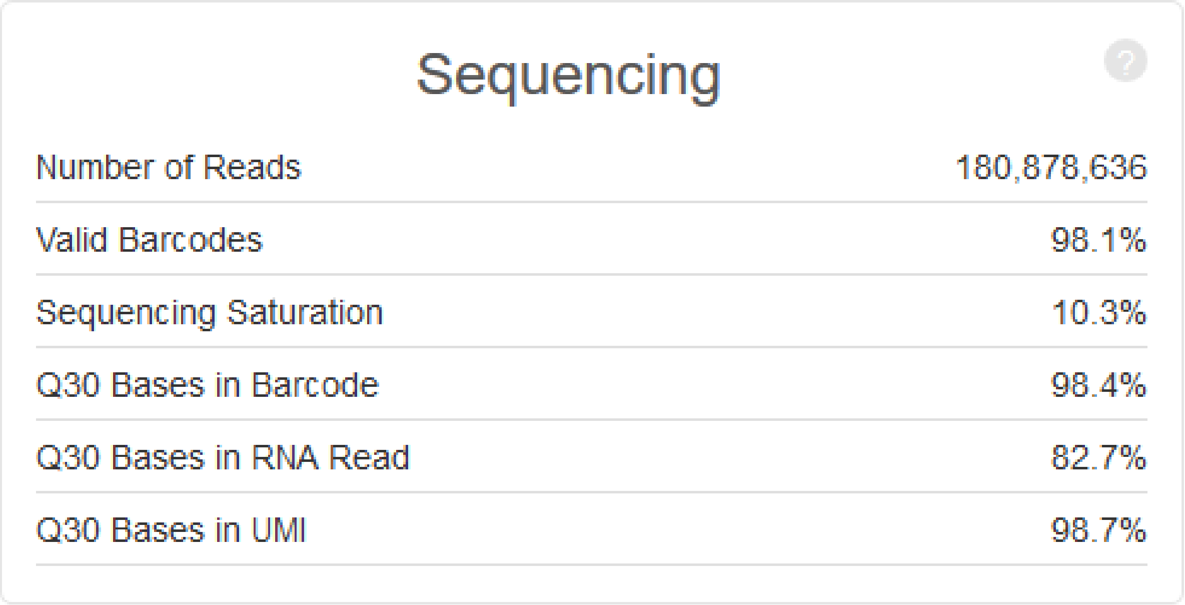
How many cells do you have?

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



Estimated Number of Cells

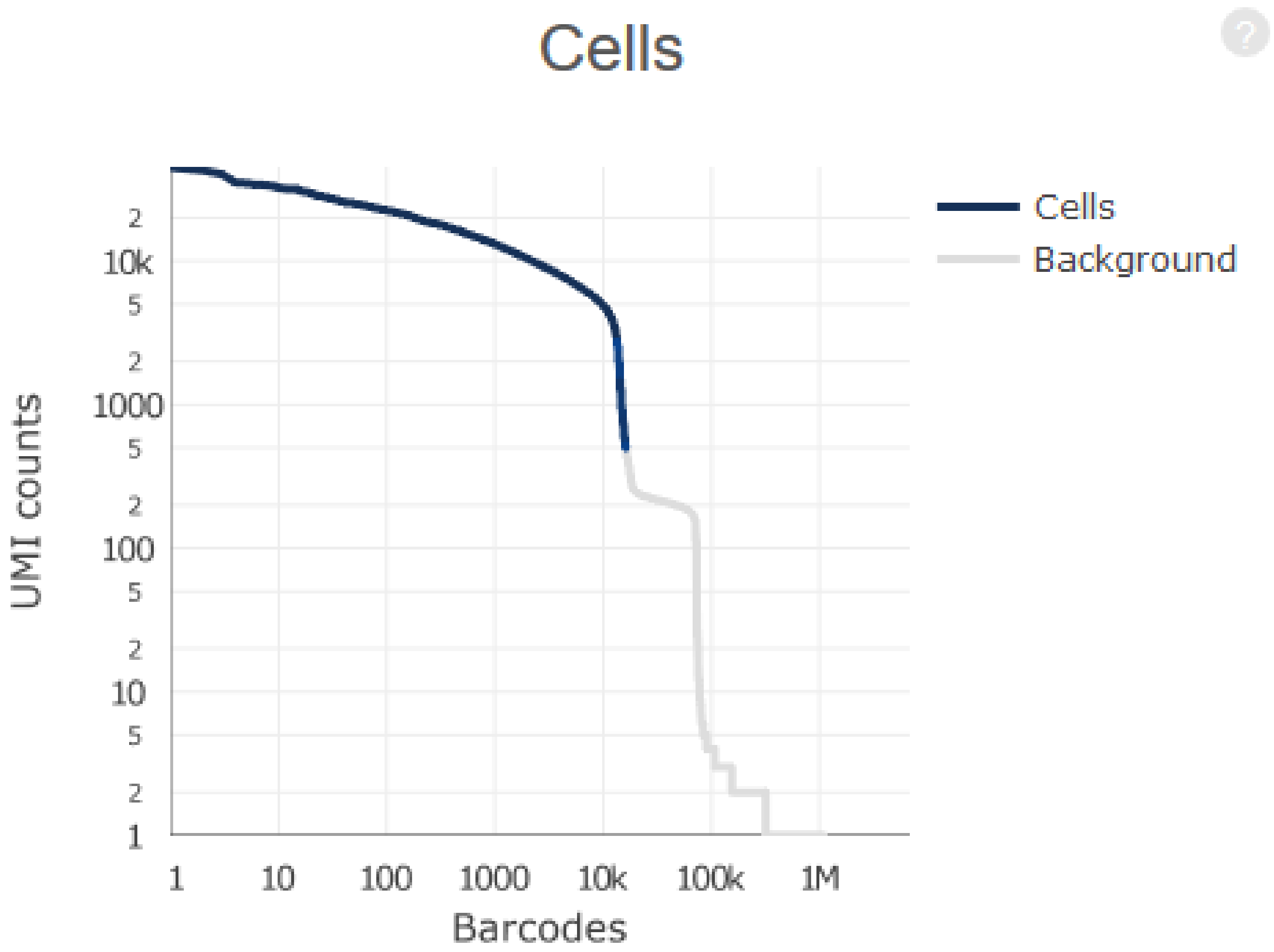
15,894



Sequencing

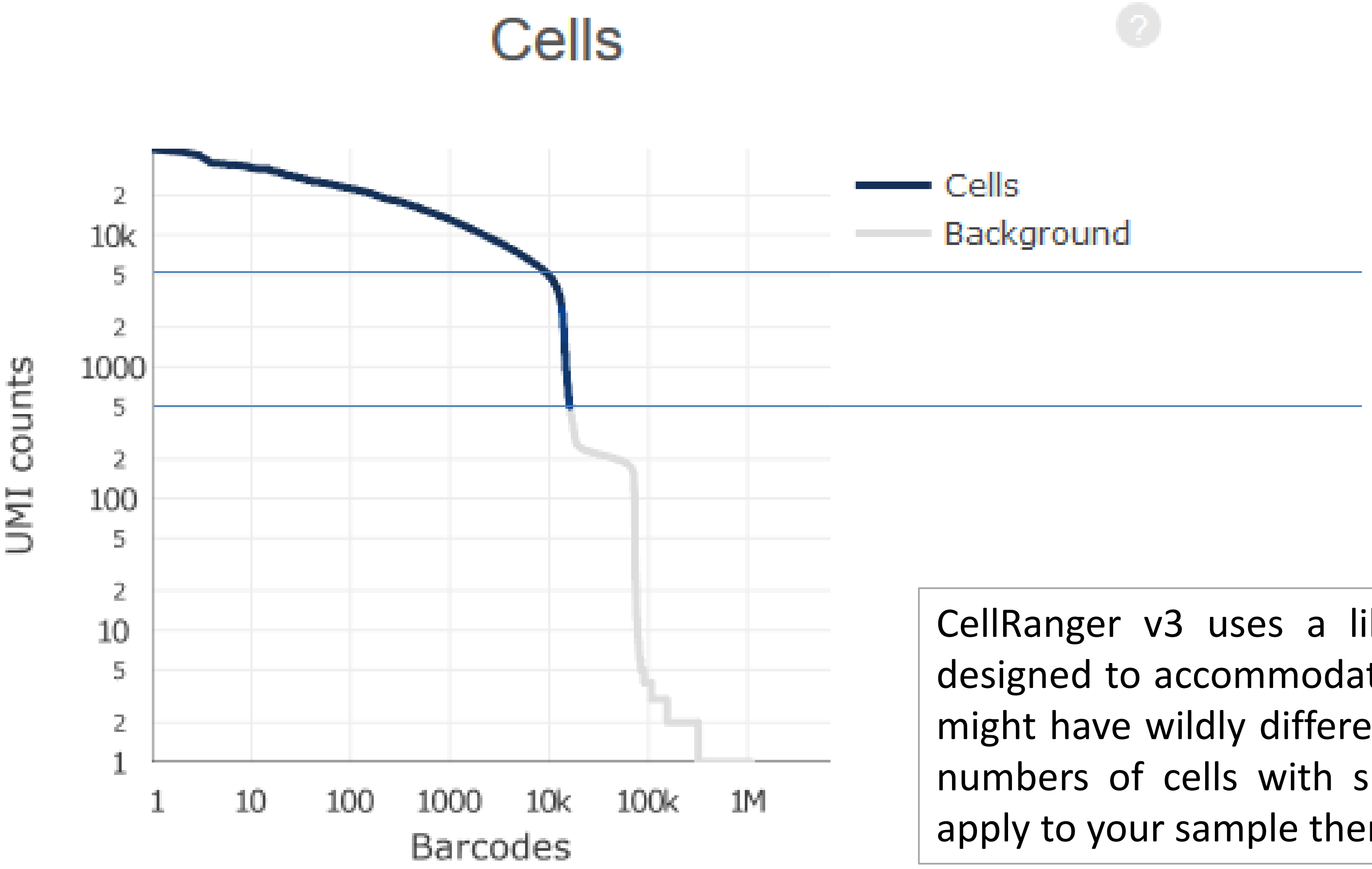
Number of Reads	180,878,636
Valid Barcodes	98.1%
Sequencing Saturation	10.3%
Q30 Bases in Barcode	98.4%
Q30 Bases in RNA Read	82.7%
Q30 Bases in UMI	98.7%

How many cells do you have



- 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

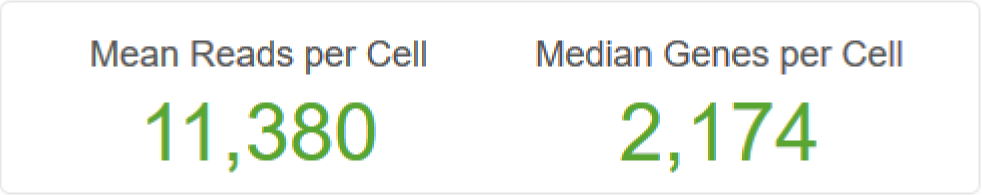


5000 reads per cell. 10k cells

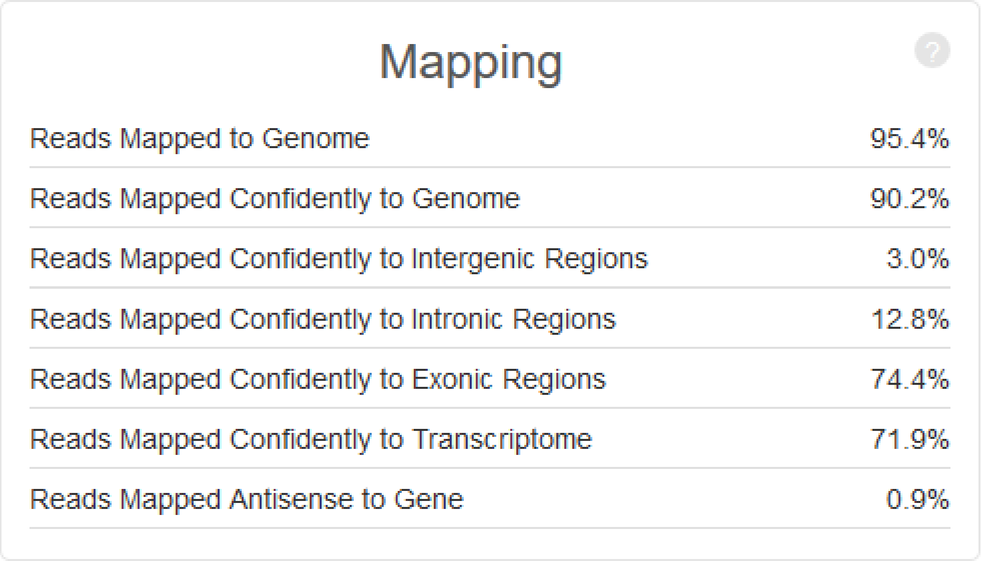
500 reads per cell. 15k cells

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?

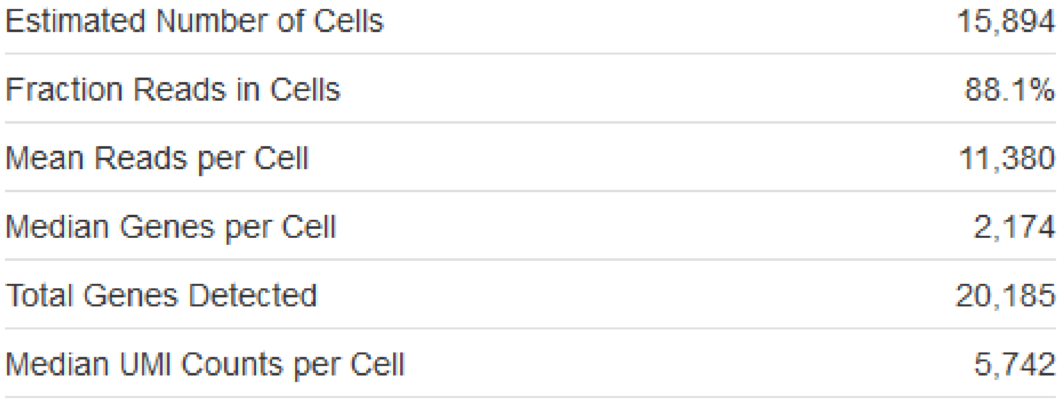


Mean Reads per Cell	Median Genes per Cell
11,380	2,174



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%

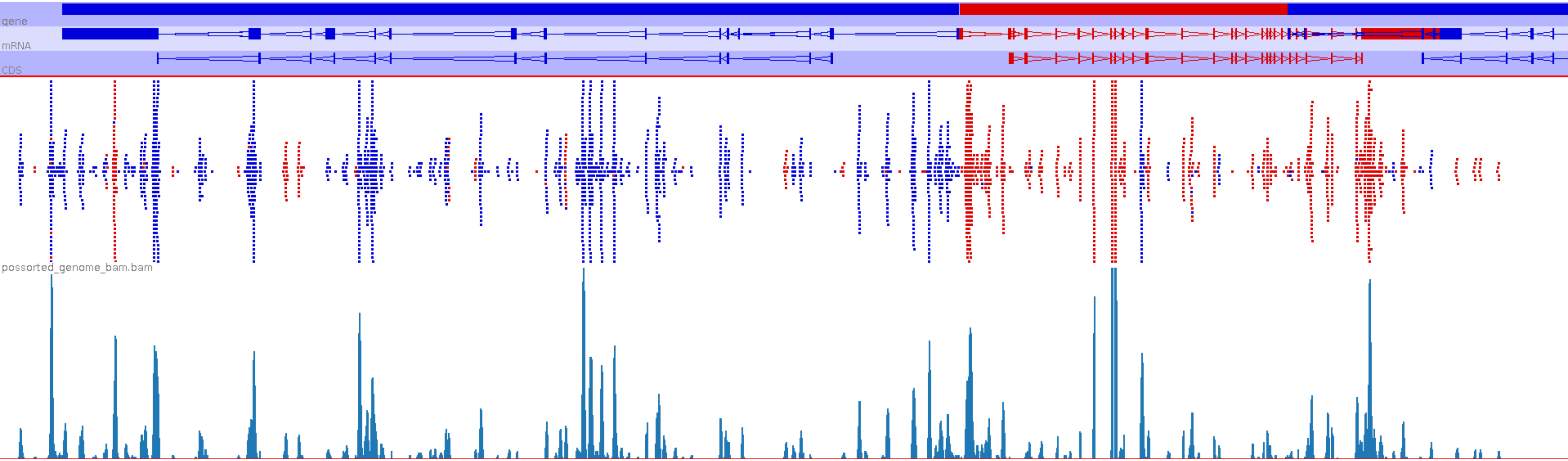


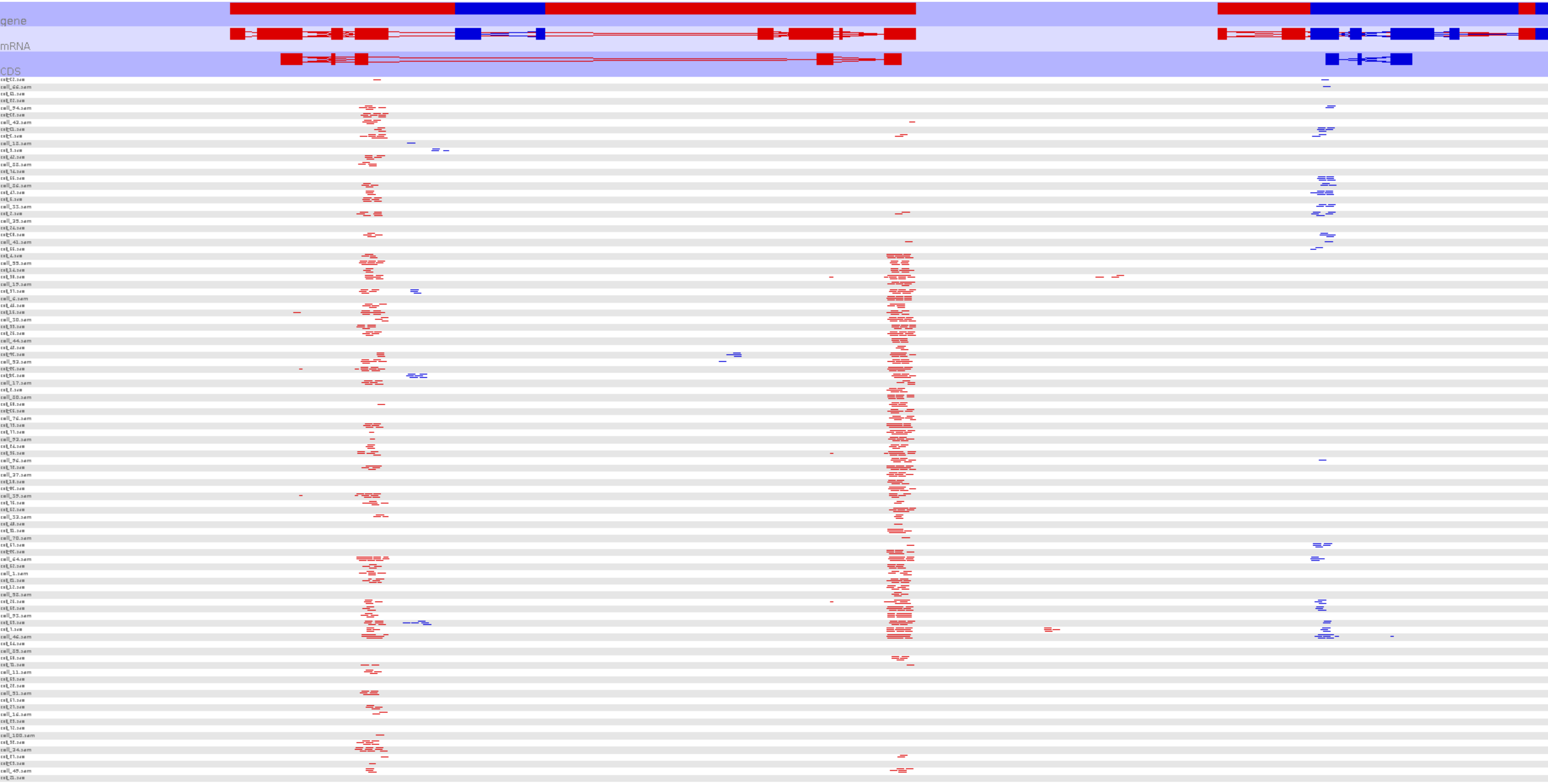
Estimated Number of Cells	15,894
Fraction Reads in Cells	88.1%
Mean Reads per Cell	11,380
Median Genes per Cell	2,174
Total Genes Detected	20,185
Median UMI Counts per Cell	5,742

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

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


gene
mRNA
CDS
possorted_genome_bam.bam

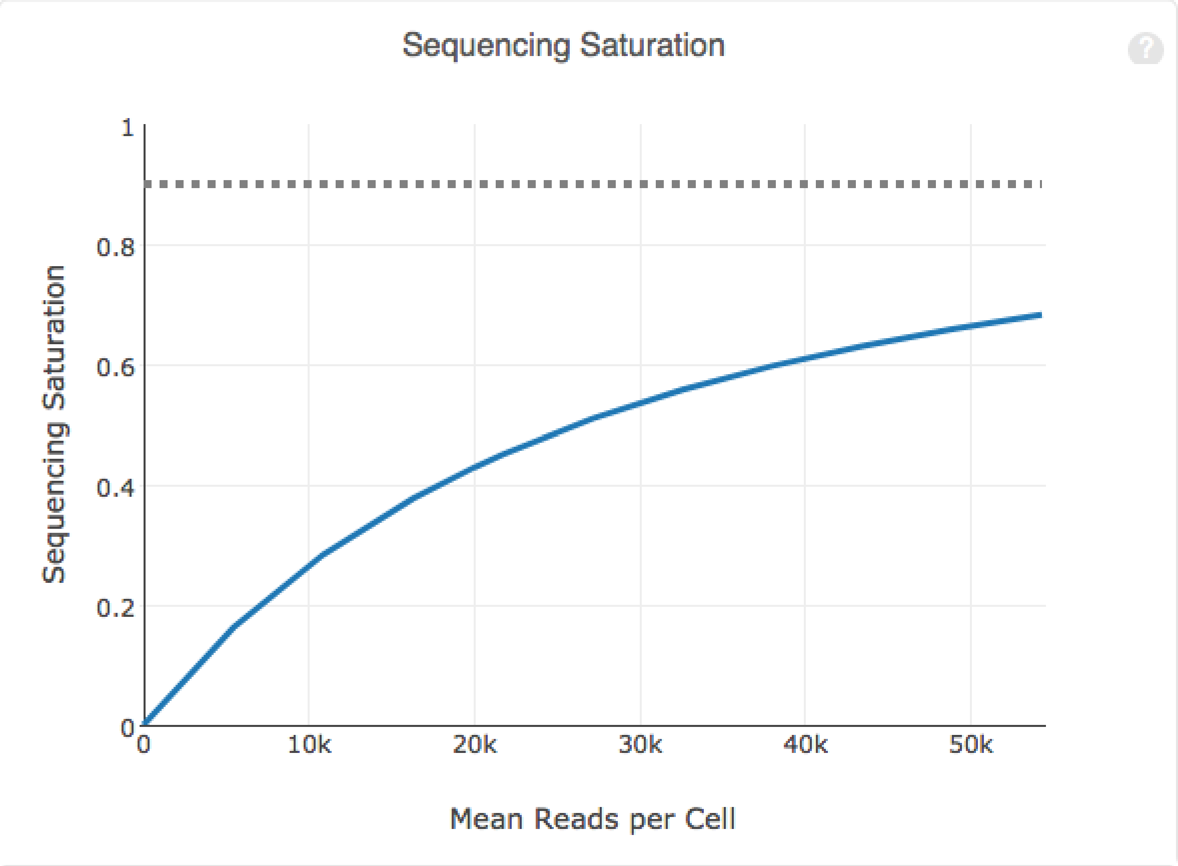

gene
mRNA
CDS

How much data do you have per cell?

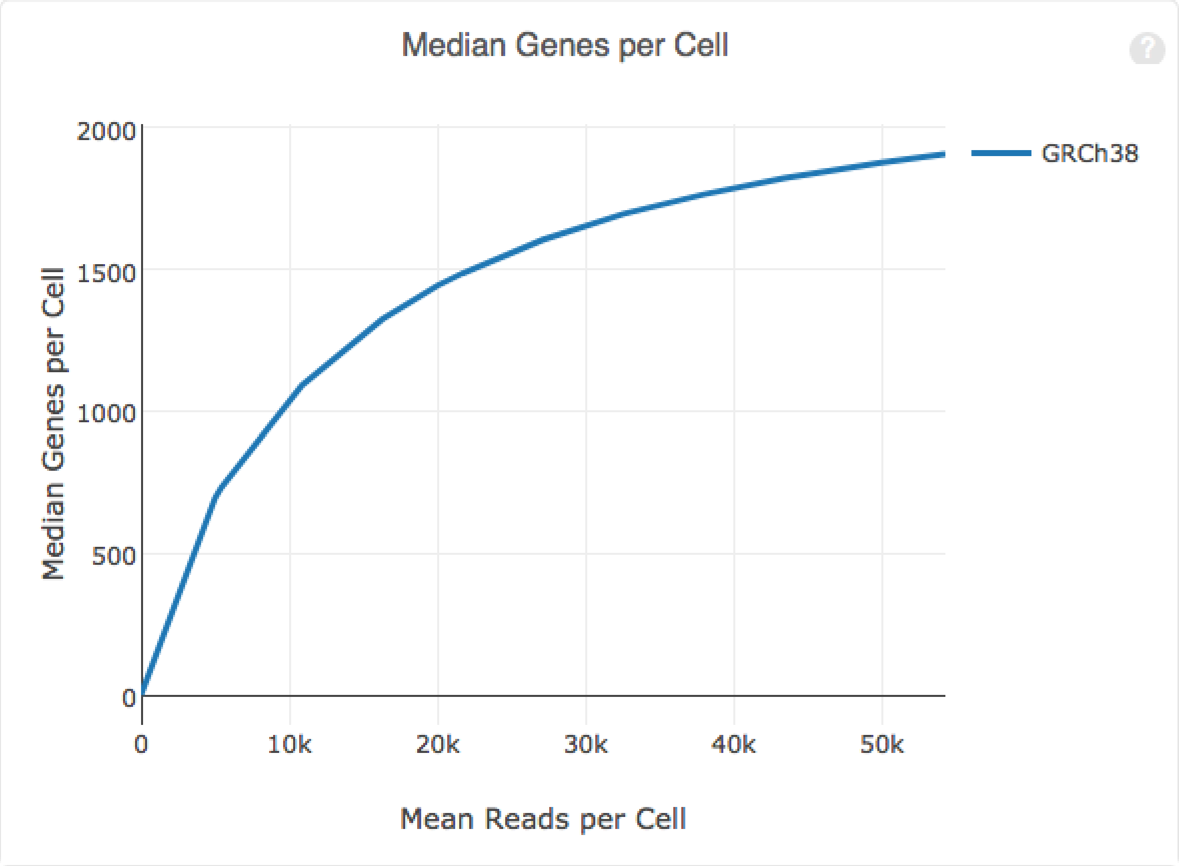
- Difficult to generalise how much data to create/expect
  - Depends on cell type, genome and other factors

- In general though, sensible numbers would be:
  - Reads per cell ~10,000
  - Genes per cell 2000 - 3000

How deeply sequenced is your library

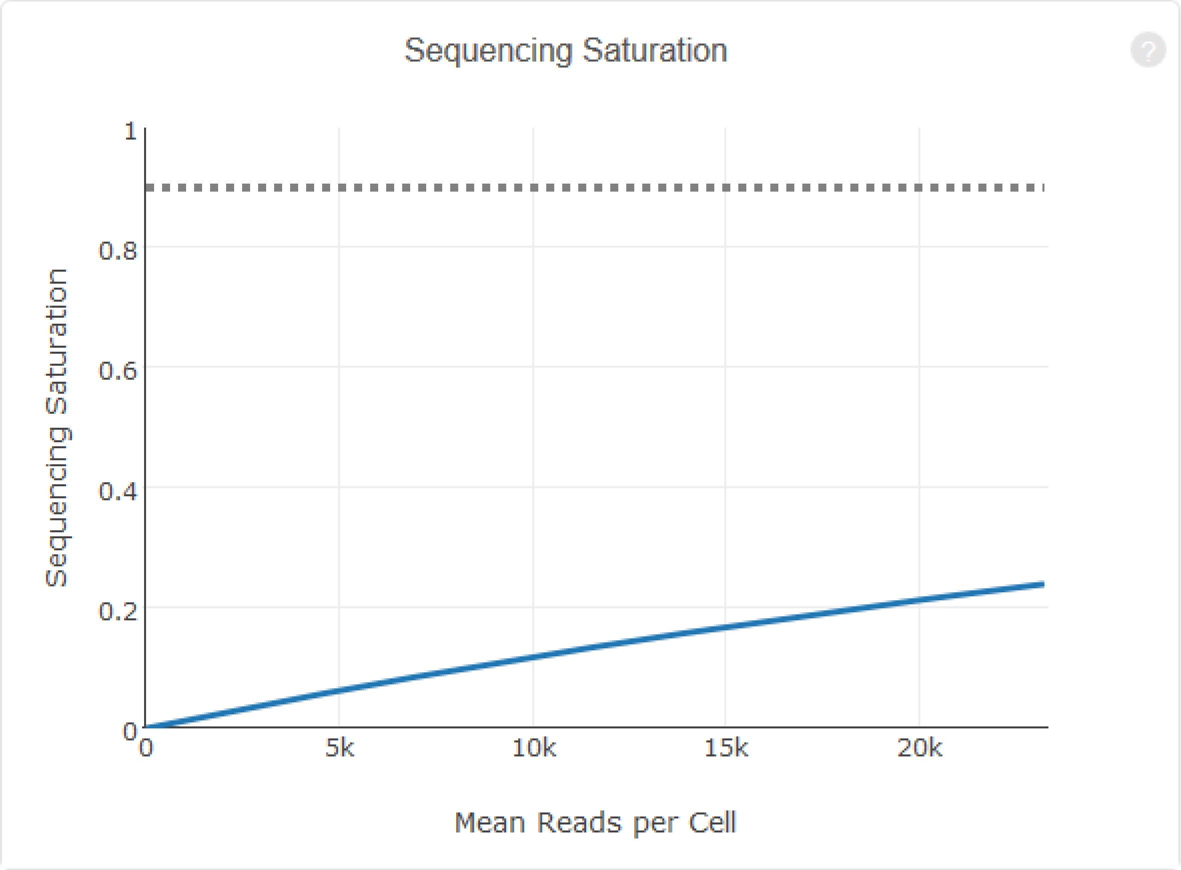


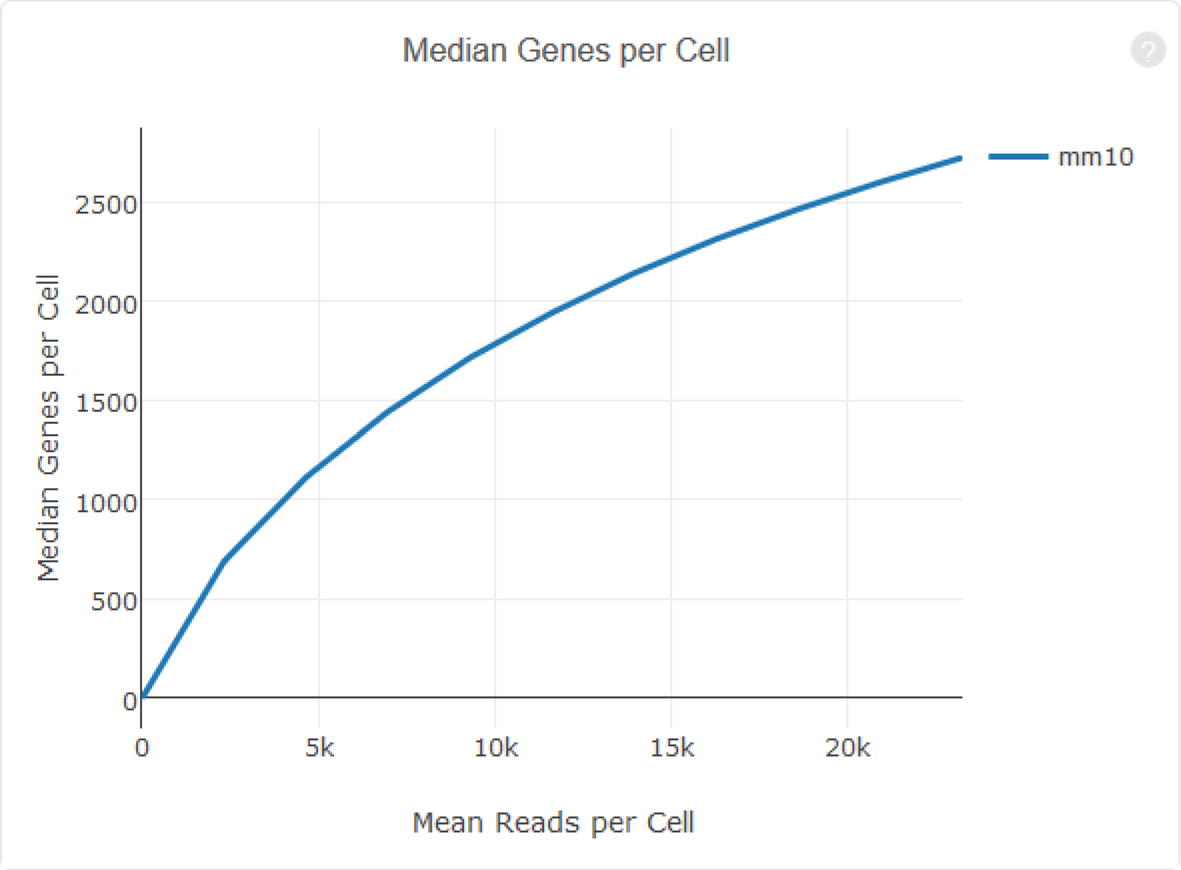
Sequencing Saturation



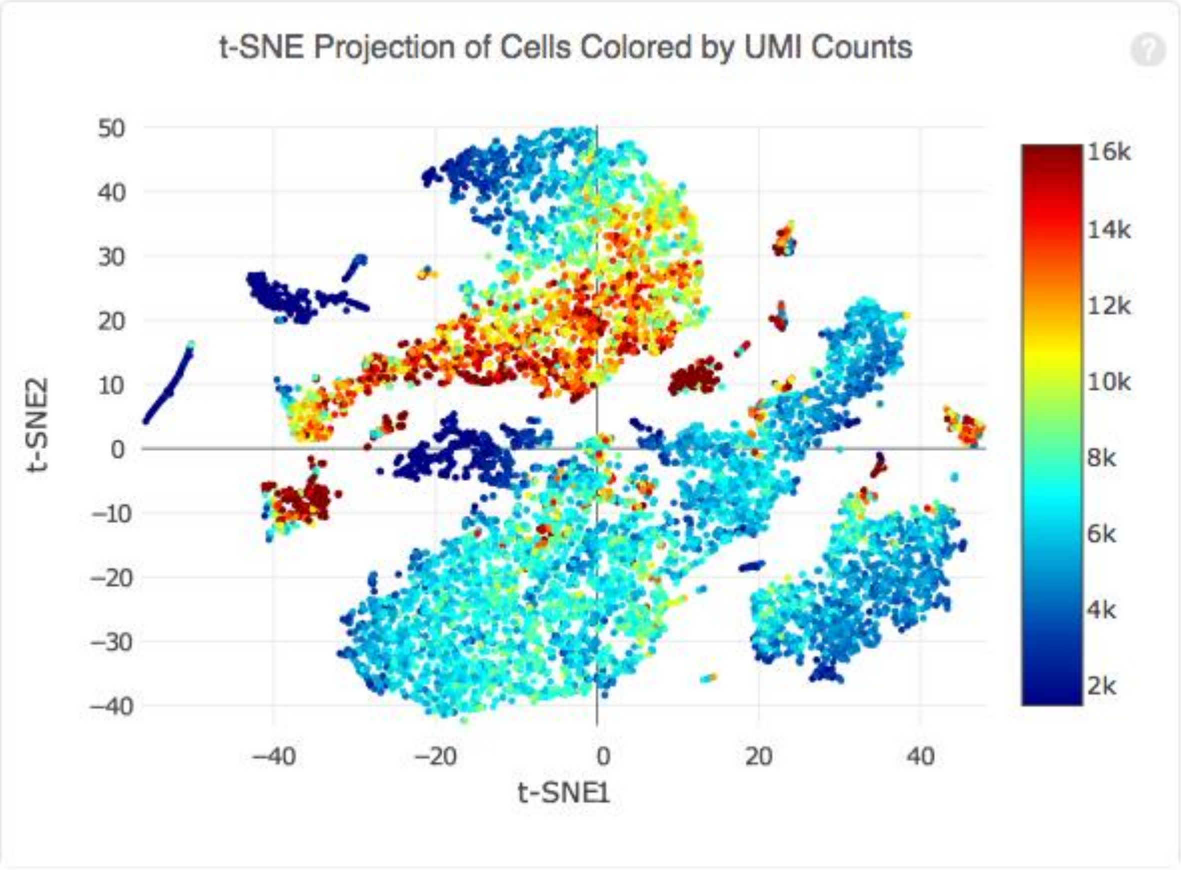
Median Genes per Cell

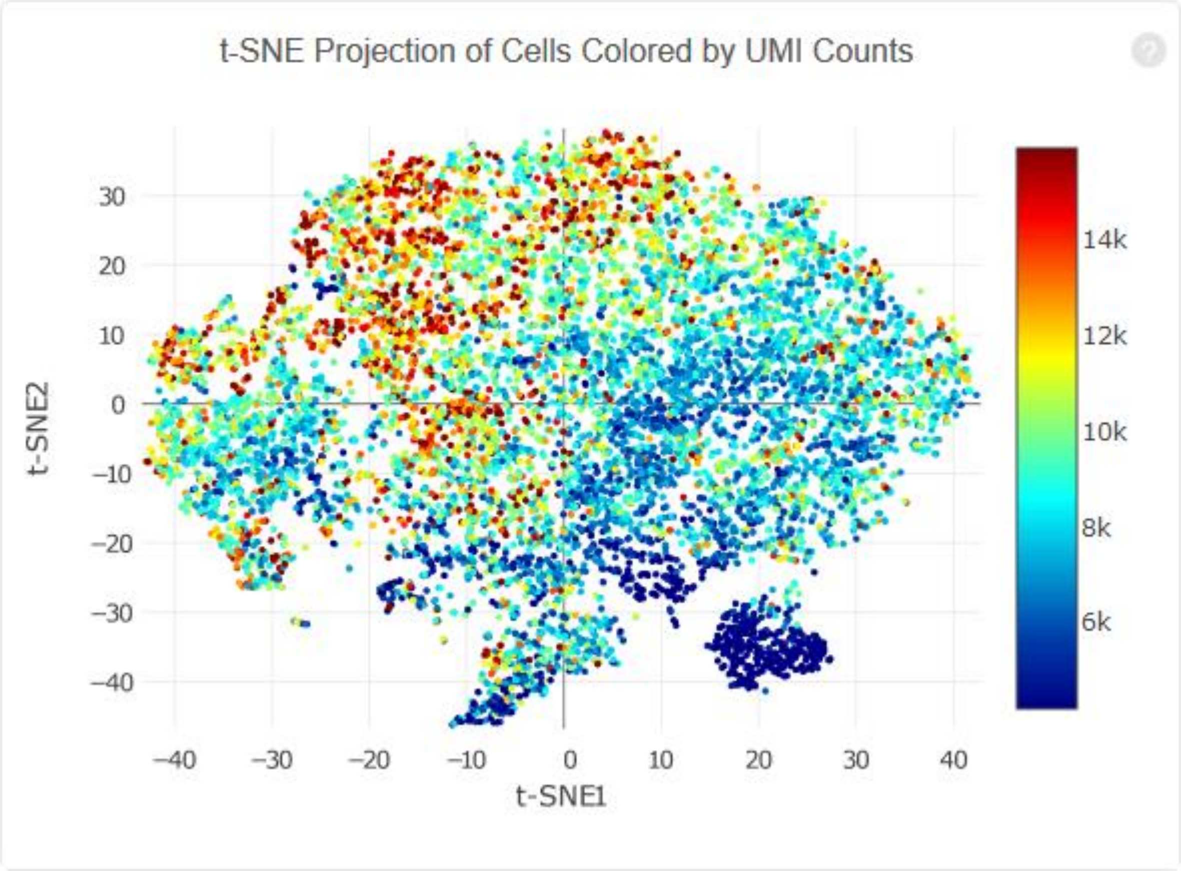
How deeply sequenced is your library





Is coverage variation affecting your data?



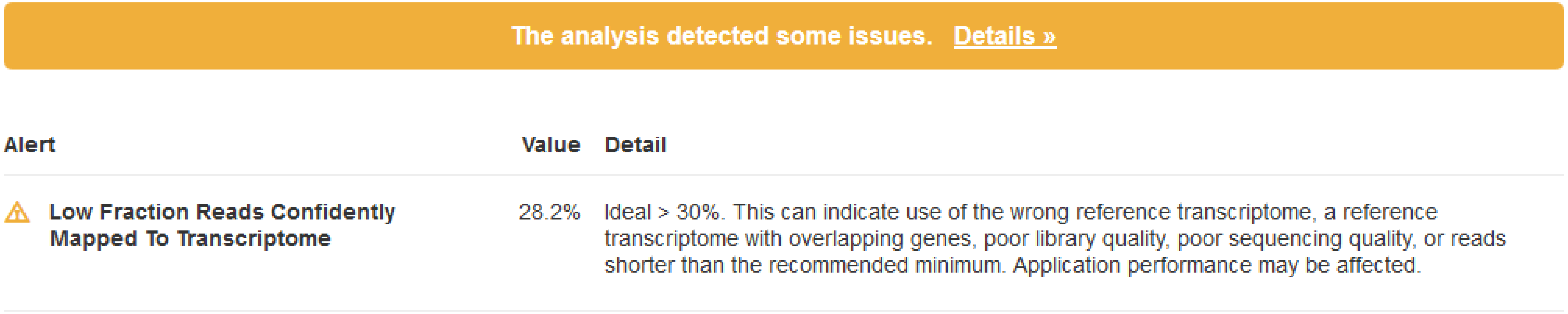


Exercise – Evaluating CellRanger Reports

- Look at the selection of CellRanger reports to get an idea for the metrics they provide

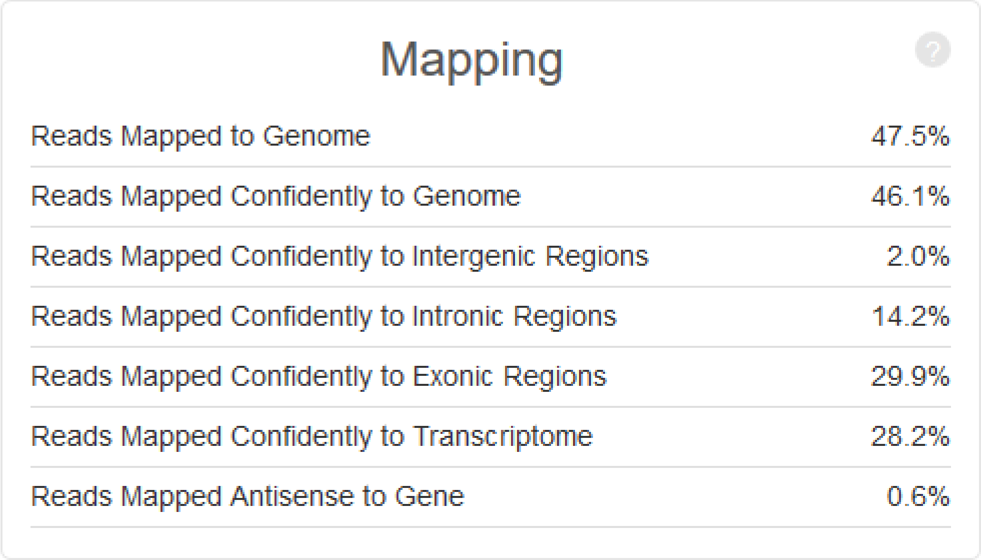
- The data we're going to use for the rest of the day is in "course_web_summary.html", do you see any problems which would concern us with this data at this stage?

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.



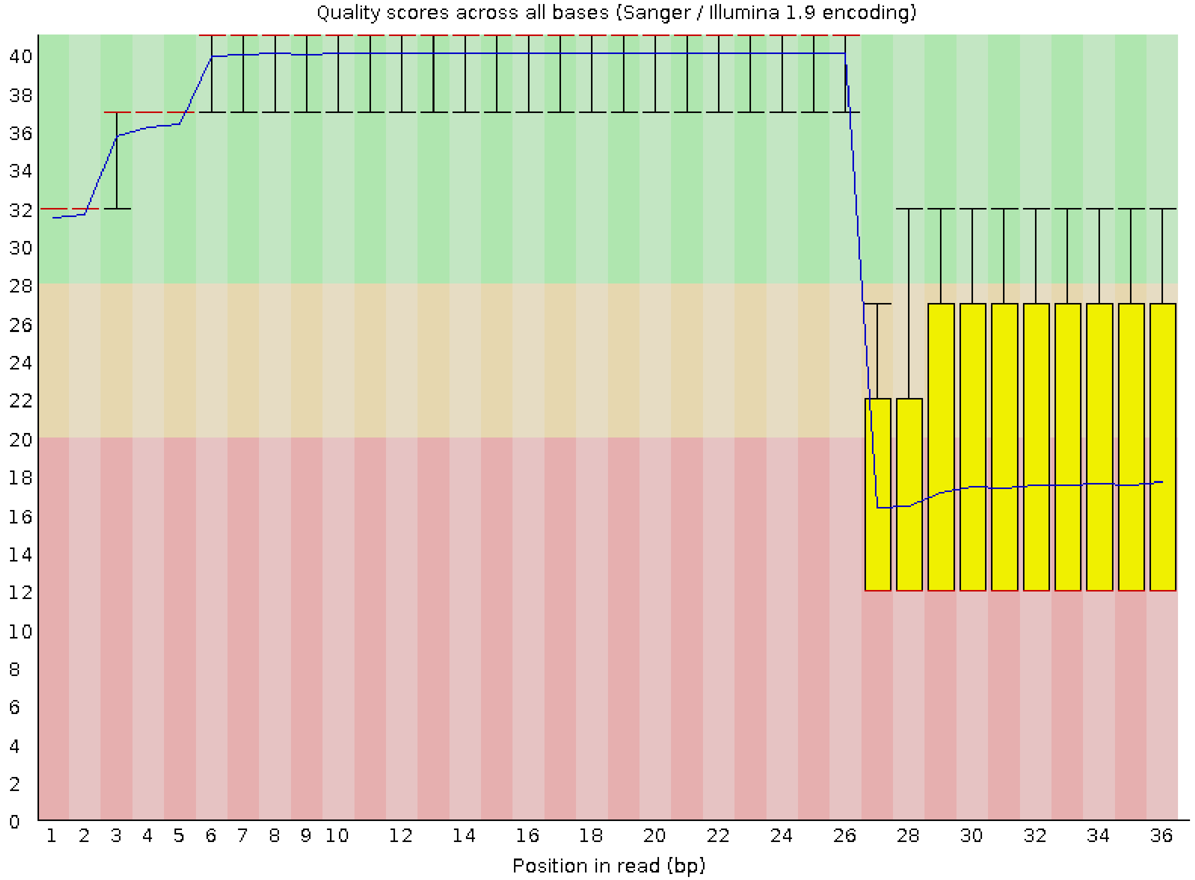
Mapping

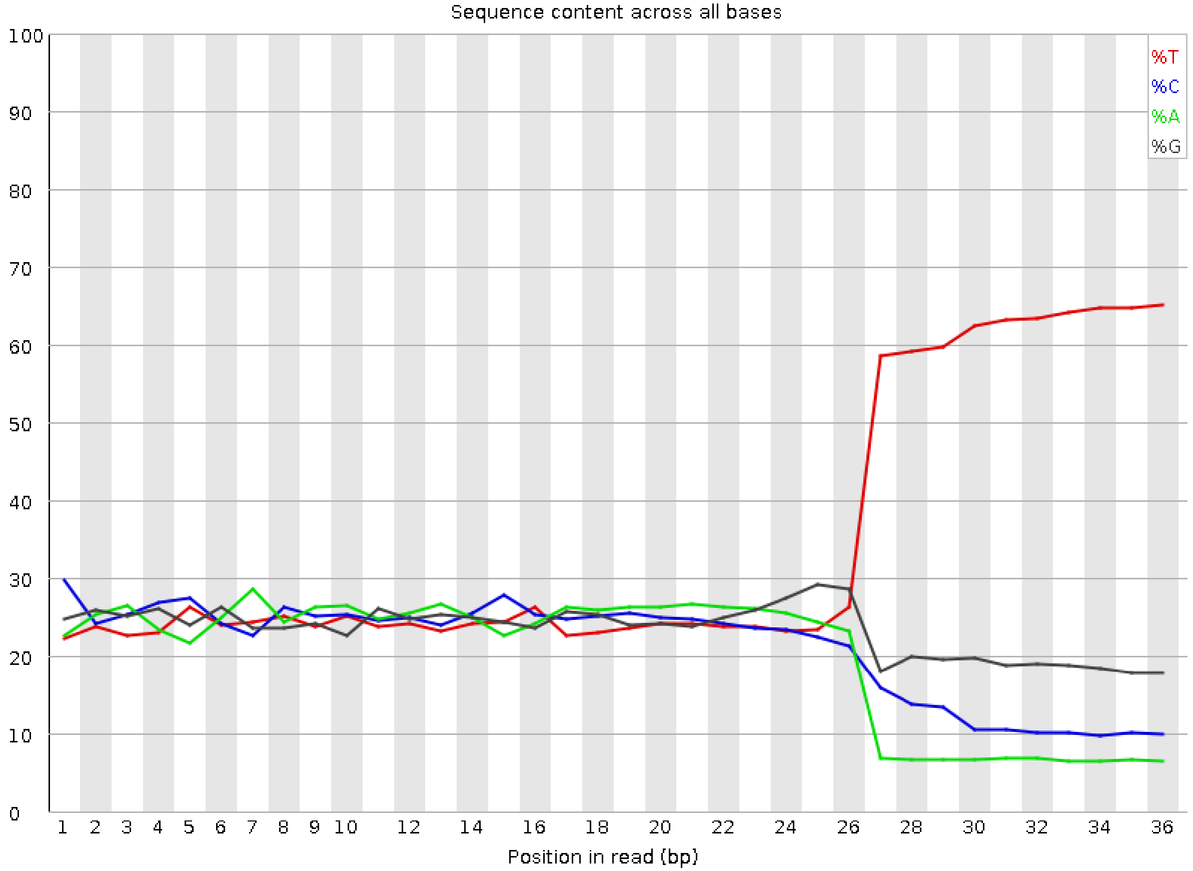
Reads Mapped to Genome	47.5%
Reads Mapped Confidently to Genome	46.1%
Reads Mapped Confidently to Intergenic Regions	2.0%
Reads Mapped Confidently to Intronic Regions	14.2%
Reads Mapped Confidently to Exonic Regions	29.9%
Reads Mapped Confidently to Transcriptome	28.2%
Reads Mapped Antisense to Gene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

← Actual Problem

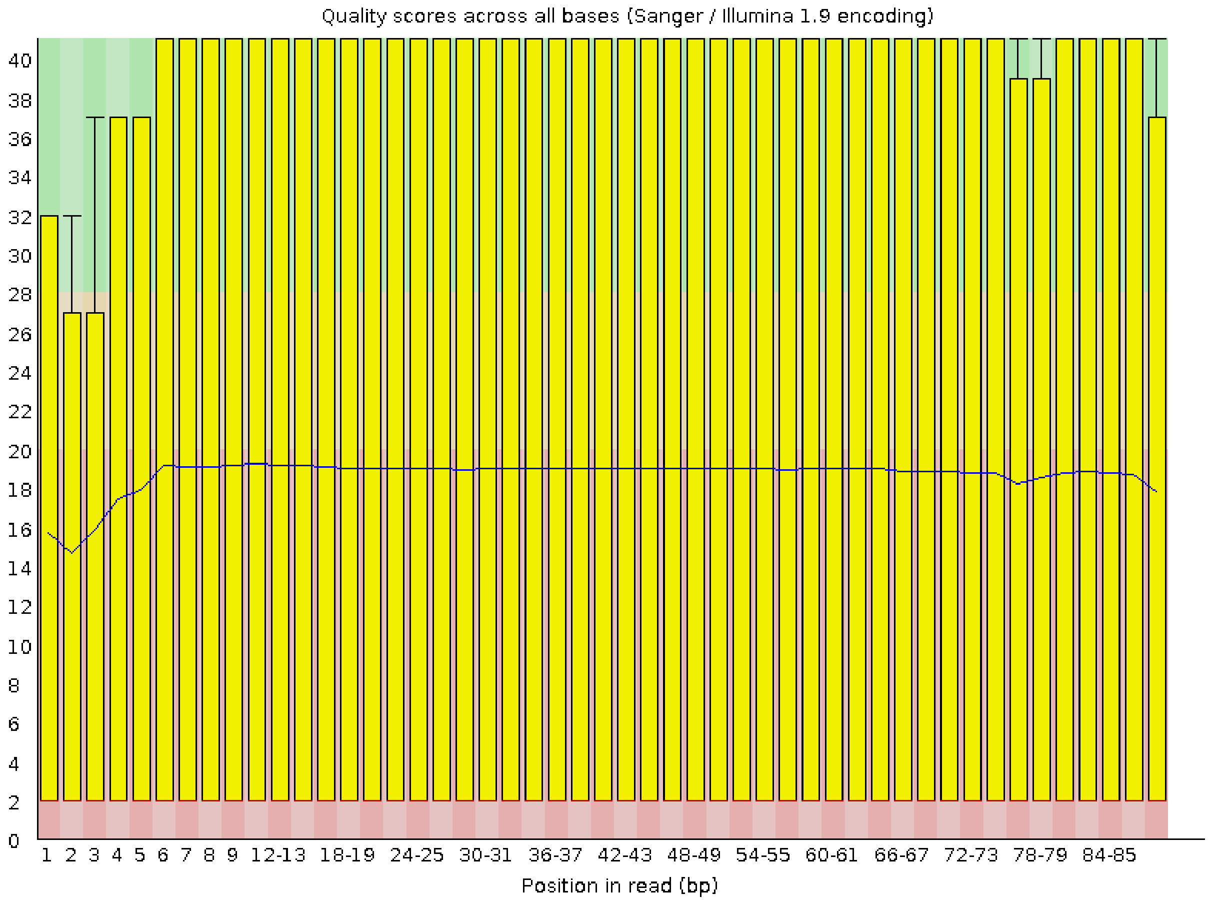
← Value Reported

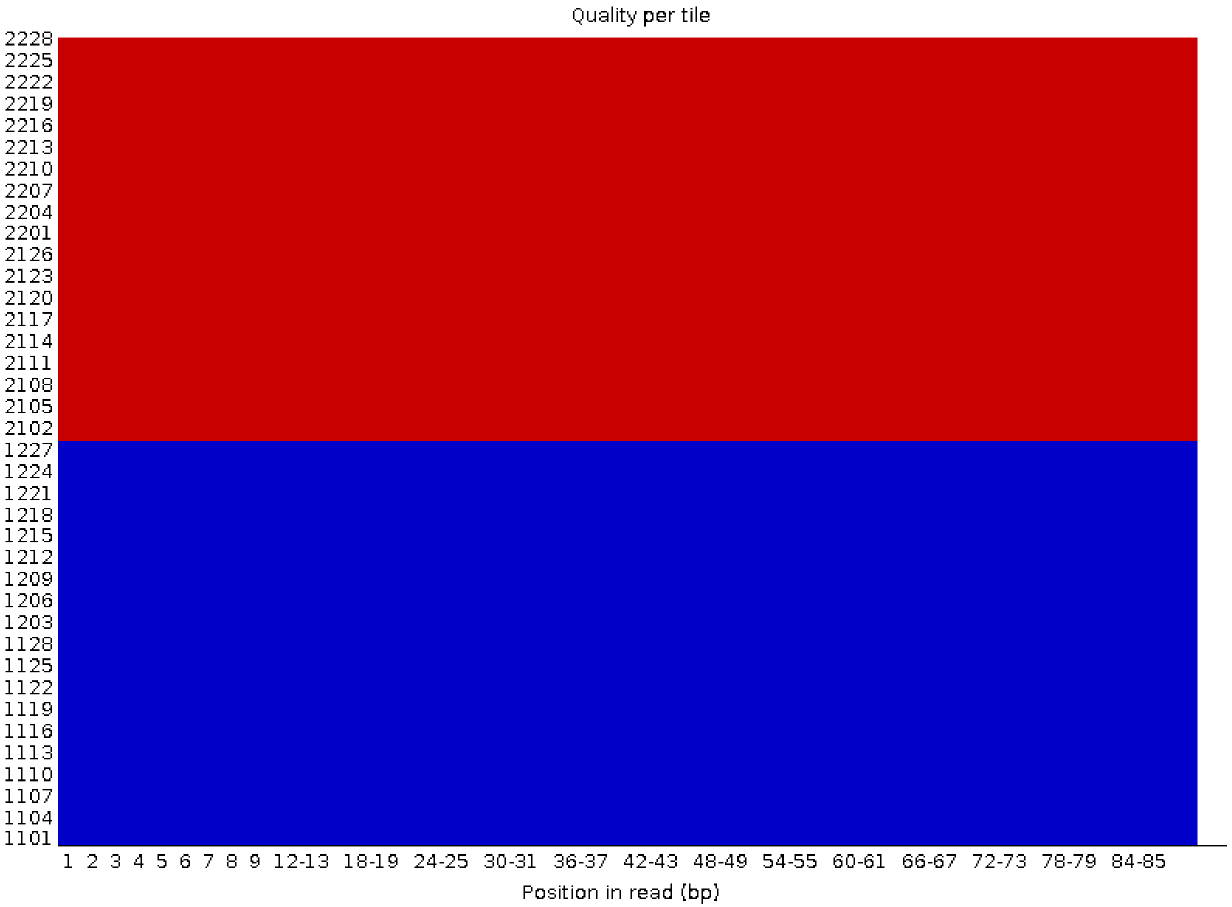
Course Data QC – Read1 (Barcodes)





Course Data QC – Read2 (RNA)


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Position in read (bp)


Quality per tile
Position in read (bp)