Swiss Institute of Bioinformatics

SINGLE-CELL TRANSCRIPTOMICS WITH R

Single-cell RNA-seq and Cell Ranger

Deepak Tanwar

November 12-14, 2025

Adapted from previous year courses

elixir SWITZERLAND

What is single-cell RNA-seq?

Single-cell RNA-seq (scRNA-seq) allows us to evaluate the transcriptome at the level of individual cells.

Why scRNAseq?

Why scRNAseq?

- To explore which cell types are present in a tissue

Why scRNAseq?

- To explore which cell types are present in a tissue
- To identify unknown/rare cell types or states

Why scRNAseq?

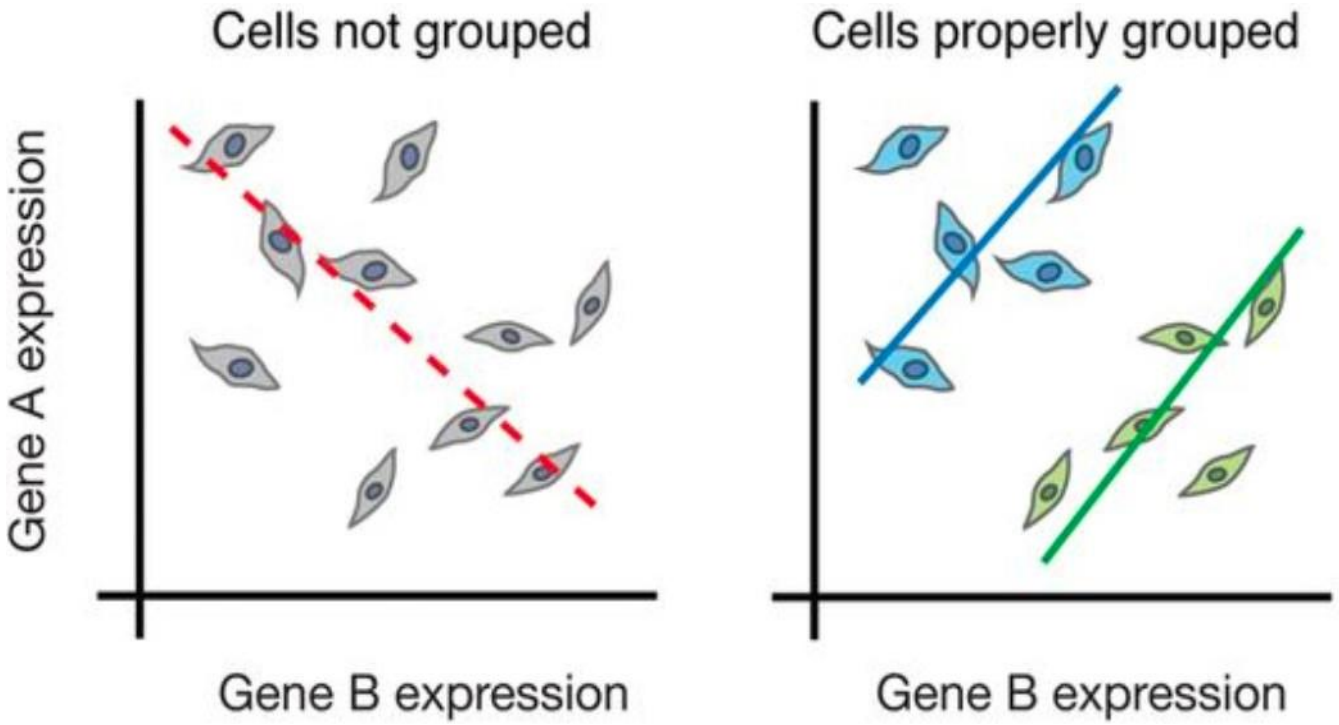
- To explore which cell types are present in a tissue
- To identify unknown/rare cell types or states
- To elucidate the changes in gene expression during differentiation processes or across time or states

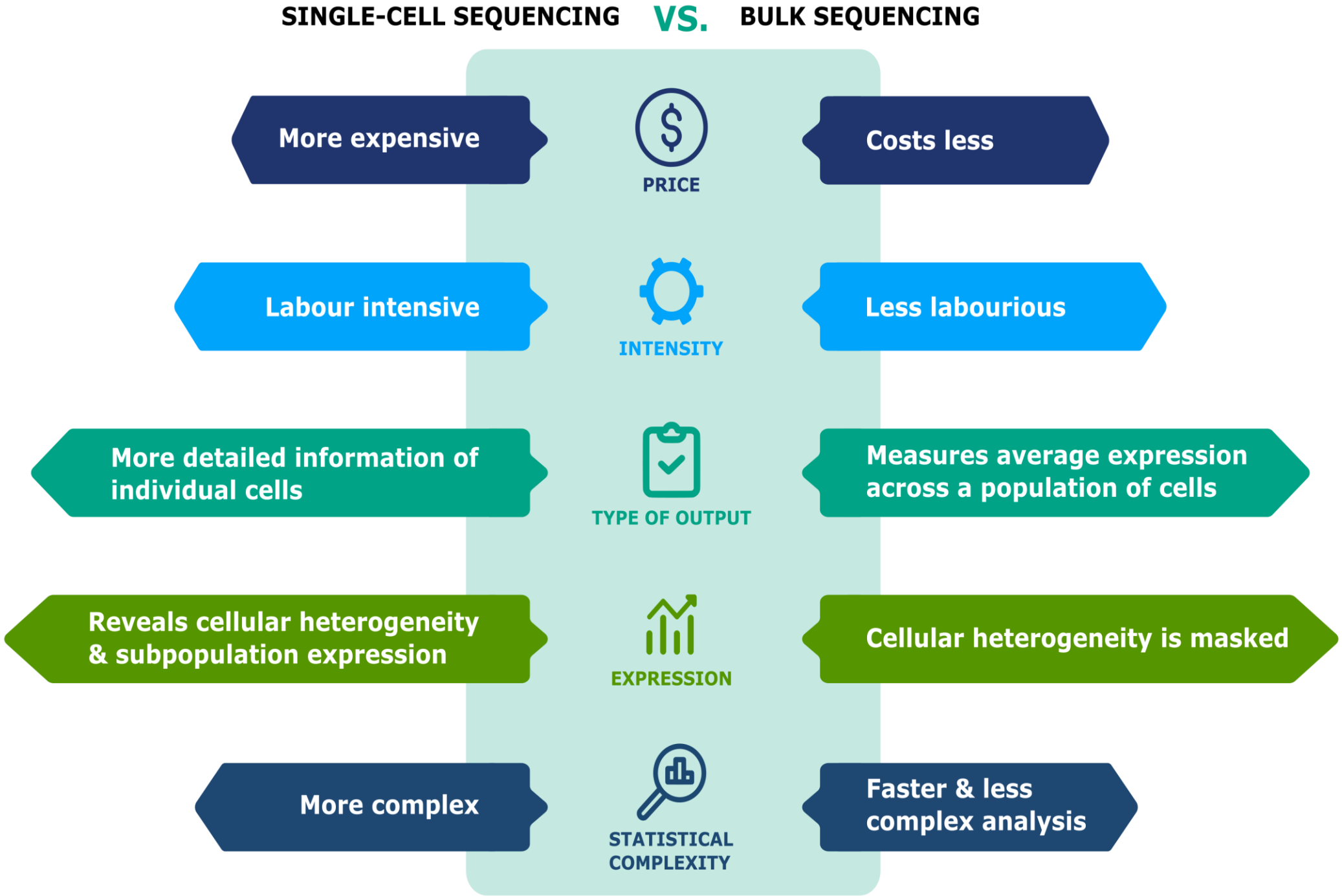
Why scRNAseq?

- To explore which cell types are present in a tissue
- To identify unknown/rare cell types or states
- To elucidate the changes in gene expression during differentiation processes or across time or states
- To identify genes that are differentially expressed in particular cell types between conditions (e.g. treatment or disease)

Bulk vs. Single-cell RNA-seq

- Bulk RNA-seq provides an overview of **average differences in gene expression**. For certain scenarios this has proven to be sufficient (i.e. biomarkers for cancer).
- Bulk RNA-seq is useful if you are not expecting or not concerned about cellular heterogeneity




SINGLE-CELL SEQUENCING
VS.
BULK SEQUENCING
More expensive
PRICE
Costs less
Labour intensive
INTENSITY
Less labourious
More detailed information of individual cells
TYPE OF OUTPUT
Measures average expression across a population of cells
Reveals cellular heterogeneity & subpopulation expression
EXPRESSION
Cellular heterogeneity is masked
More complex
STATISTICAL COMPLEXITY
Faster & less complex analysis

Challenges with scRNA-seq data

- Large volume of data
- Low depth of sequencing per cell (zero-inflation)
  - Often detecting only 10-50% of the transcriptome per cell

Challenges with scRNA-seq data

- Large volume of data
- Low depth of sequencing per cell (zero-inflation)
  - Often detecting only 10-50% of the transcriptome per cell

Increased complexity and richer datasets, means more room for misinterpretations and deriving wrong conclusions!

Recommendations

- Do not perform single-cell RNA-seq unless it is necessary for the experimental question of interest.

  Could you answer the question using bulk sequencing, which is simpler and less costly? Perhaps FACS sorting the samples could allow for bulk analysis?

Recommendations

- Understand the details of the experimental question you wish to address.

  Library preparation method and analysis workflow can vary based on the specific experiment and tissue

Recommendations

- Avoid technical sources of variability, if possible:
  - Discuss experimental design with experts prior to the initiation of the experiment
  - Isolate RNA from samples at same time
  - Prepare libraries at same time or alternate sample groups to avoid batch confounding

Experimental design

Replication, randomization and blocking

Be aware of confounding factors, e.g.:

- Person performing handling
- Reagents
- Sequencing lane/library

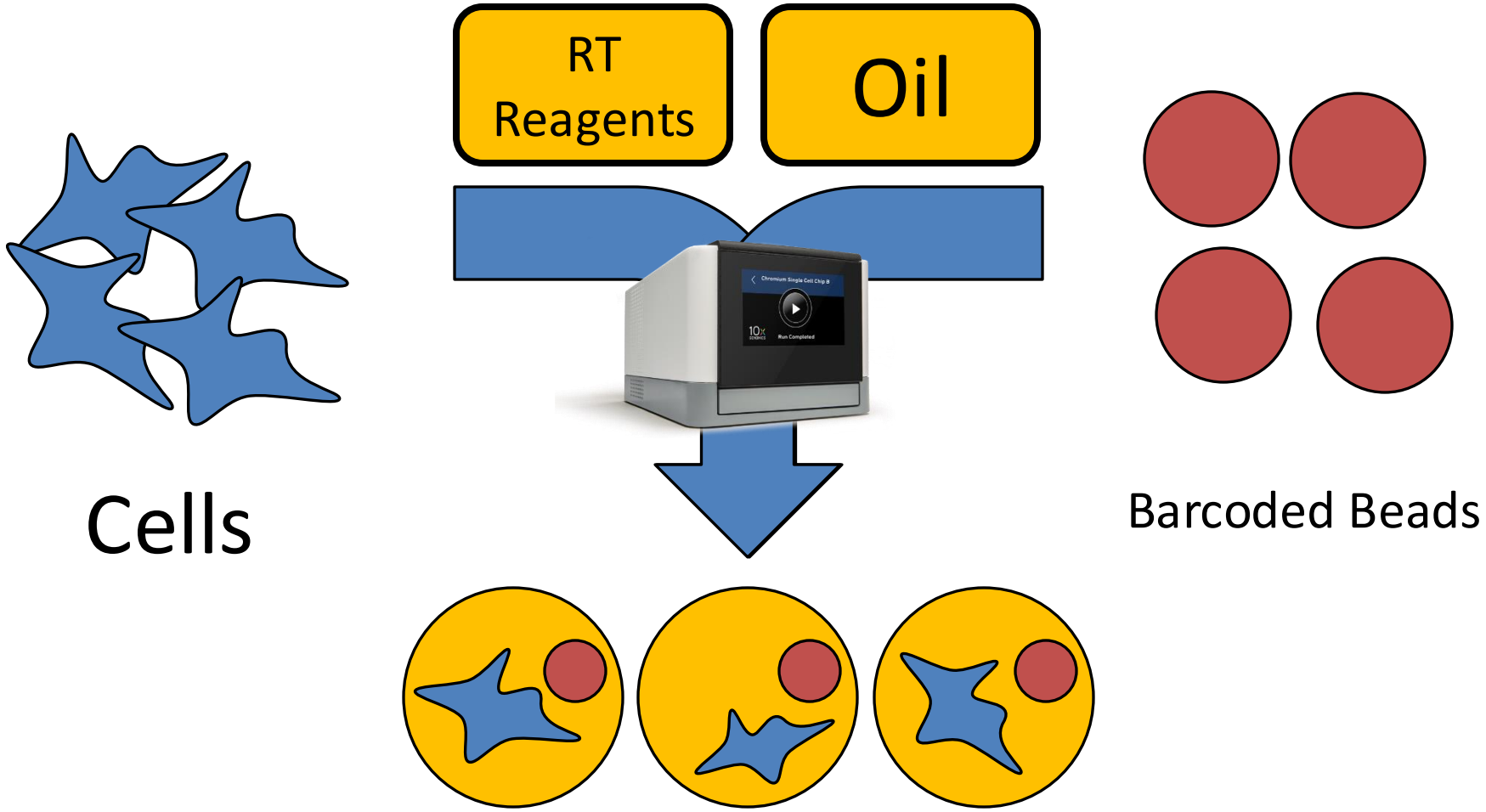
Record any factor for downstream correction: Taking notes = the best QC

Different cells of one replicate ≠ replicates!

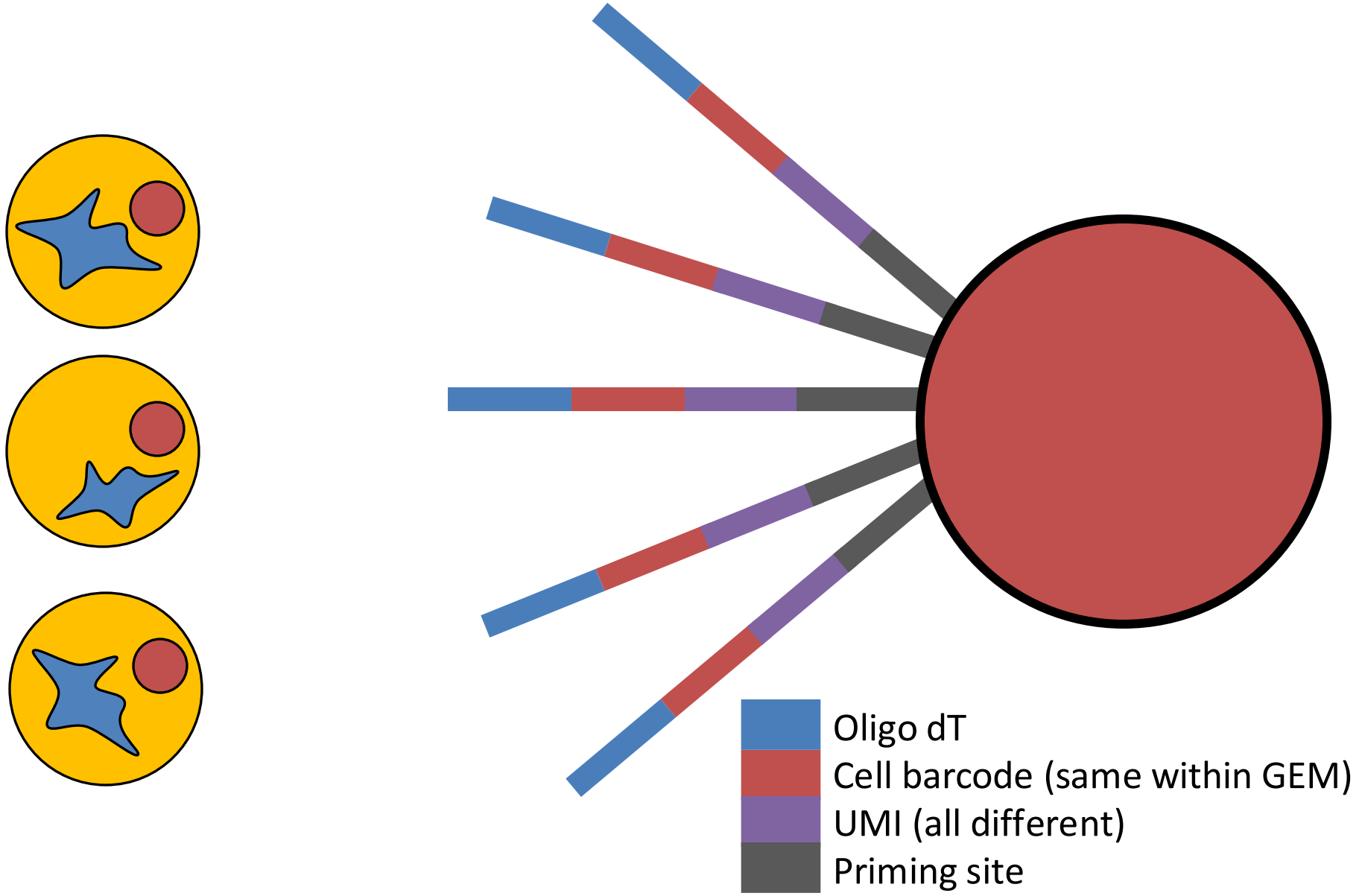
Further reading:

- https://doi.org/10.3389/fcell.2018.00108
- https://doi.org/10.1093/bib/bby007
- https://doi.org/10.1093/bfgp/elx035

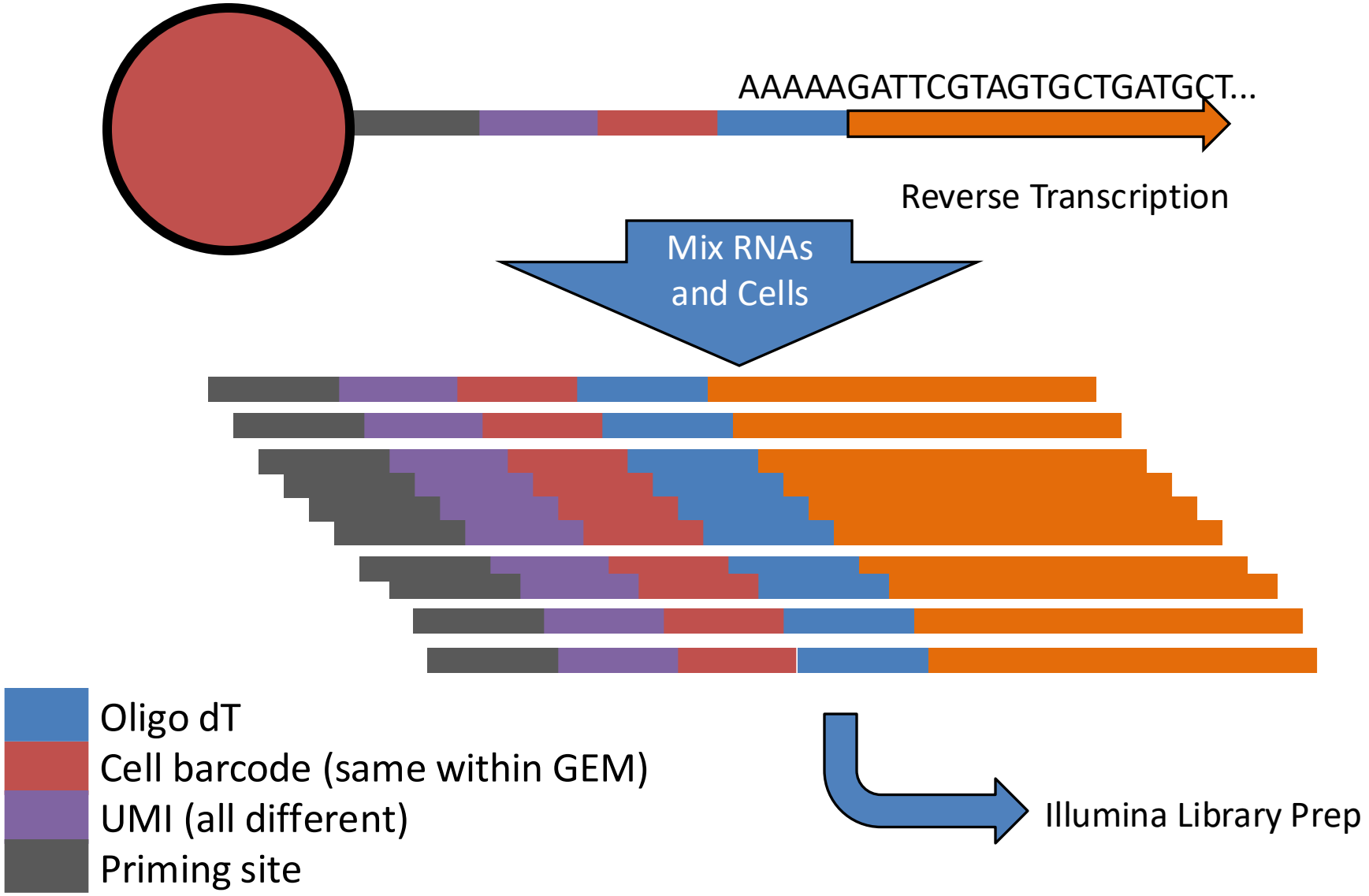
How 10X RNA-Seq Works?



How 10X RNA-Seq Works?



How 10X RNA-Seq Works?



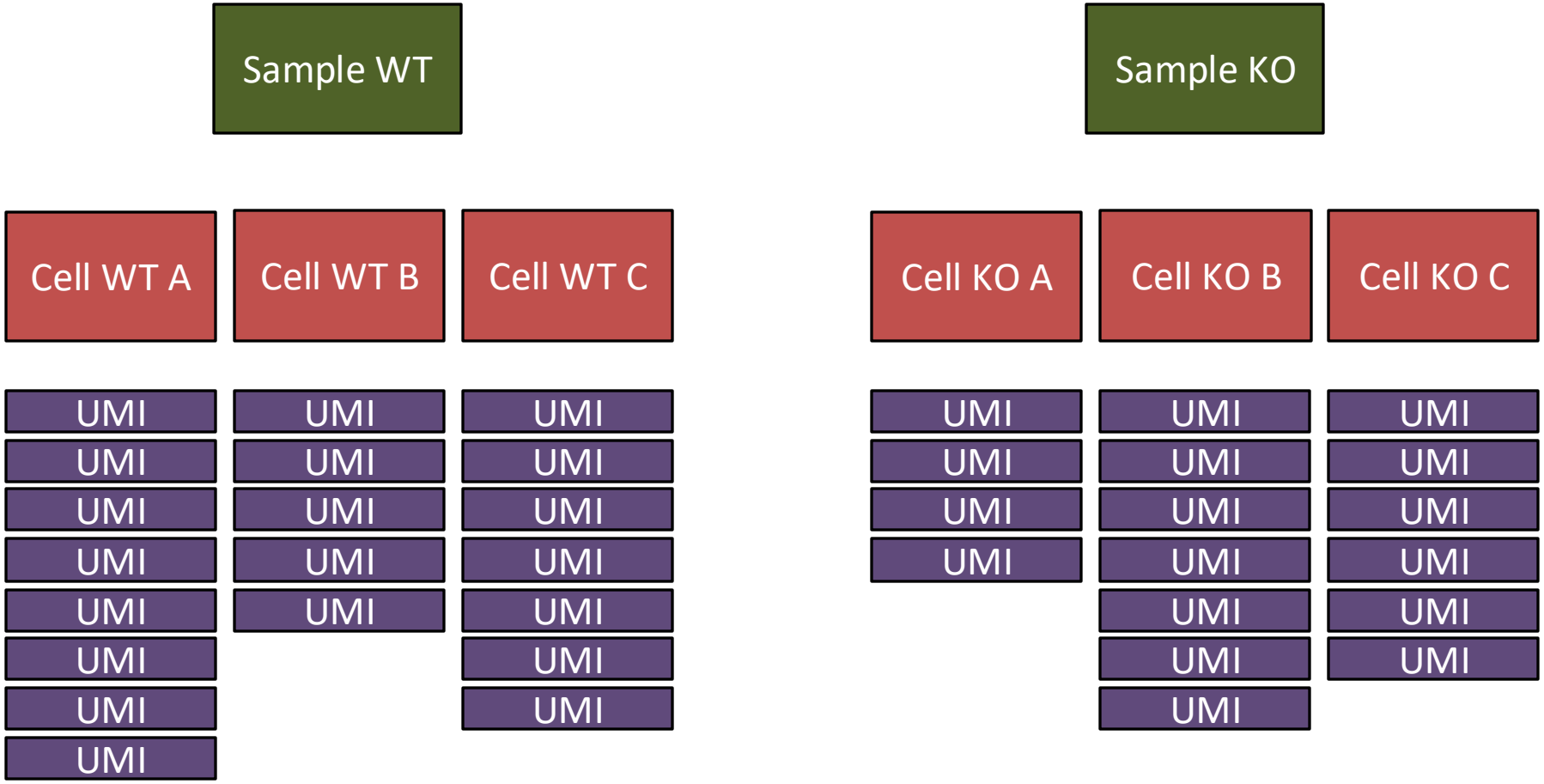
How 10X RNA-Seq Works?

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

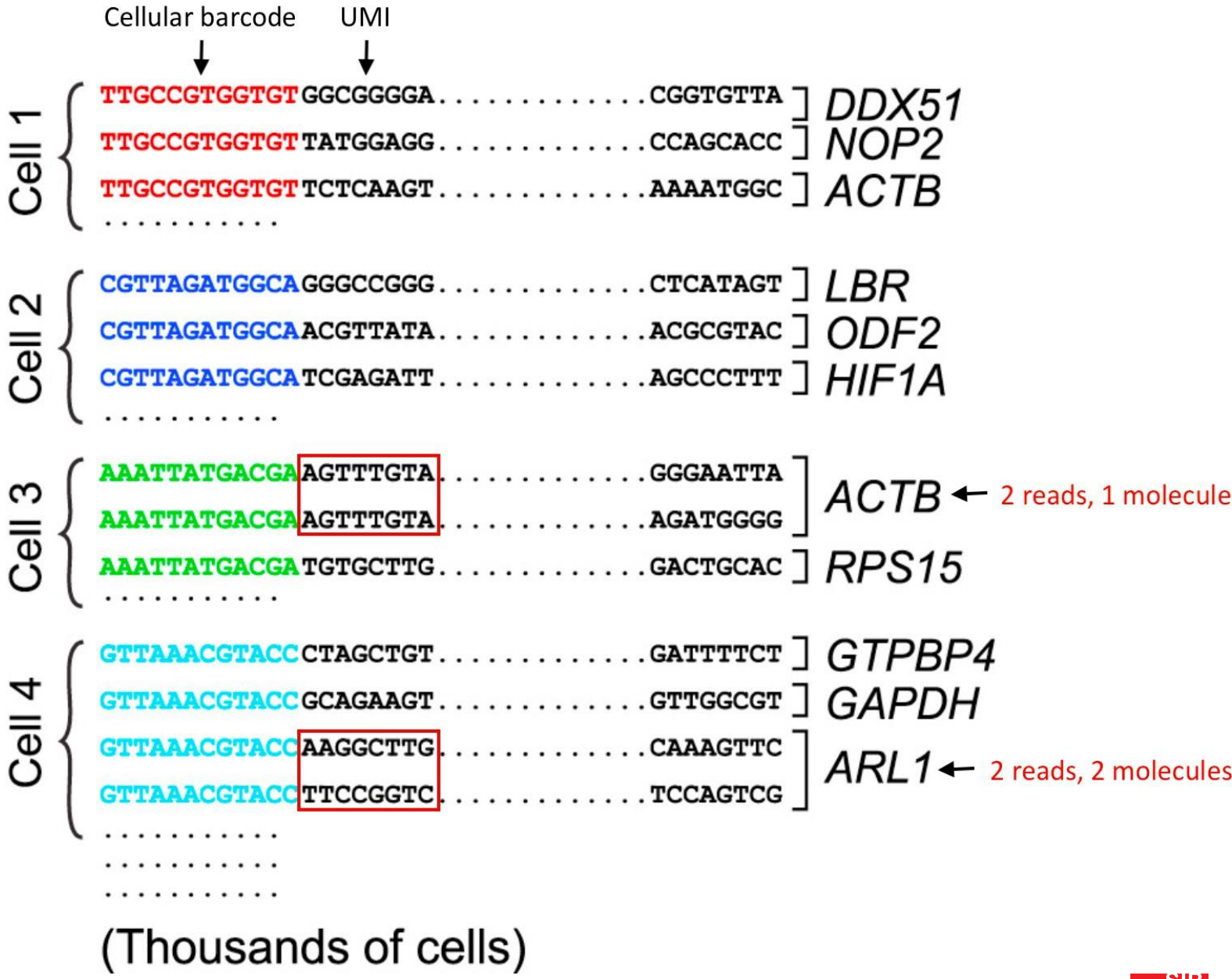


UMIs are finally related to genes to get per-gene counts

Understanding UMIs

- Reads with **different UMIs** mapping to the same transcript were derived from **different molecules** and are biological duplicates - each read should be counted.

- Reads with the **same UMI** originated from the **same molecule** and are technical duplicates - the UMIs should be collapsed to be counted as a single read.



Why count UMI (and not read alignments?)

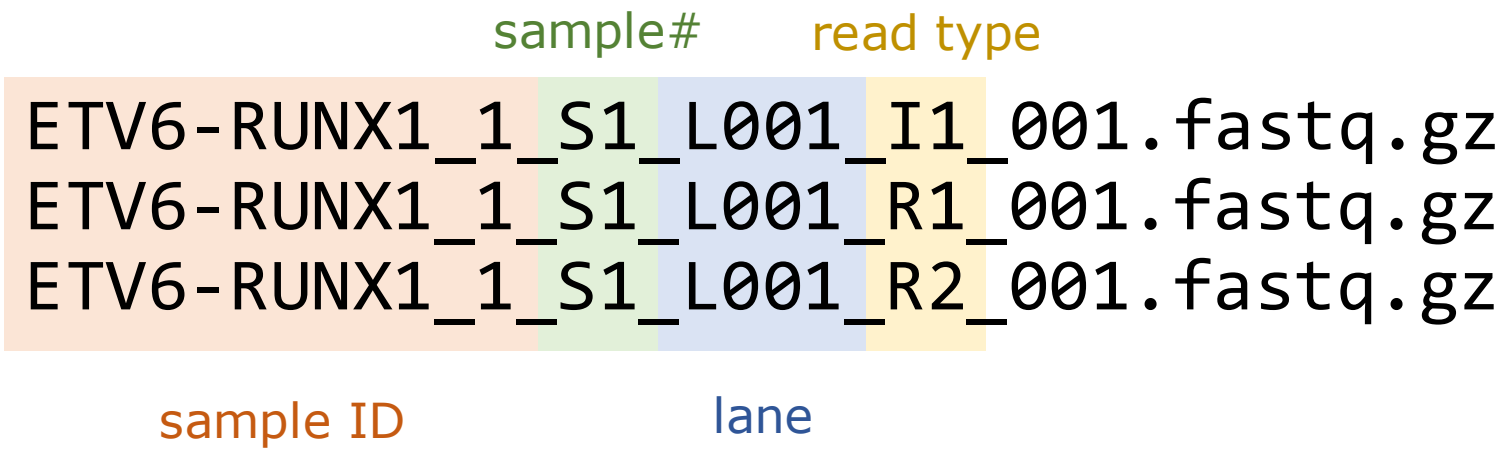
UMI: Unique Molecular Identifier:

- Identifies each molecule (i.e. sequence) uniquely

Molecules from a common PCR template -> carry the same UMI

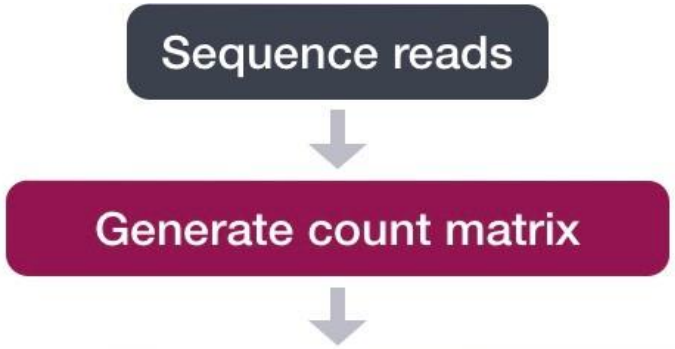
By counting UMI: correct for PCR duplicates

Sequencing output



Index Read (I1) - contains cell barcodes + UMI
Read 1 (R1) - contains cell barcode + UMI (same as I1 in 10x v2/v3)
Read 2 (R2) - contains cDNA insert (gene/transcript sequence)

From FASTQ to counts



Tools for this part of the workflow include Alevin, UMI-tools, and Cell Ranger (10X data). While each tool will do things slightly differently the steps below are common to all:

1. Assign reads to cell
2. Alignment
3. Quantification: # UMI/gene

The 10X Software Suite

Pipeline for mapping, filtering, QC and quantitation of libraries

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

Cell Ranger command

Cell Ranger Count (quantitates a single run)

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

Cell Ranger references

Human & mouse: download pre-built from 10x website

Other organisms: custom reference with **`cellranger mkref`**

extensive documentation:
https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger

Output files generated by Cell Ranger

`web_summary.html` - Web format QC report

`raw/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

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

Cellranger report

ETV6-RUNX1_1

Alerts

The analysis detected ⚠ 1 warning.

	Alert	Value	Detail
⚠	Fraction of RNA read bases with Q-score >= 30 is low	59.4%	Fraction of RNA read bases with Q-score >= 30 should be above 65%. A lower fraction might indicate poor sequencing quality. This is Read 1 for the Single Cell 3' v1 chemistry and Single Cell 5' paired end, Read 2 for the Single Cell 3' v2/v3 chemistry and Single Cell 5' R2-only)

Summary | Analysis

3,091
Estimated Number of Cells

68,259
Mean Reads per Cell

1,717
Median Genes per Cell

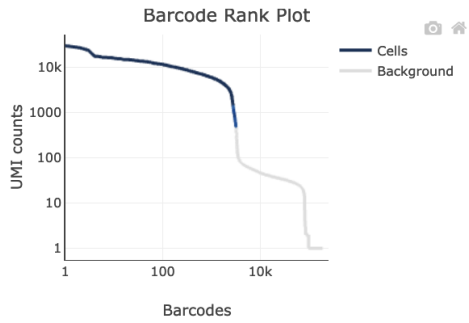
Sequencing

Number of Reads	210,987,037
Number of Short Reads Skipped	0
Valid Barcodes	98.2%
Valid UMIs	100.0%
Sequencing Saturation	84.4%
Q30 Bases in Barcode	96.4%
Q30 Bases in RNA Read	59.4%
Q30 Bases in UMI	96.5%

Mapping

Reads Mapped to Genome	95.8%
Reads Mapped Confidently to Genome	92.9%
Reads Mapped Confidently to Intergenic Regions	5.2%
Reads Mapped Confidently to Intronic Regions	25.5%
Reads Mapped Confidently to Exonic Regions	62.2%
Reads Mapped Confidently to Transcriptome	58.2%
Reads Mapped Antisense to Gene	1.2%

Cells



Estimated Number of Cells	3,091
Fraction Reads in Cells	91.1%
Mean Reads per Cell	68,259
Median Genes per Cell	1,717
Total Genes Detected	18,334
Median UMI Counts per Cell	4,825

Sample

Sample ID	ETV6-RUNX1_1
Sample Description	
Chemistry	Single Cell 3' v2
Include introns	False
Reference Path	…nger/refdata-cellranger-GRCh38-3.0.0
Transcriptome	GRCh38-3.0.0
Pipeline Version	cellranger-6.0.1

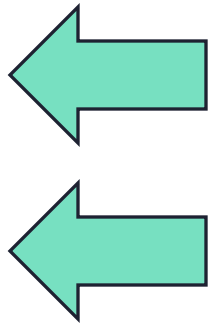
How many cells do you have?

Start by looking at the quality of the base calls in the barcodes

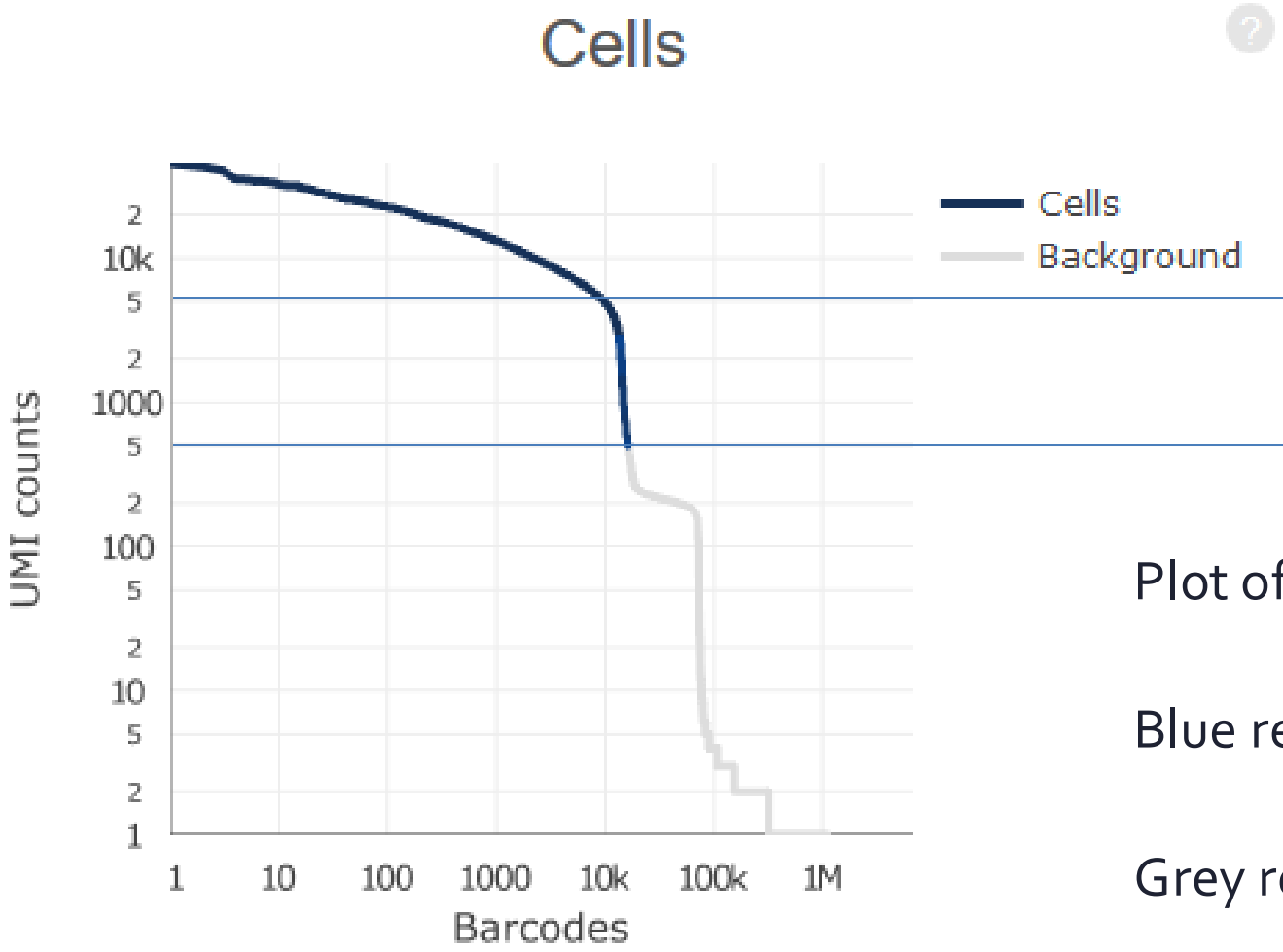
Bad calls will lead to inaccurate cell assignments

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



5000 reads per cell. 10k cells

500 reads per cell. 15k 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!!!

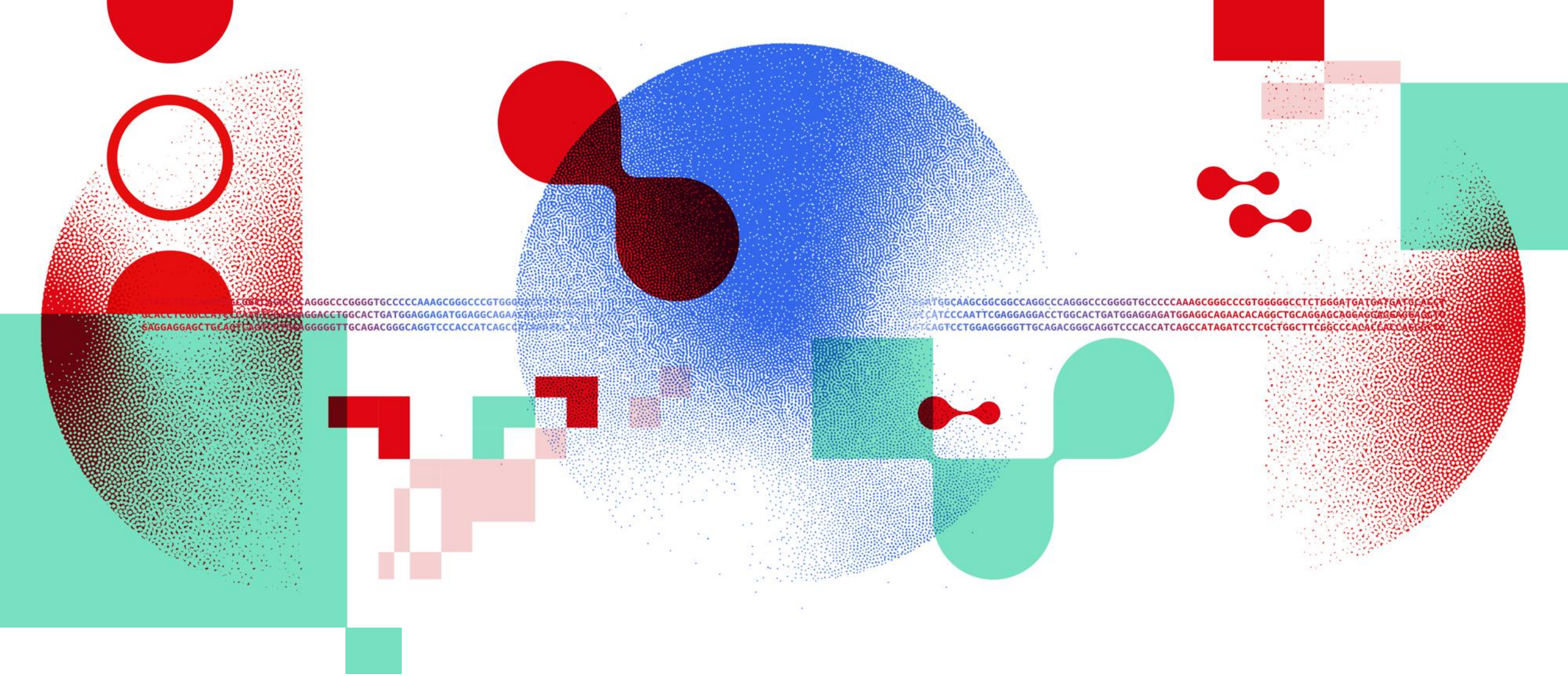
How much data do you have per cell?

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



Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss