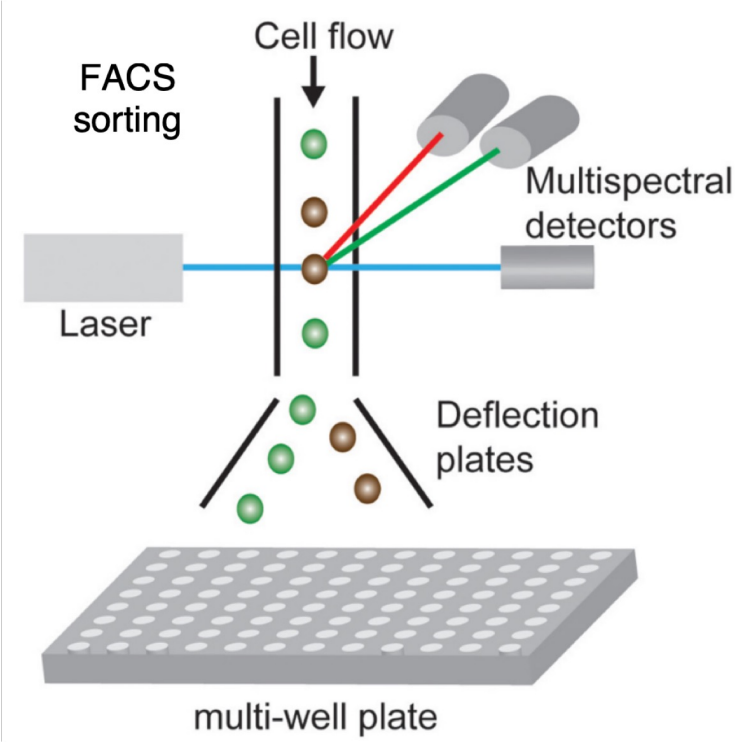
Introduction to Cell Ranger

Single Cell Transcriptomics in Python

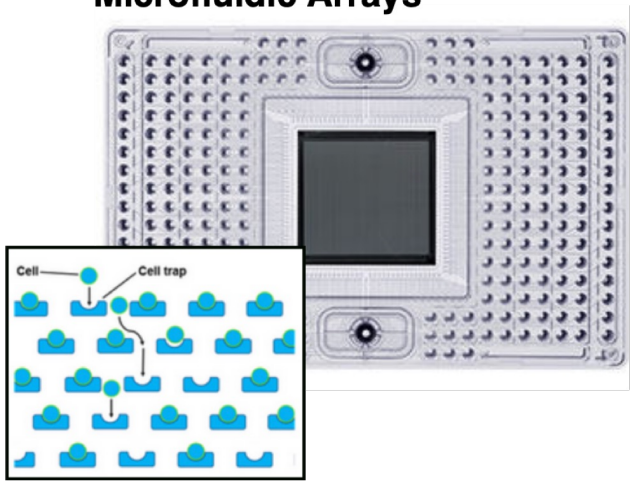
Alex Lederer

General overview of scRNA-seq technologies

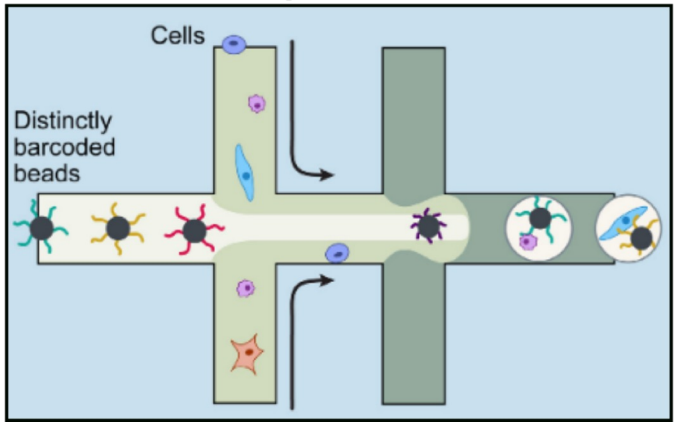
Microtitre Plates



Microfluidic Arrays

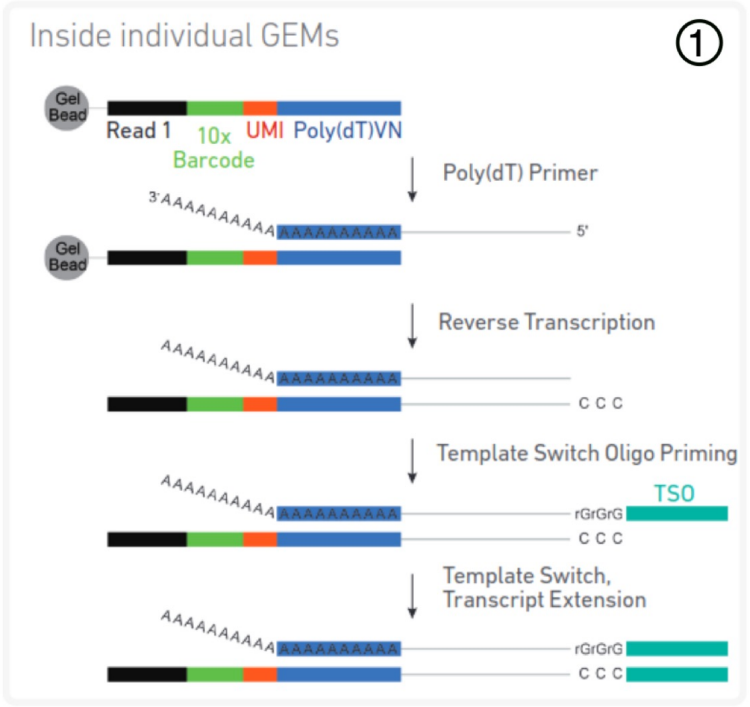


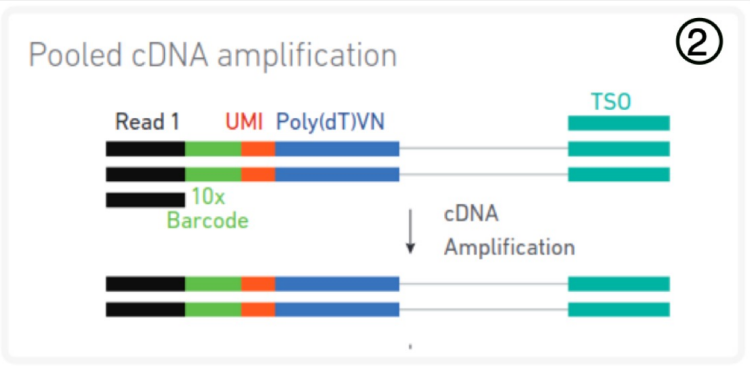
Microfluidic Droplets

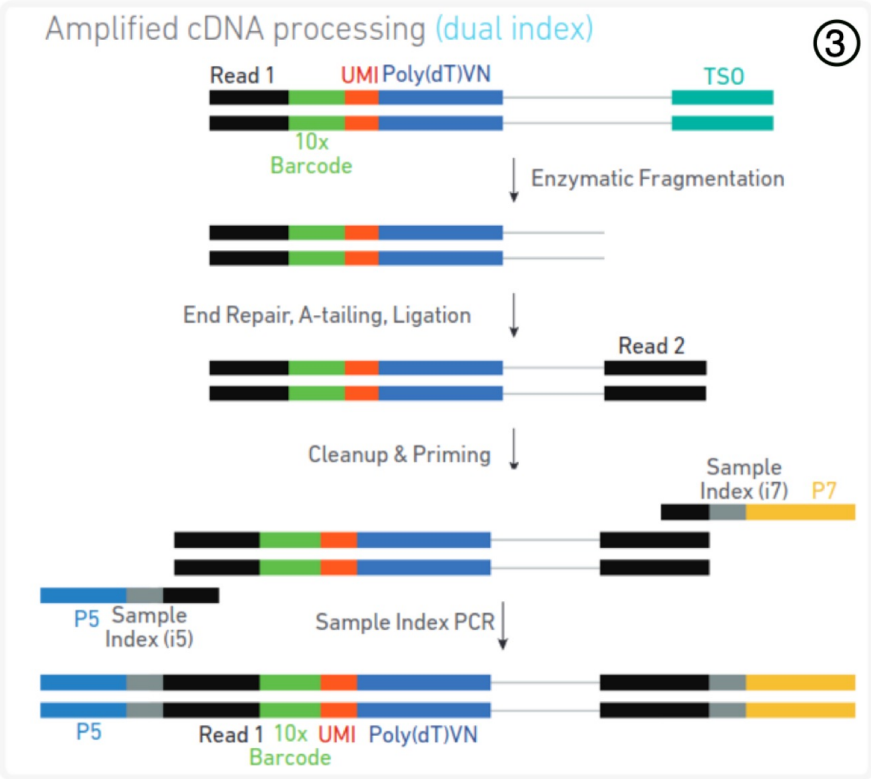


https://www.singlecellcourse.org/introduction-to-single-cell-rna-seq.html

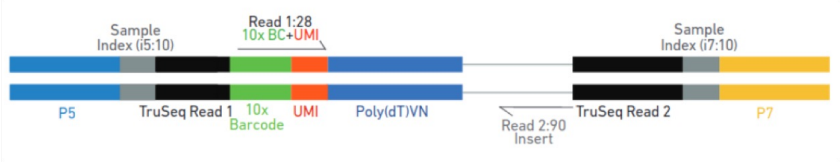
10X Genomics Single-Cell RNA-sequencing protocol overview







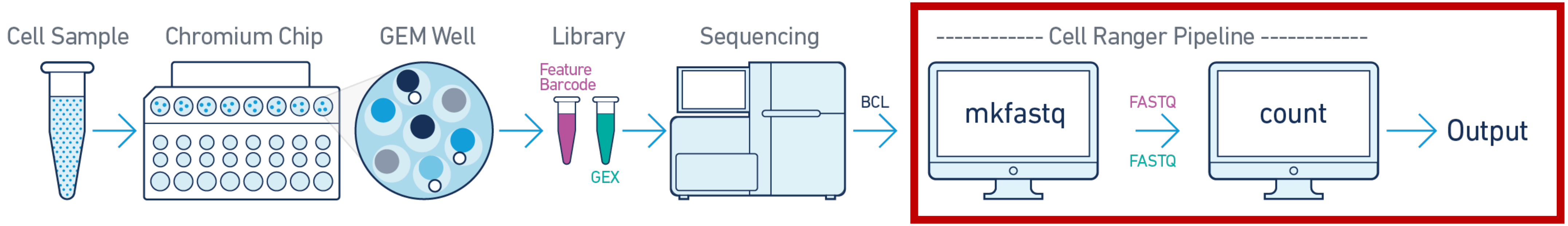
Final library



https://www.singlecellcourse.org/introduction-to-single-cell-rna-seq.html

Question: Which piece of information does Cell Ranger use to account for amplification bias?

Cell Ranger converts raw FASTQs into a cell-by-gene count matrix



FASTQ file

```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACTCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCTGAAGTACACAGAGAACAATTCTTAAAAAA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE<AEEEEEEE
```

Count matrix

	Gene0	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6	Gene7	Gene8	Gene9	...
Cell0	31.0	15.0	91.0	45.0	49.0	99.0	32.0	55.0	29.0	51.0	...
Cell1	54.0	18.0	95.0	99.0	88.0	77.0	81.0	73.0	60.0	63.0	...
Cell2	9.0	9.0	54.0	45.0	53.0	86.0	40.0	75.0	88.0	14.0	...
Cell3	90.0	85.0	28.0	11.0	92.0	99.0	2.0	44.0	61.0	18.0	...

FASTQ files explained

Link: https://support.illumina.com/bulletins/2016/04/fastq-files-explained.html

1. A sequence identifier with information about the sequencing run and the cluster. The exact contents of this line vary by based on the BCL to FASTQ conversion software used.
2. The sequence (the base calls; A, C, T, G and N).
3. A separator, which is simply a plus (+) sign.
4. The base call quality scores. These are Phred +33 encoded, using ASCII characters to represent the numerical quality scores.

Here is an example of a single entry in a R1 FASTQ file:

```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACTCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCTGAAGTACACAGAGAACAATTCTTAAAAAA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE<AEEEEEEE
```

Cell Ranger *count*

- One of multiple functions offered as part of the Cell Ranger software (*mkfastq*, *multi*, *aggr*).
- Cell ranger count operates in two main steps: (https://www.10xgenomics.com/support/software/cell-ranger/latest/analysis/running-pipelines/cr-gex-count)

Step 1: Align each sequencing read to a reference transcriptome

Convert a FASTQ file…

```
        TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
```

Into a BAM/SAM file (https://www.metagenomics.wiki/tools/samtools/bam-sam-file-format)

```
K00369:86:HKJFVBBXY:6:2218:15453:10616  0       chr1    629330  3       104M    *       0       0       TCATACACCTCCTATGAAAAAACTTCCTACCACTCACCCTAGCATTACTTATATGATATGTCTCCATACCCATTACAATCTCCAGCATTCCC
CCTCAAACCTAA AAFFFJA-7A<7FAJ<JJJJJJFJ-<-AJF7FJFFJJ<<FF<FJJJJJJ<JJFJJFFJJJFJFJJJJFJJFFJAJJFFJJJFJFJJFFFJJJJJJJJJJFFJJJJ       NH:i:2  HI:i:1  AS:i:102        nM:i:0  RG:Z:RPE_34_ctrl_rep1:0:1:HKJF
VBBXY:4   RE:A:I  xf:i:0  CR:Z:AGTGACTTCTAGGCCG   CY:Z:AAFFFJFJJJJJJJJJ   CB:Z:AGTGACTTCTAGGCCG-1  UR:Z:CATTACCAGTAA       UY:Z:JJJJJJJJJJJJ       UB:Z:CATTACCAGTAA
K00369:86:HKJFVBBXY:4:2214:17401:7732   0       chr1    629330  3       104M    *       0       0       TCATACACCTCCTATGAAAAAACTTCCTACCACTCACCCTAGCATTACTTATATGATATGTCTCCATACCCATTACAATCTCCAGCATTCCC
CCTCAAACCTAA AAFAFJJJJJJJFJJJFJJJJJJFJFJFJFJJJJJJJJJJJJFFJJJJJJJJJJJJJFJFJJJFFJJJJJFJJJJJFJJJJJJJJ<JJJFFJJJ<FJJ7FFJJJJJ       NH:i:2  HI:i:1  AS:i:102        nM:i:0  RG:Z:RPE_34_ctrl_rep1:0:1:HKJF
VBBXY:4   RE:A:I  xf:i:0  CR:Z:ATCGATGTCAGCTTCC   CY:Z:AAFFFJJJJJJJJJJJ   CB:Z:ATCGATGTCAGCTTCC-1  UR:Z:TTTCTGGATCCG       UY:Z:JJJJJJJJJJJJ       UB:Z:TTTCTGGATCCG
```

Cell Ranger *count*

- One of multiple functions offered as part of the Cell Ranger software (*mkfastq*, *multi*, *aggr*).
- Cell ranger count operates in two main steps: (https://www.10xgenomics.com/support/software/cell-ranger/latest/analysis/running-pipelines/cr-gex-count)

Step 1: Align each sequencing read to a reference transcriptome

Convert a FASTQ file…

```
        TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
```

Into a BAM/SAM file (https://www.metagenomics.wiki/tools/samtools/bam-sam-file-format)

```
K00369:86:HKJFVBBXY:6:2218:15453:10616  0       chr1    629330  3       104M    *       0       0       TCATACACCTCCTATGAAAAAACTTCCTACCACTCACCCTAGCATTACTTATATGATATGTCTCCATACCCATTACAATCTCCAGCATTCCC
CCTCAAACCTAA AAFFFJA-7A<7FAJ<JJJJJJFJ-<-AJF7FJFFJJ<<FF<FJJJJJJ<JJFJJFFJJJFJFJJJJFJJFFJAJJFFJJJFJFJJFFFJJJJJJJJJJFFJJJJ        NH:i:2  HI:i:1  AS:i:102        nM:i:0  RG:Z:RPE_34_ctrl_rep1:0:1:HKJF
VBBXY:4   RE:A:I  xf:i:0  CR:Z:AGTGACTTCTAGGCCG   CY:Z:AAFFFJFJJJJJJJJJ   CB:Z:AGTGACTTCTAGGCCG-1  UR:Z:CATTACCAGTAA       UY:Z:JJJJJJJJJJJJ       UB:Z:CATTACCAGTAA
K00369:86:HKJFVBBXY:4:2214:17401:7732   0       chr1    629330  3       104M    *       0       0       TCATACACCTCCTATGAAAAAACTTCCTACCACTCACCCTAGCATTACTTATATGATATGTCTCCATACCCATTACAATCTCCAGCATTCCC
CCTCAAACCTAA AAFAFJJJJJJJFJJJFJJJJJJFJFJFJFJJJJJJJJJJJJFFJJJJJJJJJJJJJFJFJJJFFJJJJJFJJJJJFJJJJJJJJ<JJJFFJJJ<FJJ7FFJJJJJ        NH:i:2  HI:i:1  AS:i:102        nM:i:0  RG:Z:RPE_34_ctrl_rep1:0:1:HKJF
VBBXY:4   RE:A:I  xf:i:0  CR:Z:ATCGATGTCAGCTTCC   CY:Z:AAFFFJJJJJJJJJJJ   CB:Z:ATCGATGTCAGCTTCC-1  UR:Z:TTTCTGGATCCG       UY:Z:JJJJJJJJJJJJ       UB:Z:TTTCTGGATCCG
```

Step 2: Count cell barcodes and the number of unique molecular identifiers (UMIs) per gene/cell

BAM/SAM file → **Feature-Count matrix**

Which cell barcodes are "real" and which come from "empty" droplets or sequencing errors?

5k_pbmc_protein_v3 - 5k Peripheral blood mononuclear cells (PBMCs) from a healthy donor

Summary | Analysis

5,247
Estimated Number of Cells

28,918
Mean Reads per Cell

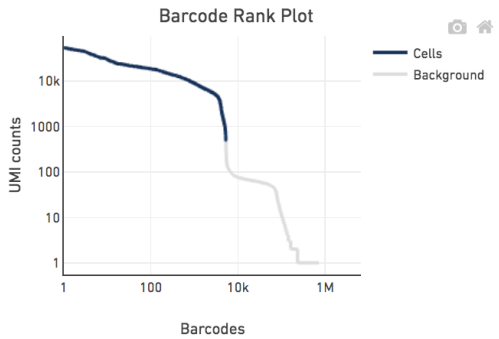
Sequencing

Number of Reads	151,731,342
Valid Barcodes	97.5%
Valid UMIs	99.9%
Sequencing Saturation	52.4%
Q30 Bases in Barcode	95.8%
Q30 Bases in RNA Read	91.9%
Q30 Bases in Sample Index	89.8%
Q30 Bases in UMI	95.4%

Mapping

Reads Mapped to Genome	94.3%
Reads Mapped Confidently to Genome	88.4%
Reads Mapped Confidently to Intergenic Regions	6.8%
Reads Mapped Confidently to Intronic Regions	25.0%
Reads Mapped Confidently to Exonic Regions	56.7%
Reads Mapped Confidently to Transcriptome	53.2%
Reads Mapped Antisense to Gene	1.3%

Cells



Estimated Number of Cells	5,247
Fraction Reads in Cells	87.7%
Mean Reads per Cell	28,918
Median Genes per Cell	1,644
Total Genes Detected	20,822
Median UMI Counts per Cell	5,496

Sample

Sample ID	5k_pbmc_protein_v3
Sample Description	5k Peripheral blood mononuclear cells (PBMCs) from a healthy donor
Chemistry	Single Cell 3' v3
Transcriptome	GRCh38-3.0.0
Pipeline Version	3.1.0

How do I know if my sample is any good?

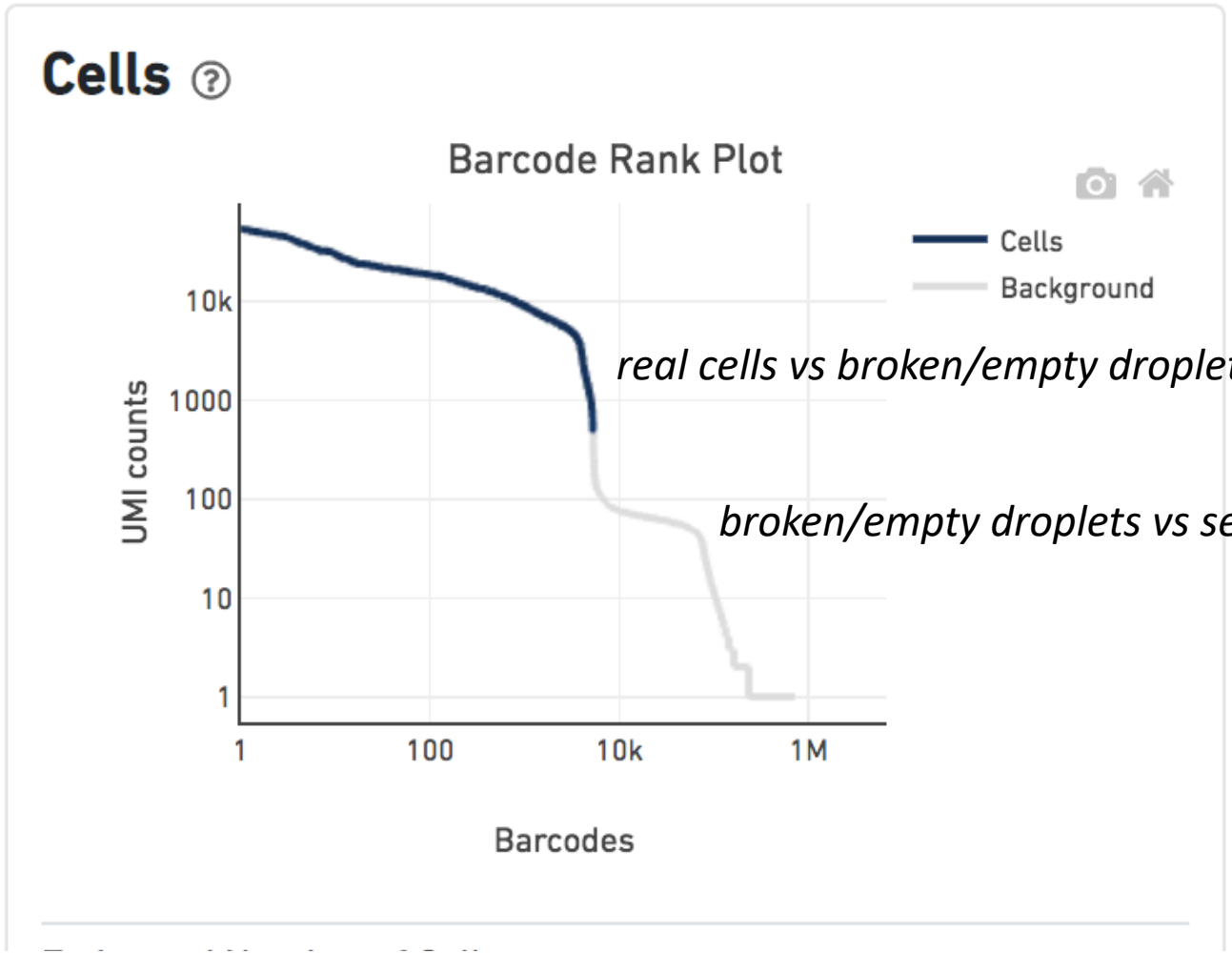
Sample Preparation
- Cell isolation and tissue dissociation
- **Potential issues:** *What if the tissue is not fully dissociated? Or agitated too much?*

Cell Encapsulation
- Gel Bead-In-Emulsions (GEMs) Formation: cells are encapsulated in droplets with gel beads coated with **barcoded cell-specific primers** and **molecule-specific UMIs**
- **Potential issues:** multiple cells per barcode, droplets without any cells, ambient RNA

Reverse Transcription
- Converts mRNA into cDNA within each droplet.
- **Potential issues:** degradation of unstable RNA molecules

Post-Emulsion Breakage Processing
- Emulsion Breakage and cDNA Cleanup: The emulsions are broken, and the pooled cDNA is cleaned up, usually using magnetic beads. The cleaned-up cDNA is PCR-amplified to generate sufficient material for library preparation (PCA amplification) and sequencing.
- **Potential issues:** incorrect size selection with magnetic beads; amplification of primers or other small DNA fragments

The barcode rank plot



real cells vs broken/empty droplets

broken/empty droplets vs sequencing errors

Question: Which piece of information are we **unable** to infer from the barcode rank plot?