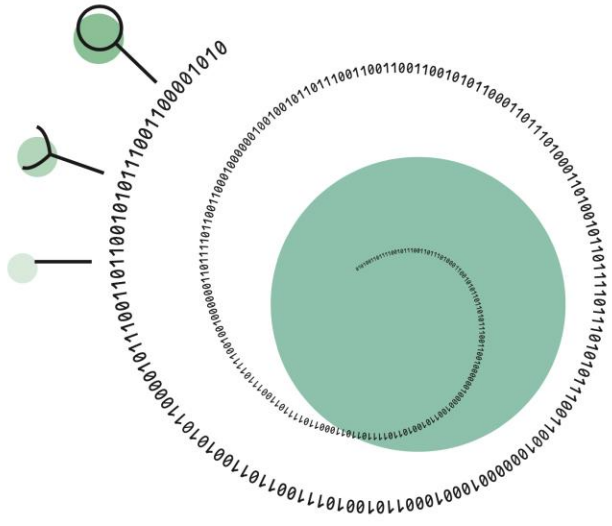


Single Cell Data Analysis for Beginners
Block course
22nd – 26th of September 2025

Lisa Buchauer
Professor of Systems Biology of Infectious Diseases
Department of Infectious Diseases and Intensive Care
Charité - Universitätsmedizin Berlin

Anika Neuschulz
Postdoctoral Researcher
Division of Translational Immunology
BIH@Charité

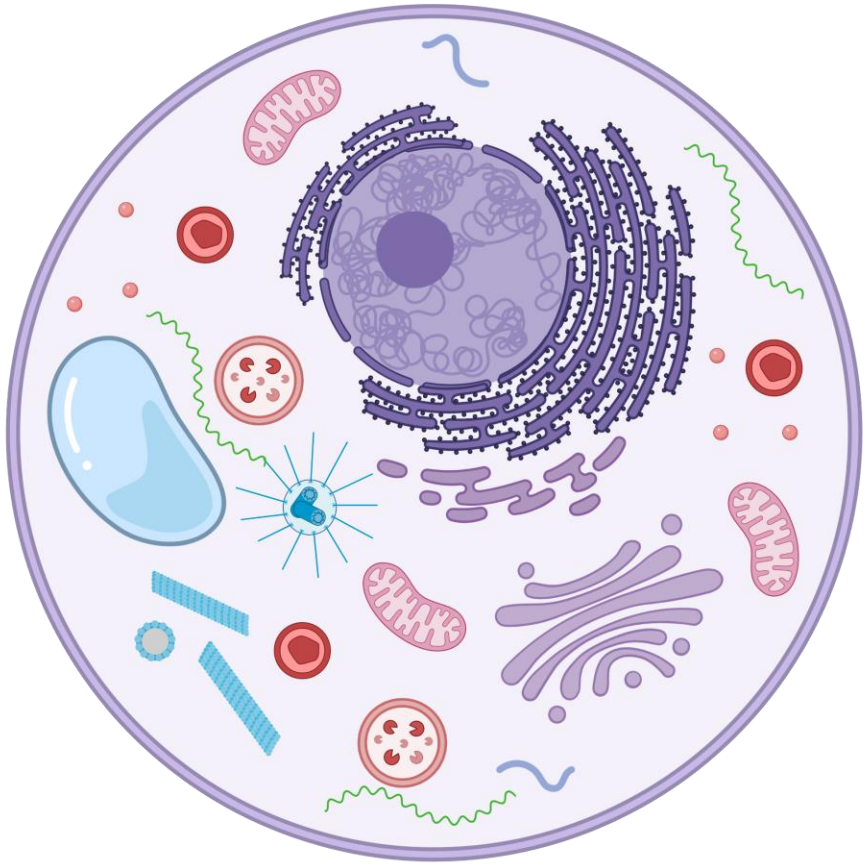
Welcome

Schedule day 1

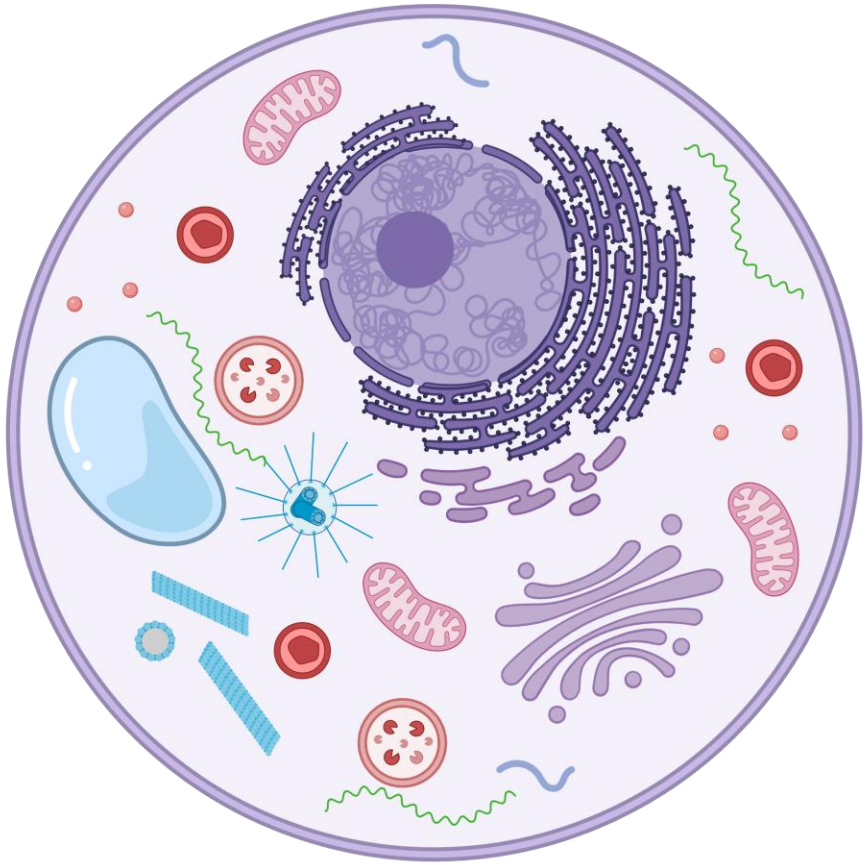
9:30 – 10:00	Opening, Icebreaker, Introductions
10:00-10:30	Intro to single cell sequencing resulting raw data types, bcl + fastq
10:30-11:30	Mini-intro how to cluster at Charité Live alignment session with cellranger, introduction of main parameters, inspection of output
11:00-12:00	inspect cellranger QC reports, assign sections to groups, study and present to everyone
12:00-13:00	Lunch break
13:00-14:30	Set-up environments (R or python)
14:30-16:00	Basic data wrangling introduction

Single-Cell Experimental Background

Why do we sequence RNA?



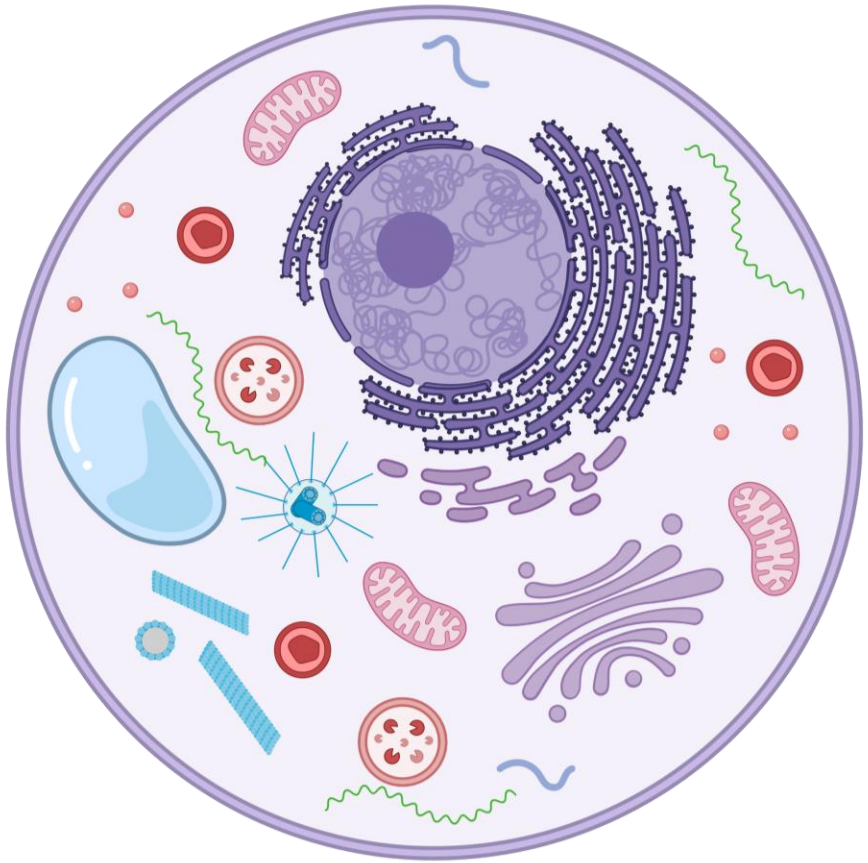
Why do we sequence RNA?



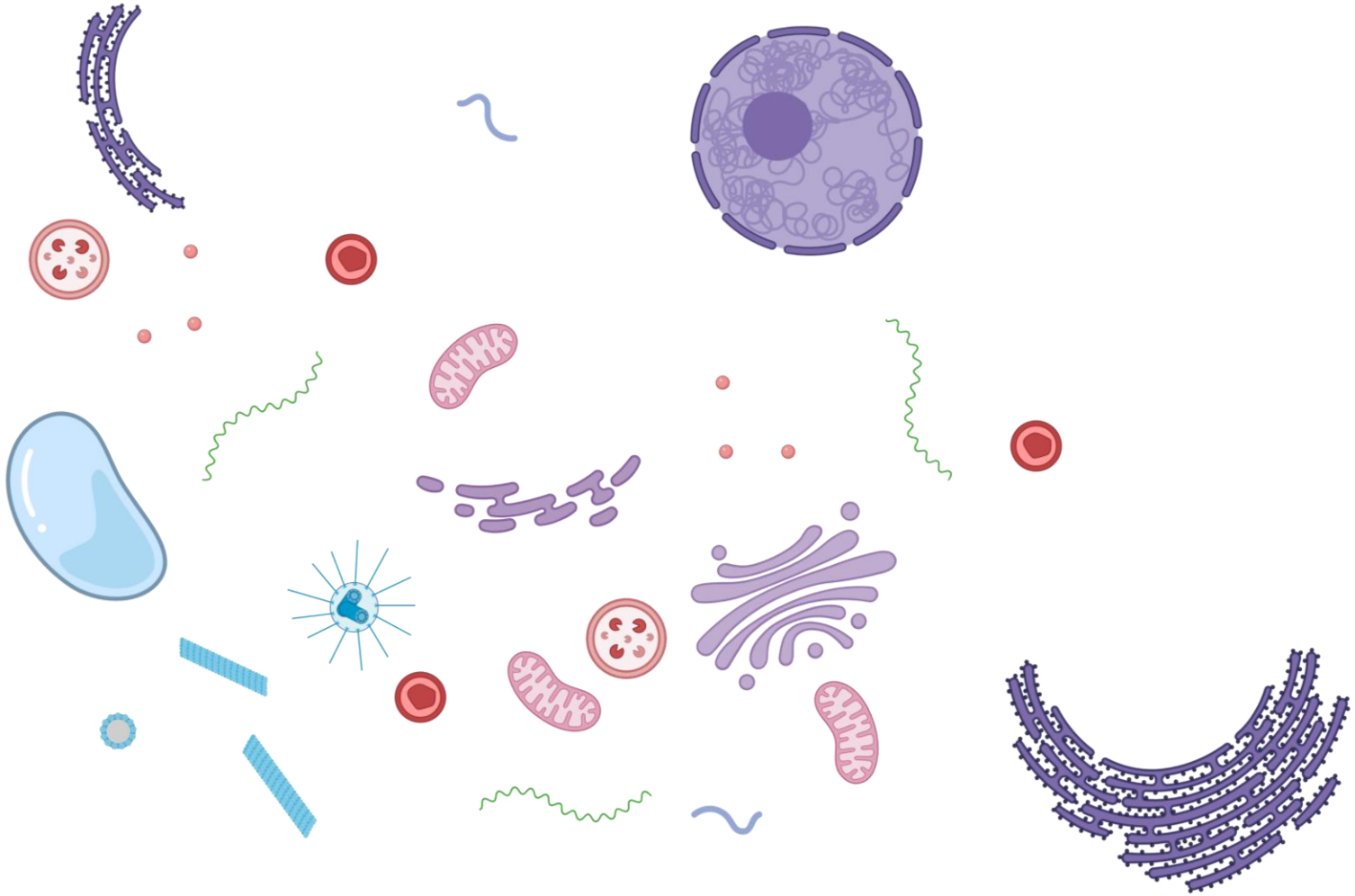
understand gene expression

→ the cell’s identity
→ life functions
→ reactions to the environment

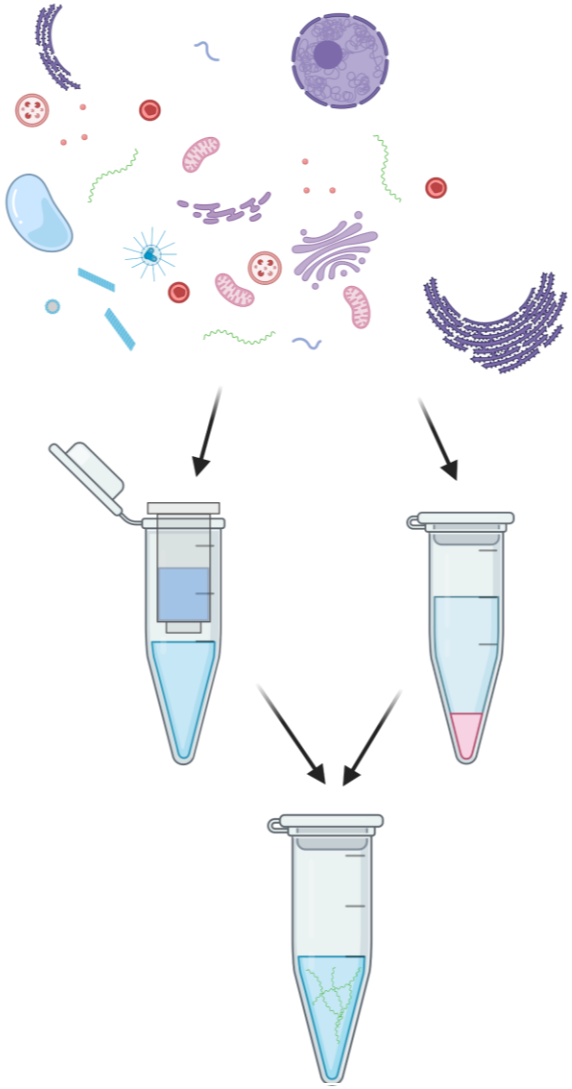
How to get to the RNA?



How to get to the RNA?



How to get to the RNA?



RNA (often enriched for mRNA)

Reverse transcription

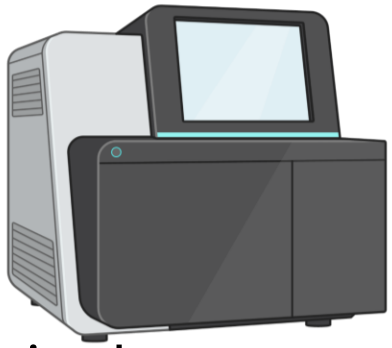
cDNA

We can sequence DNA!

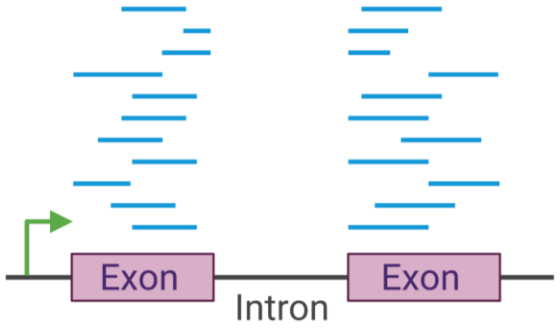
cDNA

fragment cDNA in shorter pieces

add sequencing adapters, amplify, enrich for appropriate length,…

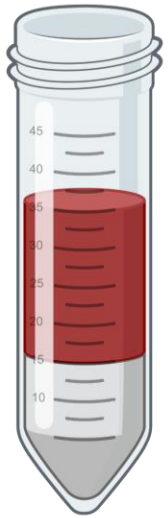


amplify fragments in the sequencer while observing which bases are used

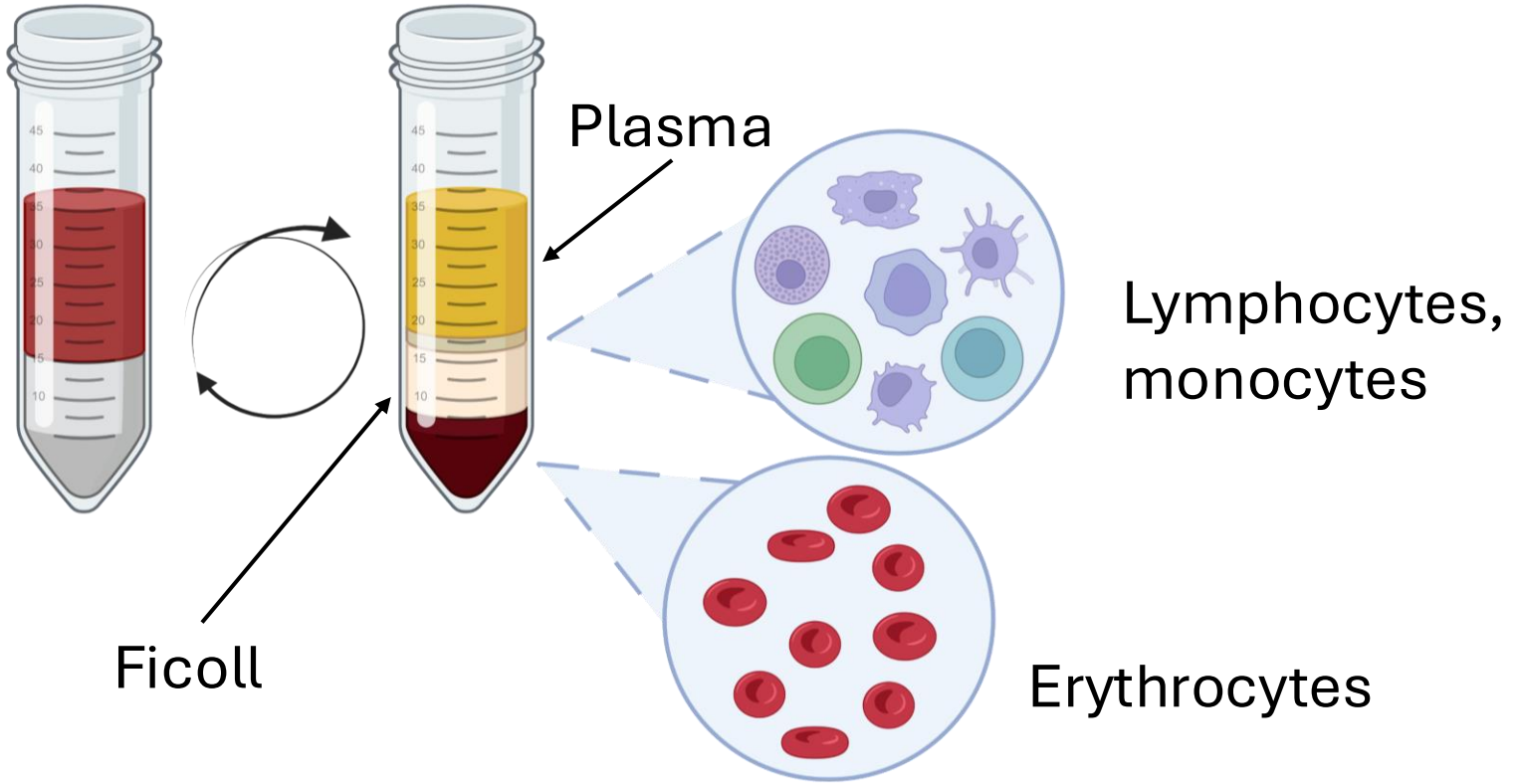


align observed base sequences to reference genome

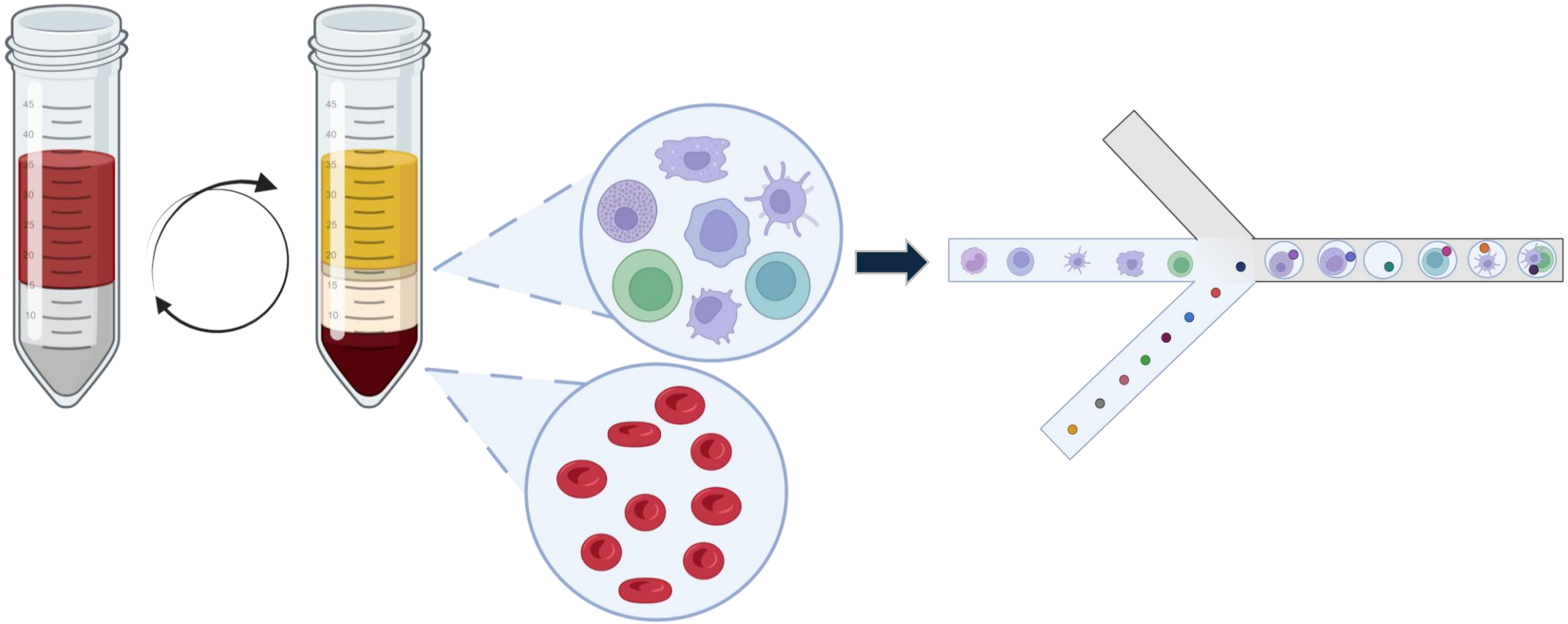
How do we sequence immune cells from blood?



How do we sequence immune cells from blood?

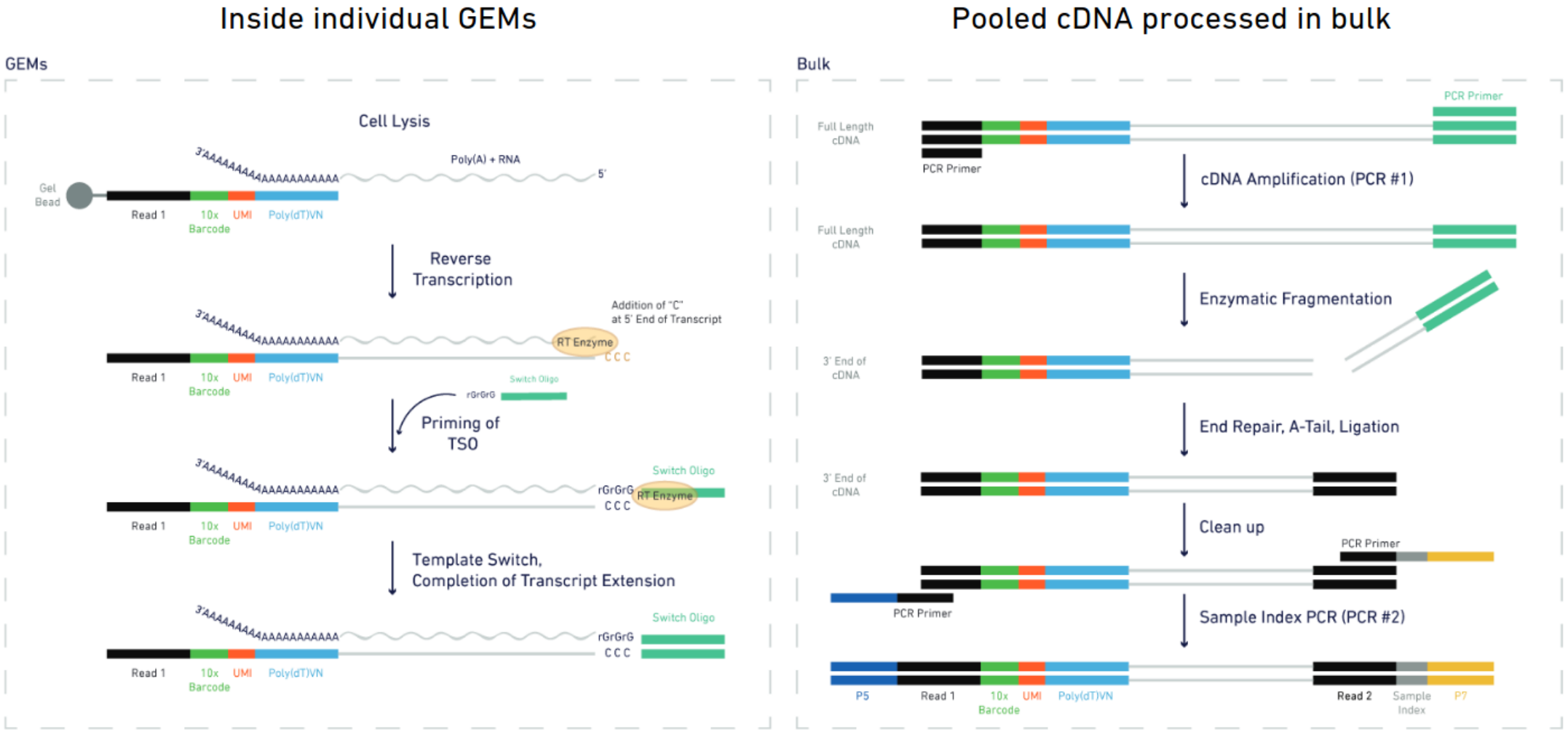


How do we sequence immune cells from blood?



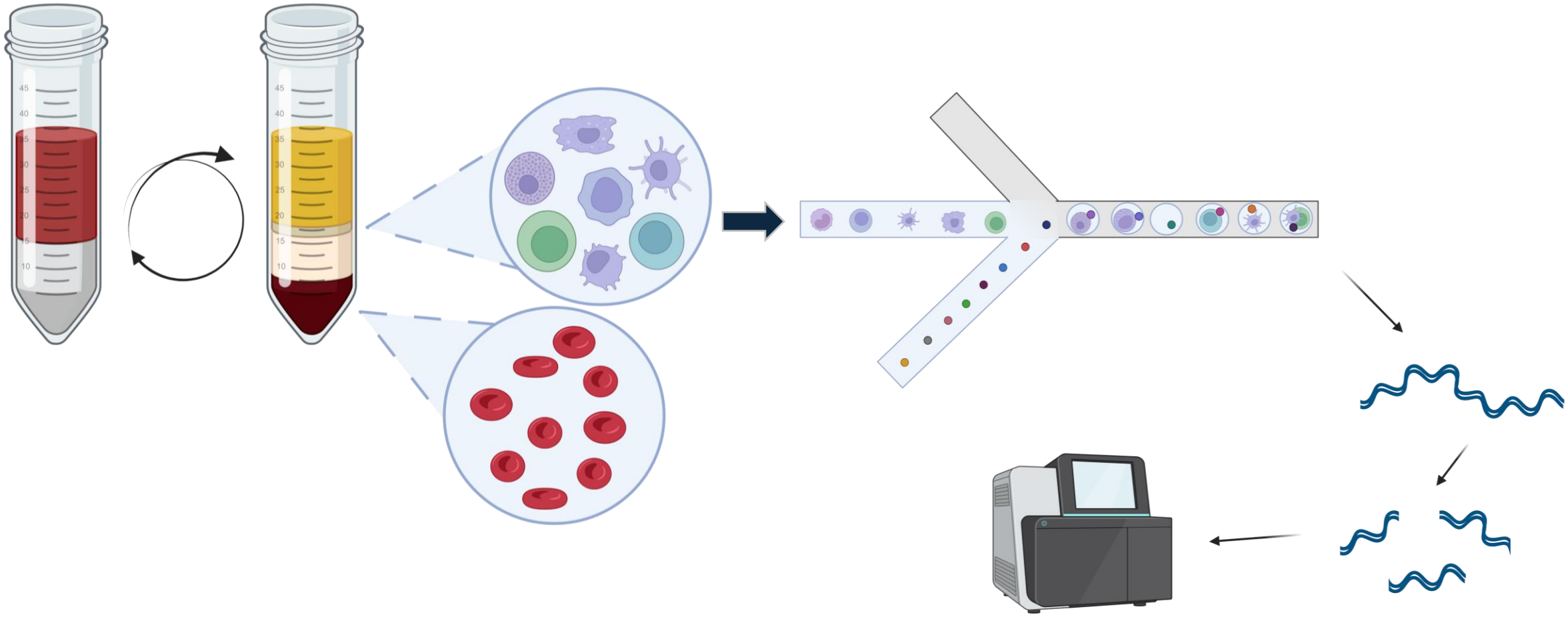
Example: 10x Genomics 3' library preparation

(how our first training data set was generated)



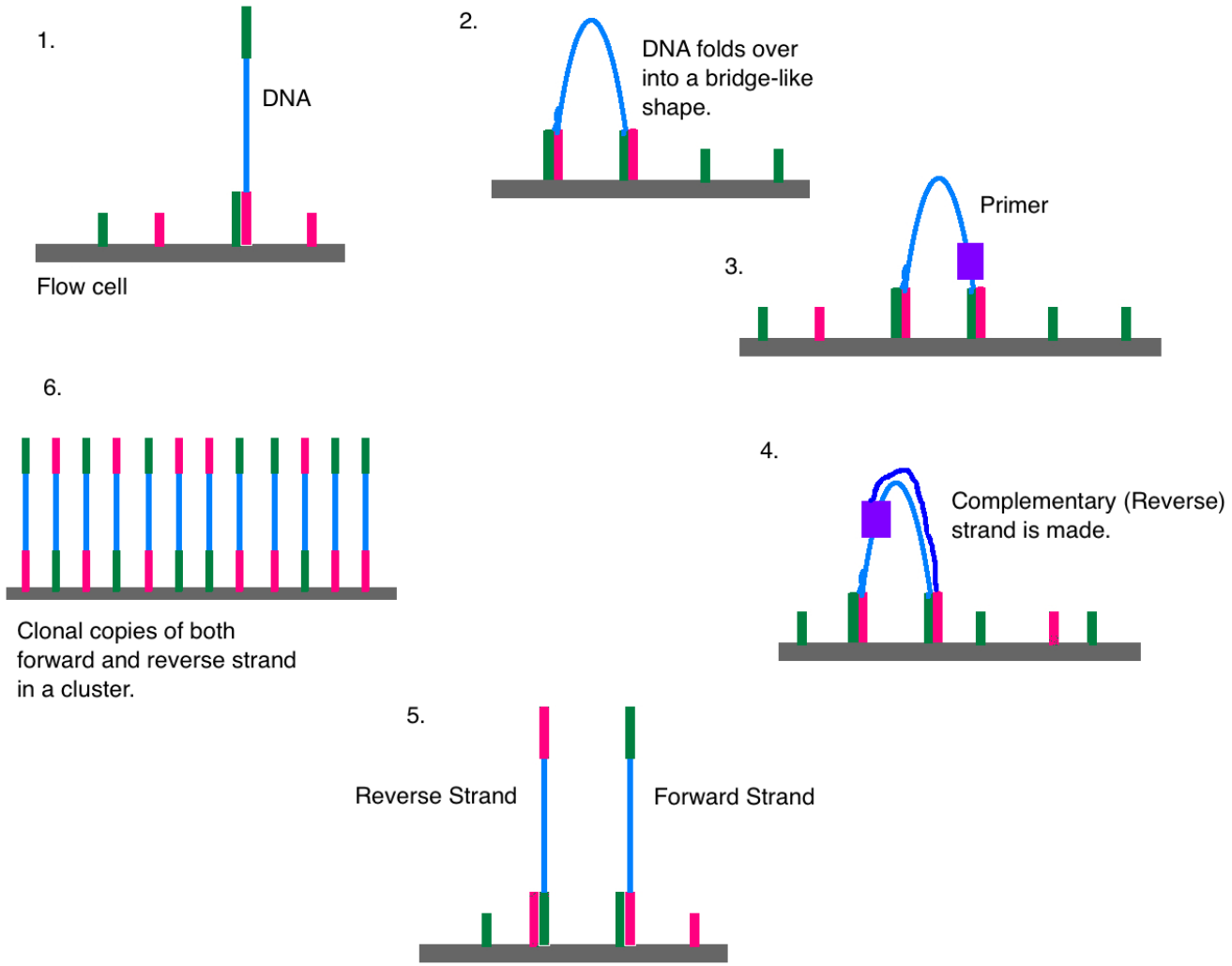
Source: CG000108 Rev A - Technical Note –Assay Scheme and Configuration of Chromium™ Single Cell 3' v2 Libraries

How do we sequence immune cells from blood?



Next Generation sequencing

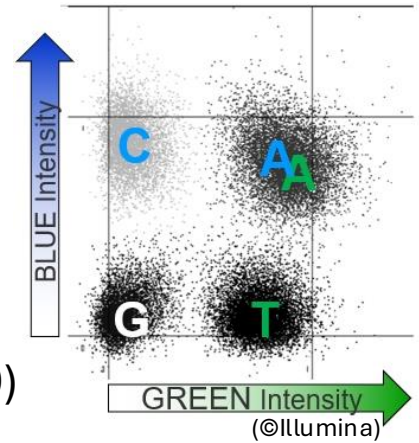
(sequencing by synthesis)



https://en.wikipedia.org/wiki/File:Cluster_Generation.png

After cluster generation:

- reverse strands are cleaved and washed away
- Polymerase attaches at forward strands & dye labelled nucleotides are provided
- Complementary strand is synthesised one nucleotide at a time, as dye blocks another nucleotide from being added
- Dyes are imaged and cleaved before the next cycle



Fluorescence emissions in a modern Illumina sequencer (NextSeq2000)

Data from the sequencer: bcl files

- Bcl (basecall) files are the binary output of the sequencing machine
  - They represent
    - which fluorophore was observed
    - in which location
    - during which cycle
    - with which certainty (-> base quality)
  - Bcl files are not human readable
- Bcl files need to be converted to fastq files for further processing
  - Either by you or the sequencing facility

Turning bcl into fastq files (demultiplexing)

- Remember the sample index?
  - Sample index sequences are used to separate data from all samples that shared the flowcell after sequencing
  - Which sequence belongs to which sample needs to be noted in a so-called sample sheet before sequencing / at library preparation time
  - Samples with the same index are impossible to separate if they are processed on the same flowcell
- If you use a user-operated sequencer (e.g. NextSeq500 / NextSeq2000 at the genomics core) you will need to demultiplex yourself using bcl2fastq (not part of this course)
- If you submitted your samples for sequencing, you can usually request fastq files if you provide a sample sheet

The anatomy of a fastq file

```
@VH01346:159:AAFNW5TM5:1:1101:64472:1000 2:N:0:CTAGTACG
GCTTCCCGACCCGCCCCCTCGCCCCCACCCCTGTGTGTTTCGCCAGTTAAGCTCCTGTGACTCCAGTACCTACTTCTGGTTTTGGGTTGGTTGTTCTGTCTTTTTTTTAATTAAATAAAAACAATTTTTAA
+
C-CCCCCCCCCC-CCCCC;-;;CCCC-;CCC-CCCCC-CC-CCC;--C-;CC-;C;CCC;;C--;CC;CCCC;C-;C;C-CC;-C;;CCC-C;;CCC;-CC---;C;;;CCC;C-CC-CC;-C;C-CCC--
@VH01346:159:AAFNW5TM5:1:1101:64510:1000 2:N:0:CTAGTACG
CCCCTGCTAGAATATTCTCTTGCTATAAACTAAGTATTGAGAACTAATGATAGTCTCTTGTTCTTTTAAATGGTGTCCTAATCCAAAATTATTACATTTTAACCTCTCAAAGAATAAATTACAAGATCTAA
+
-;CCCCCCCC;C-CC--CCCCC;CCC;C;C--CC--;C-CC-C-C-;C;-CCC;CCC;CCCC;-;CCC;CCCC;;CC-CCCC-CC-;CCC;CC----C;CCC-;-C;;;CCC-C-C;-;CC-C---;--
```

Read name (instrument name, flowcell ID, position of the read on the flowcell, library index)

Read sequence (this one is a gene read)

\+ sign, otherwise not used by illumna-style files

Quality score in ASCII

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

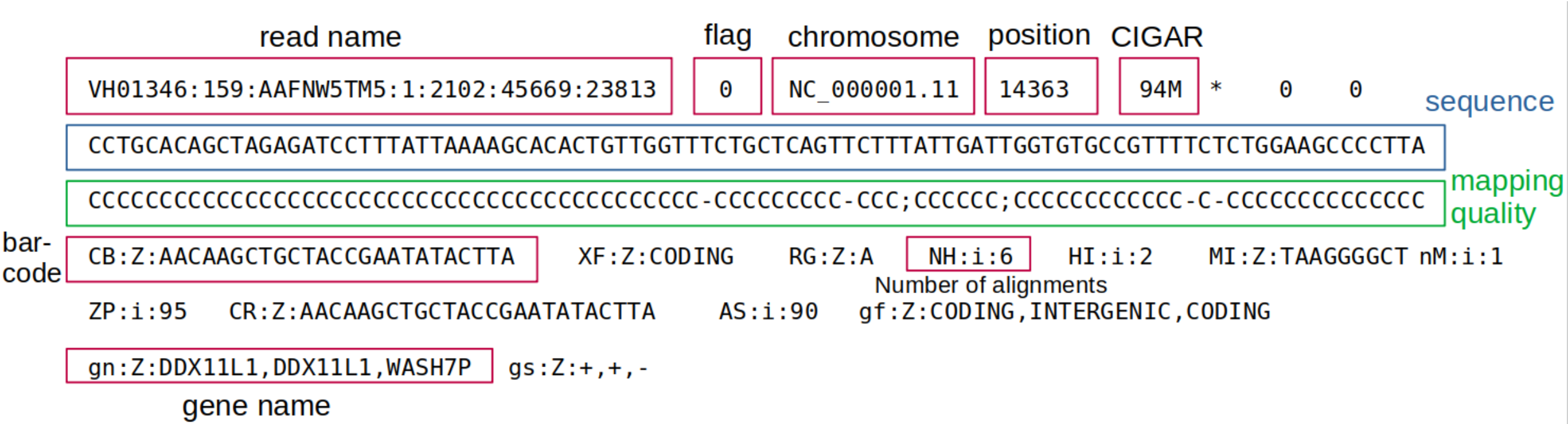
ASCII_BASE=64 Old Illumina

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	64 @	11	0.07943	75 K	22	0.00631	86 V	33	0.00050	97 a
1	0.79433	65 A	12	0.06310	76 L	23	0.00501	87 W	34	0.00040	98 b
2	0.63096	66 B	13	0.05012	77 M	24	0.00398	88 X	35	0.00032	99 c
3	0.50119	67 C	14	0.03981	78 N	25	0.00316	89 Y	36	0.00025	100 d
4	0.39811	68 D	15	0.03162	79 O	26	0.00251	90 Z	37	0.00020	101 e
5	0.31623	69 E	16	0.02512	80 P	27	0.00200	91 [	38	0.00016	102 f
6	0.25119	70 F	17	0.01995	81 Q	28	0.00158	92 \	39	0.00013	103 g
7	0.19953	71 G	18	0.01585	82 R	29	0.00126	93]	40	0.00010	104 h
8	0.15849	72 H	19	0.01259	83 S	30	0.00100	94 ^	41	0.00008	105 i
9	0.12589	73 I	20	0.01000	84 T	31	0.00079	95 _	42	0.00006	106 j
10	0.10000	74 J	21	0.00794	85 U	32	0.00063	96 `			

https://www.drive5.com/usearch/manual/quality_score.html

The anatomy of a .sam/.bam file

(after mapping)



Flag has encoded information on how the read mapped (e.g. forward or reverse), if it is a PCR duplicate,…

CIGAR string describes the alignment of the read (if it has gaps, insertions, non-matching portions at the beginning or end,…)

Warming up with your programming environment for data analysis: Learning Objectives

By the end of the morning session, you will have learned how to

- Apply for Charité HPC cluster access
- Map (10x genomics) single cell RNA sequencing data to a genome
- Interpret the cellranger report after mapping

By the end of the afternoon session, you will be able to

- Set up a workspace for data analysis
- Load and inspect tabular datasets
- Perform basic data transformations (standardization)
- Filter and subset data based on conditions
- Create basic visualizations

(and you will see who these little guys are)