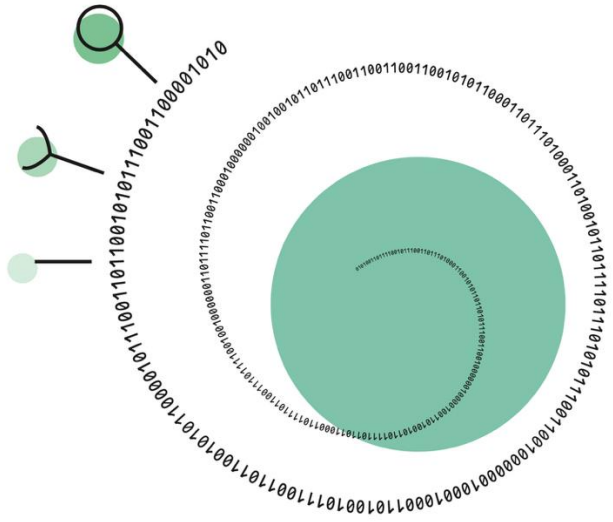
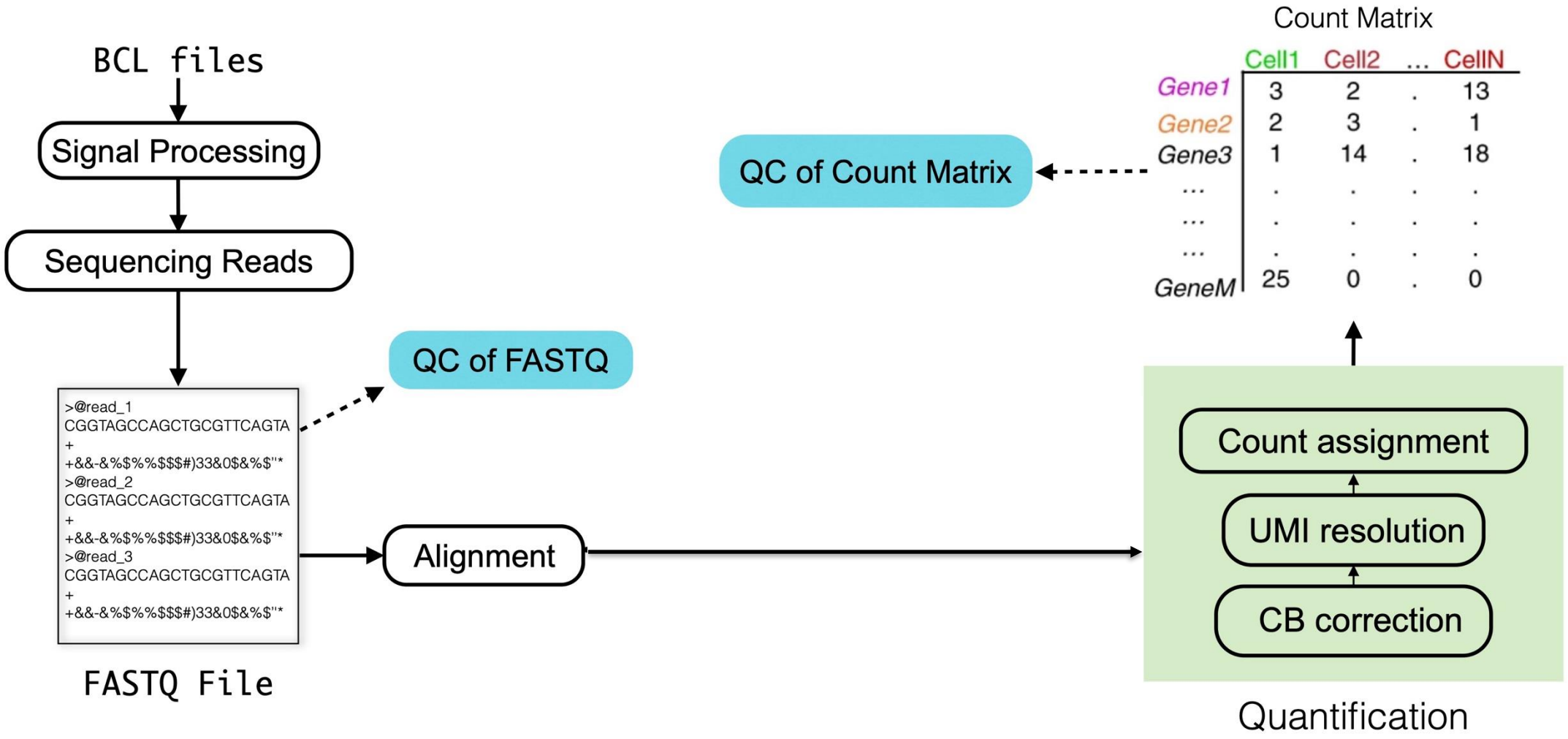


Single-Cell Data Analysis Course

What is a count matrix and how can we work with it?

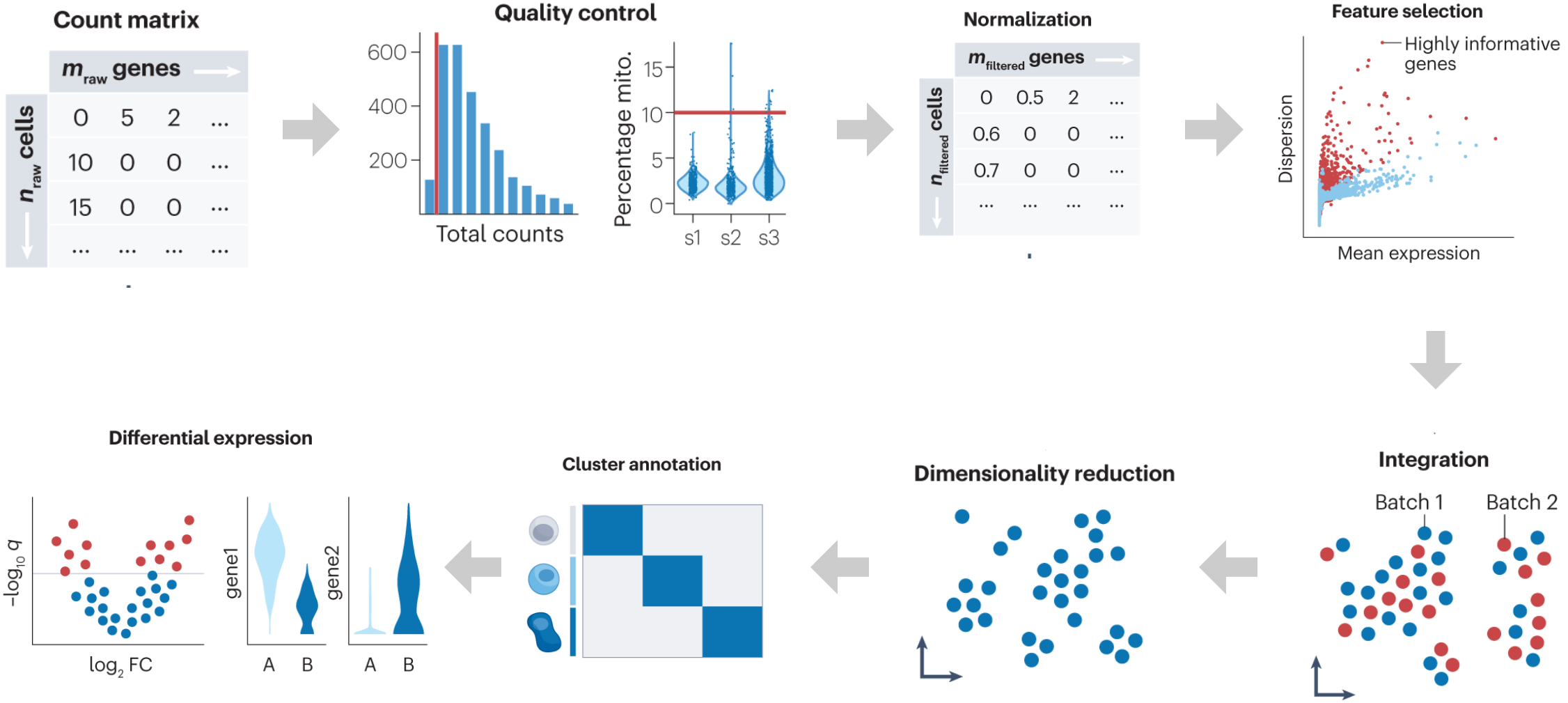
Lisa Buchauer
Professor of Systems Biology of Infectious Diseases
Department of Infectious Diseases and Intensive Care
Charité - Universitätsmedizin Berlin

From raw sequencing data to count matrix



https://www.sc-best-practices.org/_images/overview_raw_data_processing.jpg

Processing overview



Heumos, L., Schaar, A.C., Lance, C. et al. Best practices for single-cell analysis across modalities. Nat Rev Genet 24, 550–572 (2023). https://doi.org/10.1038/s41576-023-00586-w

The count matrix

Cells – “barcodes”

Genes/Transcripts – “features”

	A	B	C	D	E
G1	1	4	0	1	4
G2	1	4	2	3	2
G3	0	0	4	3	2

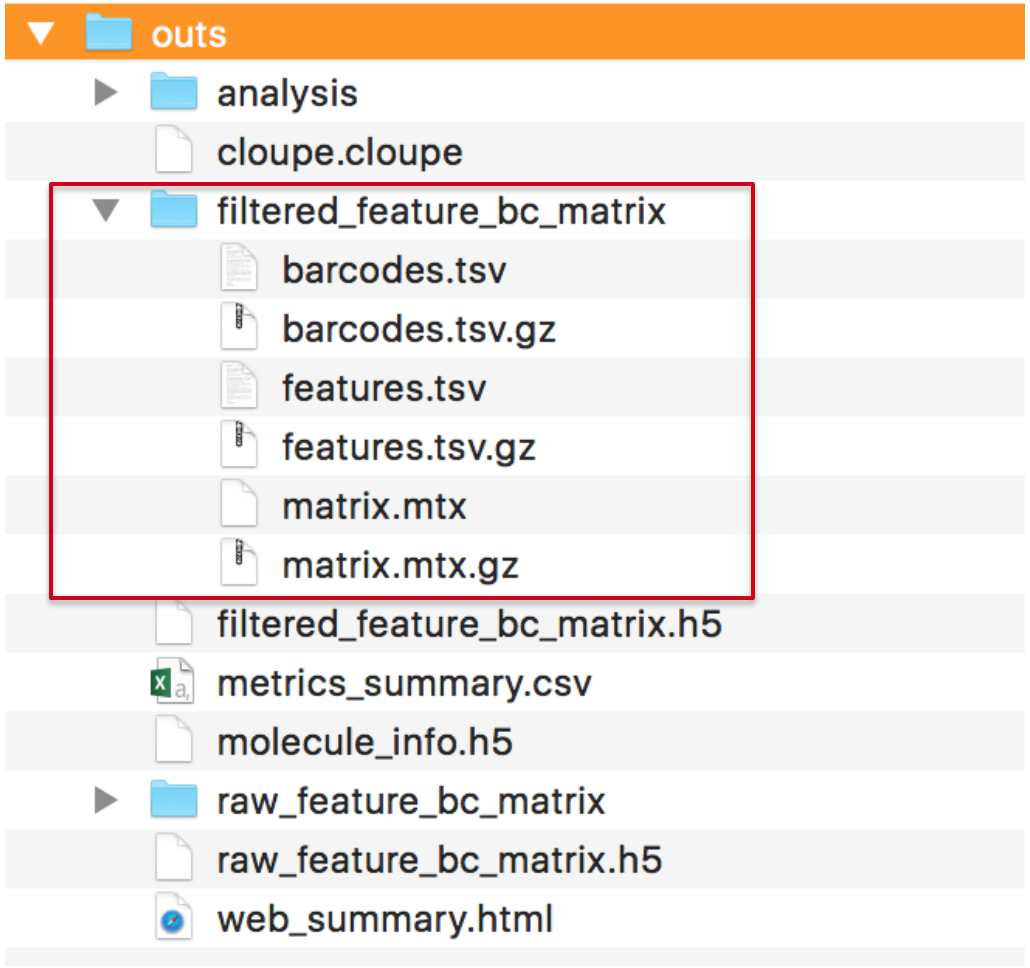
Count matrix

R/Seurat/SingleCellExperiment: cells are columns, genes are rows

Python/scanpy: cells are rows, genes are columns

https://training.galaxyproject.org/training-material/topics/single-cell/tutorials/scrna-intro/slides-plain.html

Count matrix, features and barcodes in cellranger output



Calculating summary statistics

	A	B	C	D	E
G1	1	4	0	1	4
G2	1	4	2	3	2
G3	0	0	4	3	2

Per Column (Cell)

	A	B	C	D	E
Library Size (Σ)	2	8	8	7	8
Genes Detected (> 0)	2	2	2	3	3
Genes Detected (> 1)	0	2	2	2	3
Genes Detected (> 2)	0	2	1	2	1
Genes Detected (> 3)	0	2	1	0	1
Genes Detected (> 4)	0	0	0	0	0

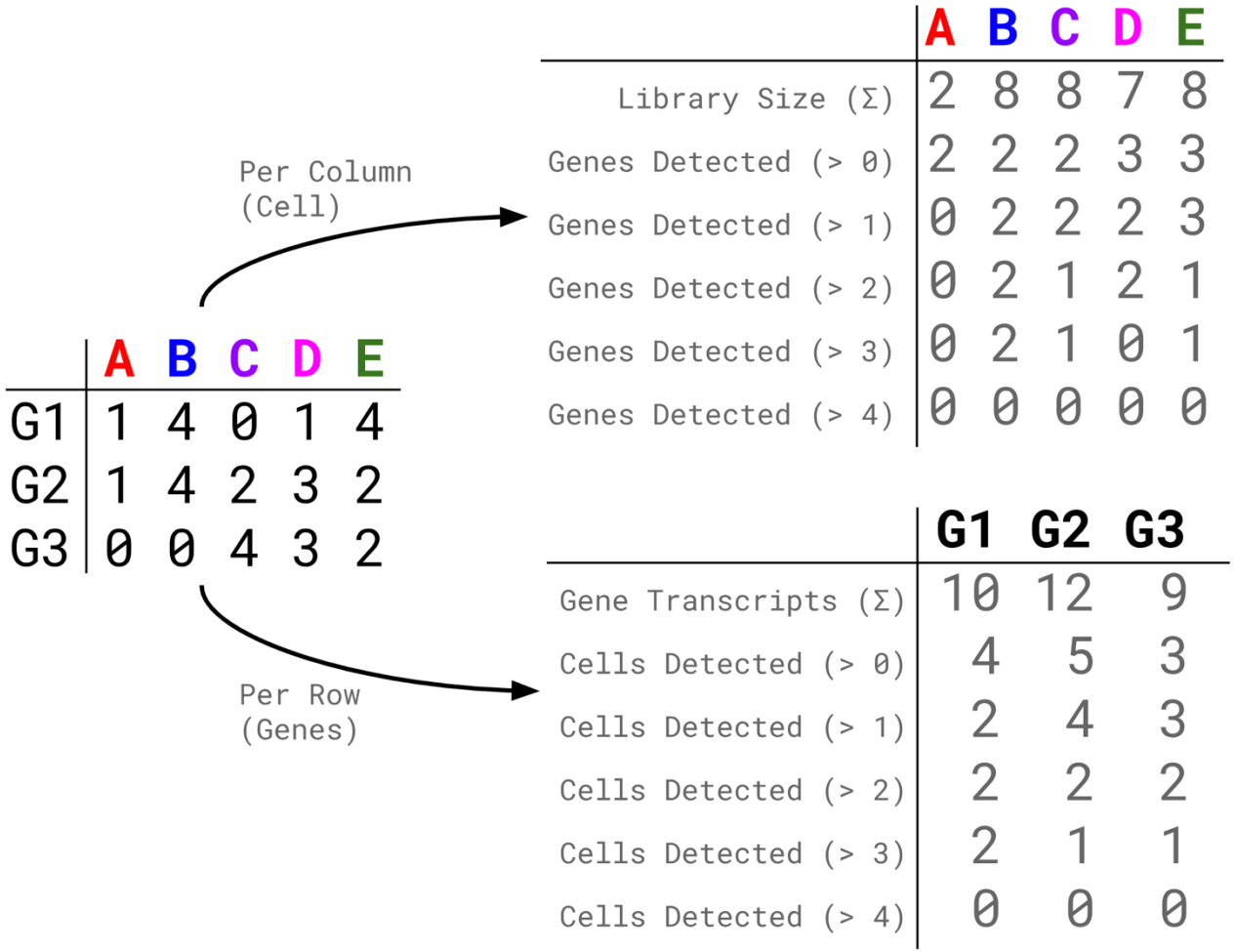
Per Row (Genes)

	G1	G2	G3
Gene Transcripts (Σ)	10	12	9
Cells Detected (> 0)	4	5	3
Cells Detected (> 1)	2	4	3
Cells Detected (> 2)	2	2	2
Cells Detected (> 3)	2	1	1
Cells Detected (> 4)	0	0	0

https://training.galaxyproject.org/training-material/topics/single-cell/tutorials/scrna-intro/slides-plain.html

Adding meta data …

	A	B	C	D	E
celltype	T	T	B	B	T
patient	1	1	1	2	2
gender	f	f	f	m	m
batch	A	B	A	B	A
time p.i.	7d	0d	7d	0d	7d
…					



https://training.galaxyproject.org/training-material/topics/single-cell/tutorials/scrna-intro/slides-plain.html

… and analysis results

Normalized expression

Nearest neighbor graph

UMAP

	A	B	C	D	E
celltype	T	T	B	B	T
patient	1	1	1	2	2
gender	f	f	f	m	m
batch	A	B	A	B	A
time p.i.	7d	0d	7d	0d	7d
…					

PCA

Highly variable genes

Differential expression

…

Per Column (Cell)

	A	B	C	D	E
G1	1	4	0	1	4
G2	1	4	2	3	2
G3	0	0	4	3	2

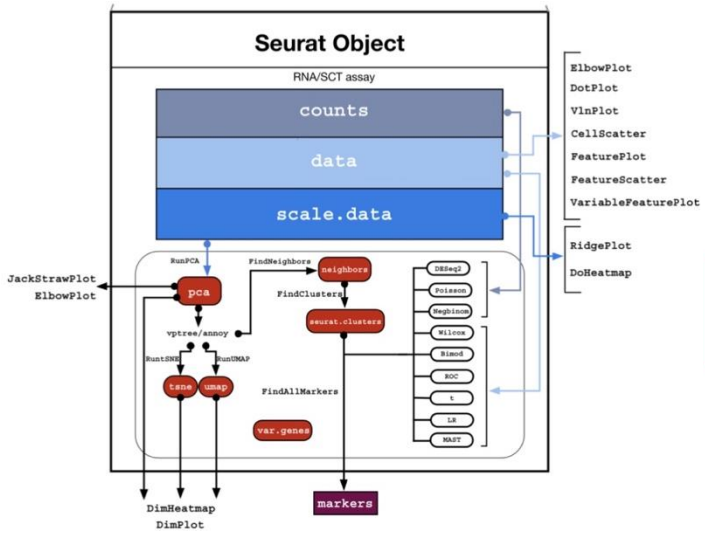
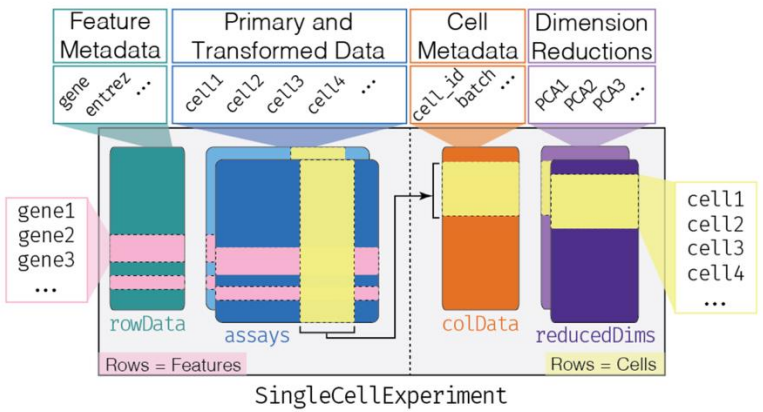
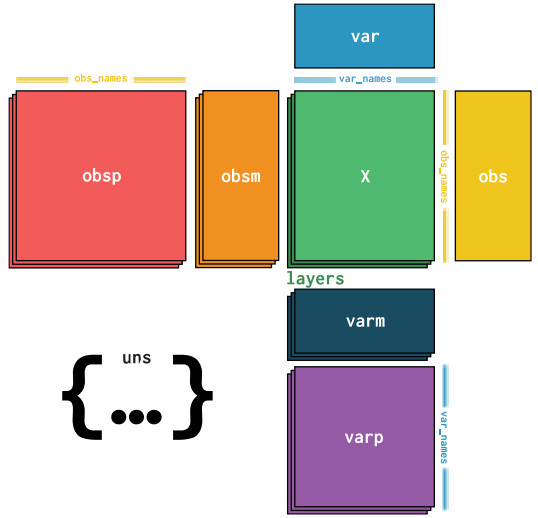
Per Row (Genes)

	A	B	C	D	E
Library Size (Σ)	2	8	8	7	8
Genes Detected (> 0)	2	2	2	3	3
Genes Detected (> 1)	0	2	2	2	3
Genes Detected (> 2)	0	2	1	2	1
Genes Detected (> 3)	0	2	1	0	1
Genes Detected (> 4)	0	0	0	0	0

	G1	G2	G3
Gene Transcripts (Σ)	10	12	9
Cells Detected (> 0)	4	5	3
Cells Detected (> 1)	2	4	3
Cells Detected (> 2)	2	2	2
Cells Detected (> 3)	2	1	1
Cells Detected (> 4)	0	0	0

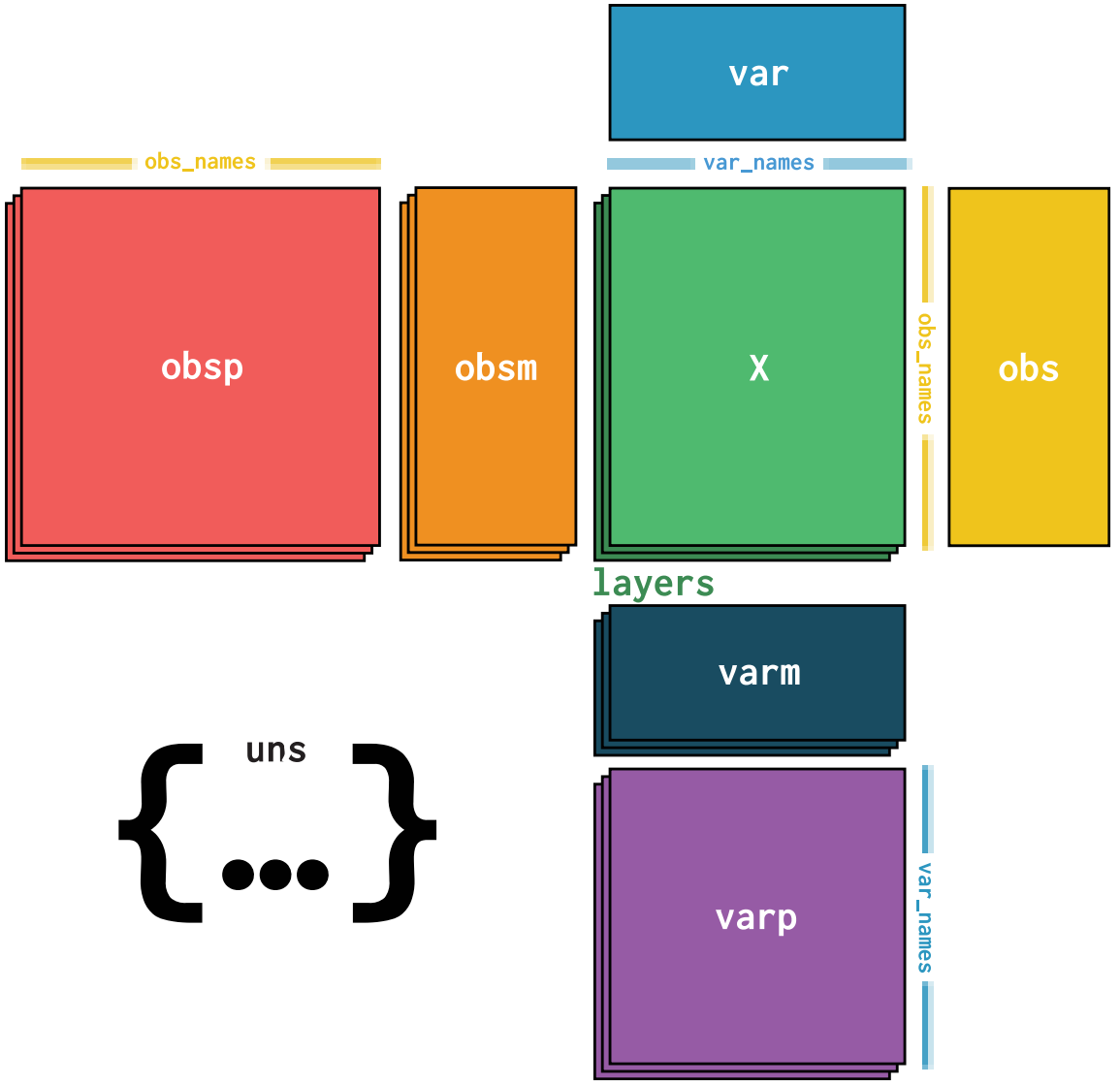
https://training.galaxyproject.org/training-material/topics/single-cell/tutorials/scrna-intro/slides-plain.html

Dedicated storage formats collect all relevant information in one place and provide easy access.

	Seurat Object	SingleCellExperiment	AnnData
			
language	R	R / python	python / R
toolkit	Seurat	[bioconductor]	scanpy
on disk	.rds, .h5Seurat	.rds, .h5, [.h5ad]	.h5ad

https://twitter.com/lpachter/status/1524413513233575936/photo/1
https://www.singlecellcourse.org/scrna-seq-analysis-with-bioconductor.html
https://anndata.readthedocs.io/en/latest/index.html

Example: AnnData objects preserve relationships between barcodes & features and meta-data & analyses



Commonly used environments for single-cell analysis in R and python

Bioconductor

scanpy

R	R	python
Starting point: "3k PBMCs guided tutorial"	**Starting point:** "Orchestrating single-cell analysis with Bionconductor"	**Starting point:** "Preprocessing and clustering 3k PBMCs"
https://satijalab.org/seurat/articles/pbmc3k_tutorial	https://bioconductor.org/books/release/OSCA/	https://scanpy.readthedocs.io/en/stable/tutorials/basics/clustering.html

YOU DO NOT HAVE TO USE A TOOLKIT!

Single-cell count matrices are sparse

- Most entries in the count matrix are zero
- Typically, 85-95% of all entries are 0
- This sets single cell count matrices apart from their bulk counterparts

```
adata = sc.read_10x_mtx("/Users/libuchauer/Projects/charite-sc-data-course/materials/Day2/healthy_PBMCs/")
```

```
adata.X.todense()
```

```
matrix([[0., 0., 0., ..., 0., 0., 0.],
        [0., 0., 0., ..., 0., 0., 0.],
        [0., 0., 0., ..., 0., 0., 0.],
        ...,
        [0., 0., 0., ..., 0., 0., 0.],
        [0., 0., 0., ..., 0., 0., 0.],
        [0., 0., 0., ..., 0., 0., 0.]], shape=(2700, 33538), dtype=float32)
```

94% of all entries are 0

Origins of sparsity in single-cell data

How much RNA does a typical mammalian cell contain?

The RNA content and RNA make up of a cell depend very much on its developmental stage and the type of cell. To estimate the approximate yield of RNA that can be expected from your starting material, we usually calculate that a typical mammalian cell contains 10–30 pg total RNA.

The majority of RNA molecules are tRNAs and rRNAs. mRNA accounts for only 1–5% of the total cellular RNA although the actual amount depends on the cell type and physiological state. Approximately 360,000 mRNA molecules are present in a single mammalian cell, made up of approximately 12,000 different transcripts with a typical length of around 2 kb. Some mRNAs comprise 3% of the mRNA pool whereas others account for less than 0.1%. These rare or low-abundance mRNAs may have a copy number of only 5–15 molecules per cell.

https://www.qiagen.com/us/resources/faq/2946

- Biological sources
  - Low RNA content per cell
  - Transcriptional bursting
  - Cell-type specific expression patterns
- Technical sources
  - “Dropout” events – limited sensitivity of sequencing platforms

Efficient storage with sparse matrices

0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	1	0	0	0
0	0	0	0	0	0
0	0	0	0	2	0

How would you remember this table?

Efficient storage with sparse matrices

Standard (dense) storage wastes resources

- 10k cells, 20k genes → 200M entries
- If 90% are zeros → storing 180M unnecessary zeros

Solution: Sparse formats store only non-zero values, e.g. via coordinates

- **R/Seurat**: Uses Matrix package (dgCMatrix format)
- **Python/Scanpy**: Uses scipy.sparse matrices
- **File formats**: HDF5, MEX format preserve sparsity

Not all numerical operations can be performed with sparse matrices, sometimes conversion to dense is necessary.

```
adata.X
```

```
<Compressed Sparse Column sparse matrix of dtype 'float32'
        with 5682040 stored elements and shape (2700, 33538)>
```