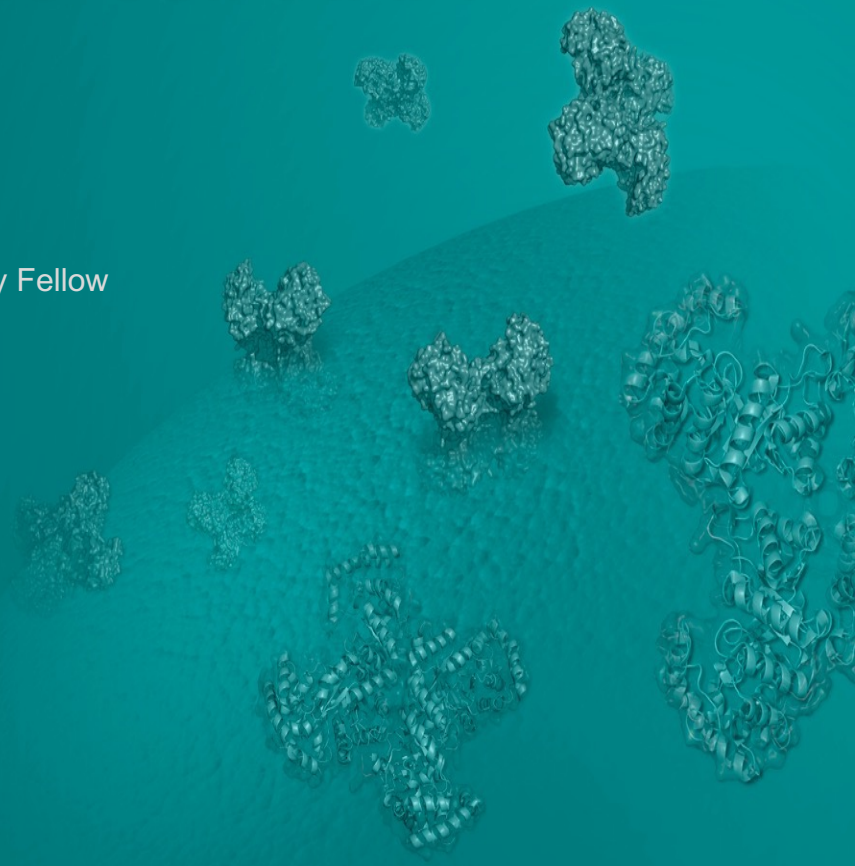
Introduction to single-cell RNA-seq data analysis

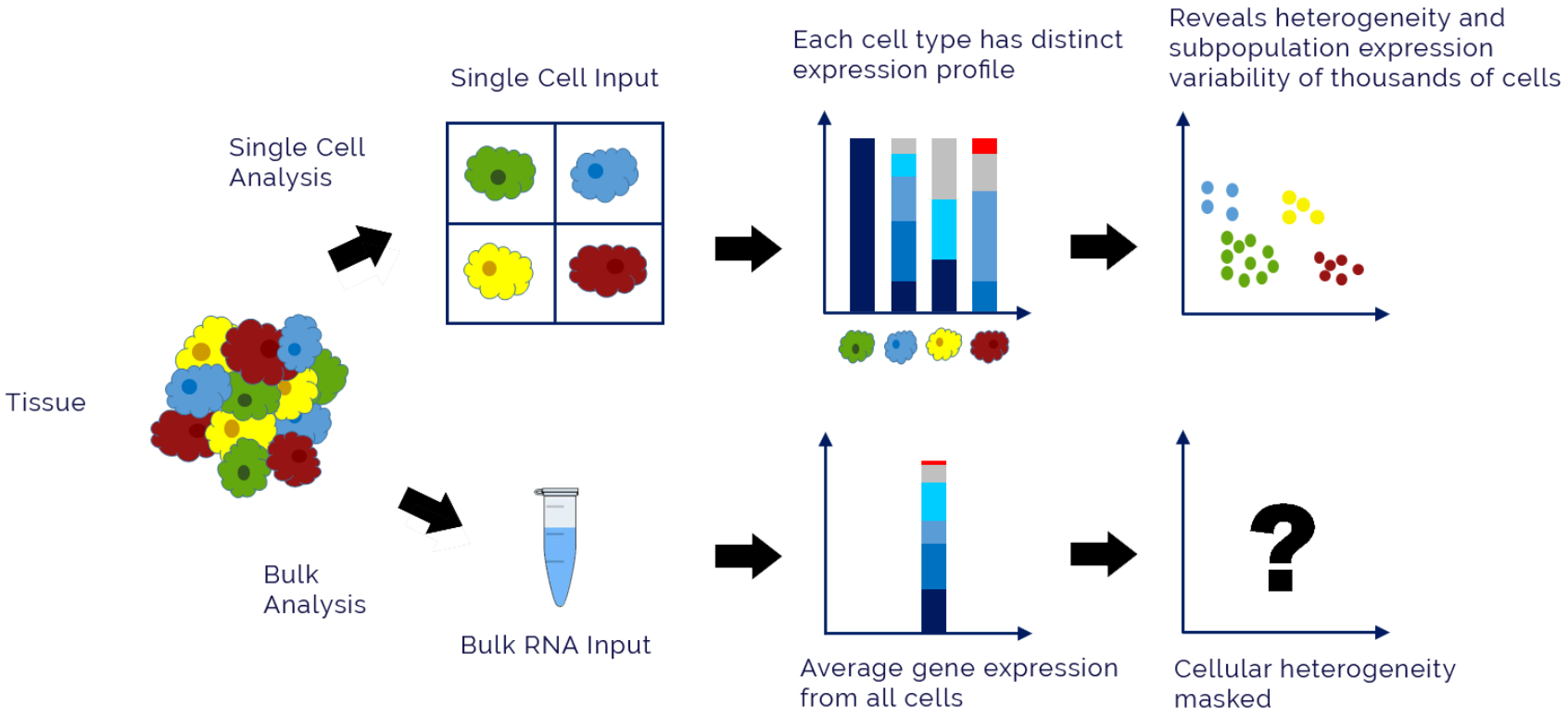
Jiawei Wang, PhD
Marie Curie Postdoctoral Fellow & EMBO Non-Stipendiary Fellow
Finn & Marioni Groups
EMBL-EBI
jwang@ebi.ac.uk
2024/07/03



EMBL-EBI

Why do we need single-cell RNA sequencing (scRNA-seq)?

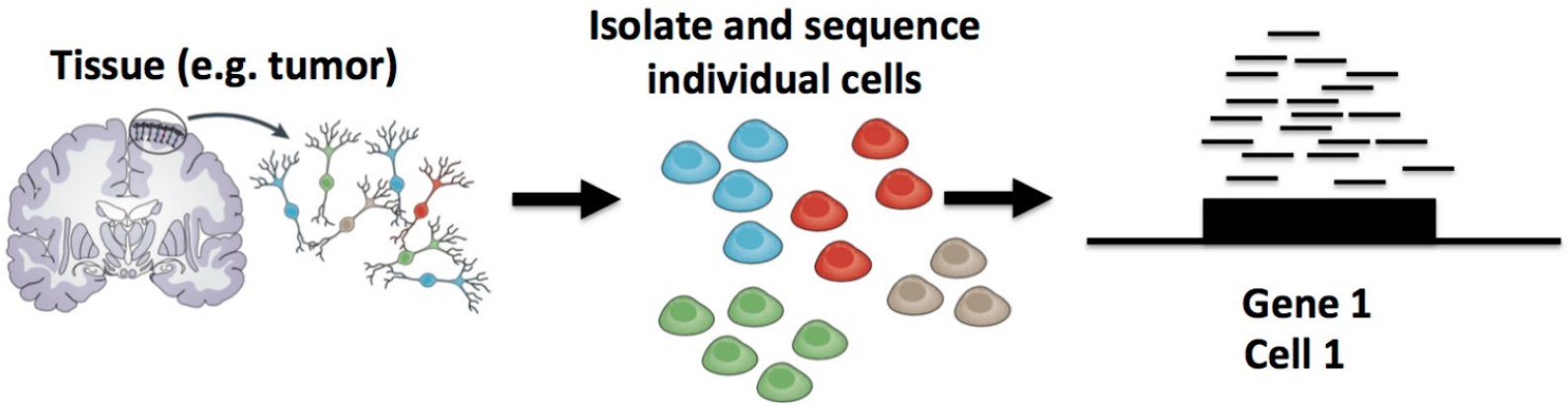
- scRNA-seq reveals cellular heterogeneity that is masked by bulk RNA-seq methods.



Credit: https://www.10xgenomics.com/blog/single-cell-rna-seq-an-introductory-overview-and-tools-for-getting-started

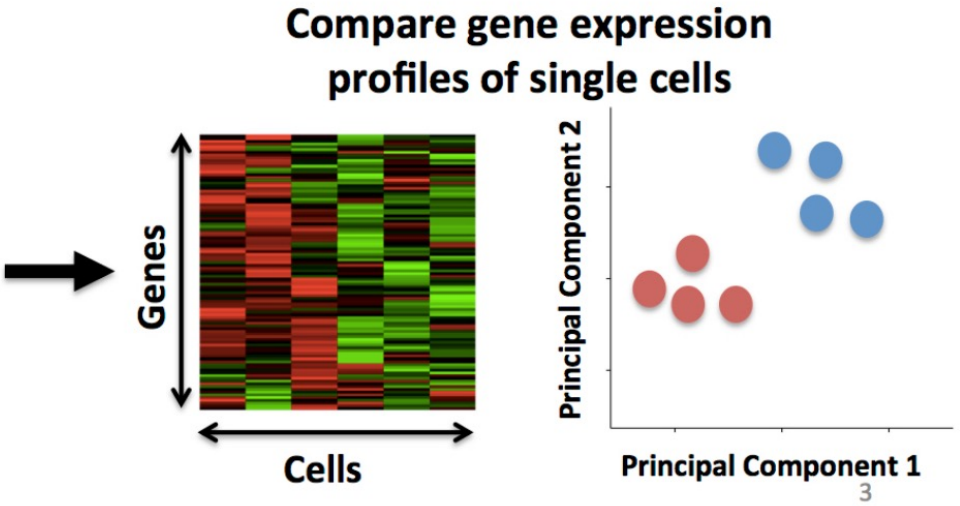
EMBL-EBI

scRNA-seq workflow



Read Counts

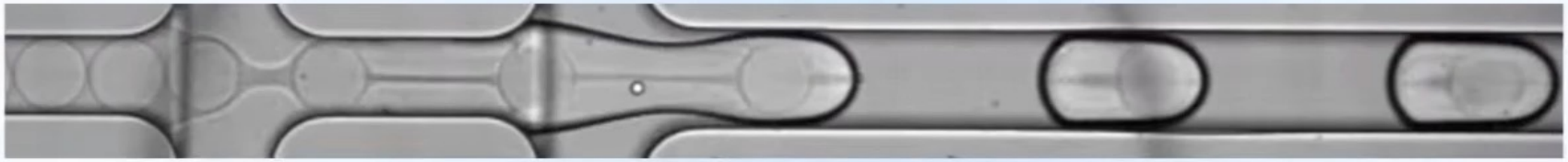
	Cell 1	Cell 2	...
Gene 1	18	0	
Gene 2	1010	506	
Gene 3	0	49	
Gene 4	22	0	
...			



Credit: https://learn.gencore.bio.nyu.edu/single-cell-rnaseq/

EMBL-EBI

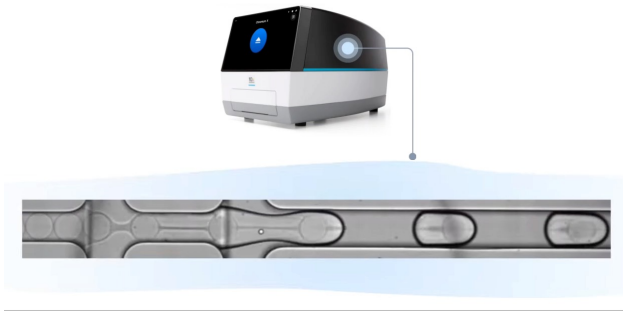
Key stage to isolate single cells (by 10x Genomics)

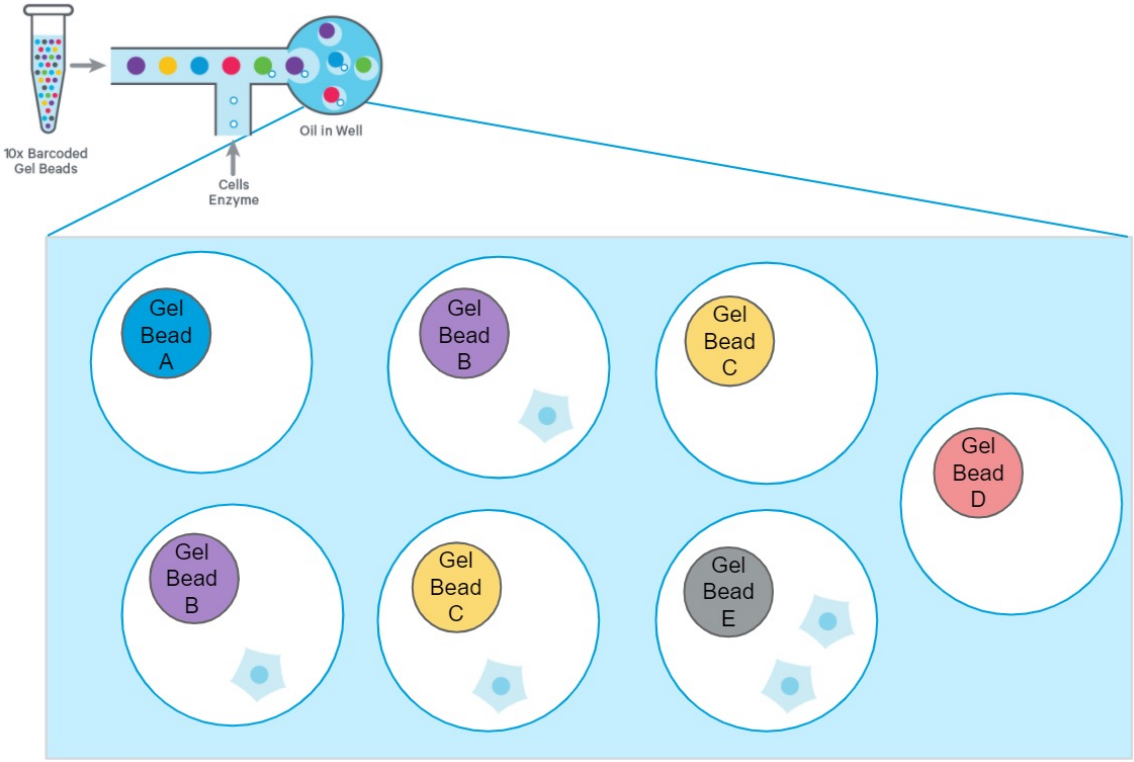


Credit: https://www.10xgenomics.com/instruments/chromium-x-series

EMBL-EBI

Key stage to isolate single cells (by 10x Genomics)





Credit: https://kb.10xgenomics.com/hc/en-us/articles/360059124751-Why-is-the-multiplet-rate-different-for-the-Next-GEM-Single-Cell-3-LT-v3-1-assay-compared-to-other-single-cell-applications

scRNA-seq data processed by Cell Ranger

Datasets

10k PBMCs from a Healthy Donor (v3 chemistry)

Single Cell Gene Expression Dataset by Cell Ranger 3.0.0

Peripheral blood mononuclear cells (PBMCs) from a healthy donor (the same cells were used to gene pbmc_10k_v3). PBMCs are primary cells with relatively small amounts of RNA (~1pg RNA/cell).

- 11,769 cells detected
- Sequenced on Illumina NovaSeq with approximately 54,000 reads per cell
- 28bp read1 (16bp Chromium barcode and 12bp UMI), 91bp read2 (transcript), and 8bp I7 sample
- run with --expect-cells=10000

Results Summary

View summary metrics for the experiment

View Summary

Download in browser | Batch download

If the file size is large, we suggest using batch download instead.

Input Files	Size	md5sum
FASTQs	51.7 GB	e0021592e209642d71f5dc420cf4c5c0

Output Files (Format details)	Size	md5sum
Genome-aligned BAM	44.2 GB	779efd29b694e27be0cb836ebcad1f70
Genome-aligned BAM index	17.3 MB	2049e9481aca84bf3410055f7eed15fa
Per-molecule read information	488 MB	835556df4d369e444dbca7c1f75a7342
Feature / cell matrix HDF5 (filtered)	37.5 MB	563fb0c8cae8a4410a26ac62bd2a3c1f
Feature / cell matrix (filtered)	94.3 MB	f6f80a4561ebd7816a9c0816f1f15e0f
Feature / cell matrix HDF5 (raw)	176 MB	b59e3542d1d2e67a717a4c18c7c6c3af
Feature / cell matrix (raw)	151 MB	7fbd2395f903173ea4c4eef6ae95818c
Clustering analysis	33.5 MB	b04b7aaf5d07649702dd67274b8e2456
Summary CSV	684 B	8acaa33c2c225287ab4937f2dff2a66d
Summary HTML	4.61 MB	a3edf5d766bd31b25eb049e161f9ec0b
Loupe Browser file	126 MB	59c51000169ae632593e1a6629c53317

Data available at: https://www.10xgenomics.com/datasets/10-k-pbm-cs-from-a-healthy-donor-v-3-chemistry-3-standard-3-0-0

EMBL-EBI

scRNA-seq data processed by Cell Ranger

Estimated Number of Cells

11,769

Mean Reads per Cell	Median Genes per Cell
54,286	1,906

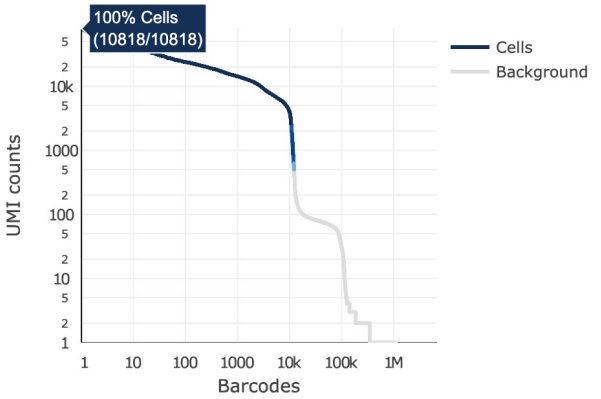
Sequencing

Number of Reads	638,901,019
Valid Barcodes	97.4%
Sequencing Saturation	68.2%
Q30 Bases in Barcode	93.7%
Q30 Bases in RNA Read	90.1%
Q30 Bases in Sample Index	90.1%
Q30 Bases in UMI	92.4%

Mapping

Reads Mapped to Genome	95.5%
Reads Mapped Confidently to Genome	92.5%
Reads Mapped Confidently to Intergenic Regions	5.0%
Reads Mapped Confidently to Intronic Regions	34.7%
Reads Mapped Confidently to Exonic Regions	52.7%
Reads Mapped Confidently to Transcriptome	49.7%
Reads Mapped Antisense to Gene	1.3%

Cells



Estimated Number of Cells	11,769
Fraction Reads in Cells	95.1%
Mean Reads per Cell	54,286
Median Genes per Cell	1,906
Total Genes Detected	23,036
Median UMI Counts per Cell	6,521

Sample

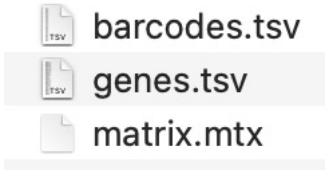
Name	pbmc_10k_v3
Description	Peripheral blood mononuclear cells (PBMCs) from a healthy donor
Transcriptome	GRCh38
Chemistry	Single Cell 3' v3
Cell Ranger Version	3.0.0

Page from: https://cf.10xgenomics.com/samples/cell-exp/3.0.0/pbmc_10k_v3/pbmc_10k_v3_web_summary.html

EMBL-EBI

Two Cell Ranger output formats

".tsv" and ".mtx" file formats:



".h5" file format: Hierarchical Data Format (HDF5 or H5)

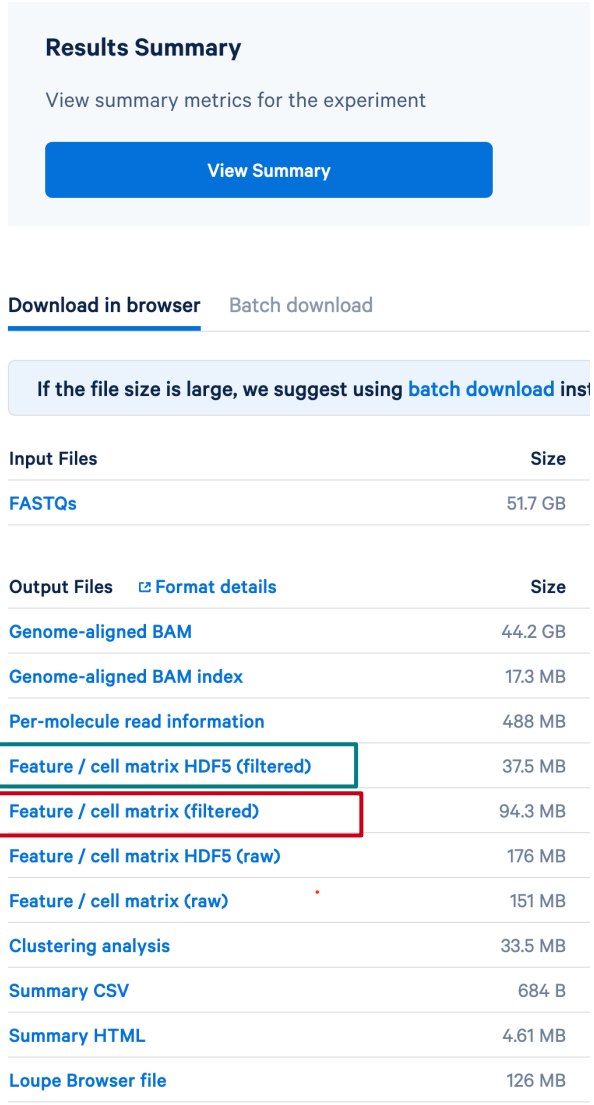
Read10X()

Read10X_h5()

A unique molecular identified (UMI) count matrix

Count Matrix

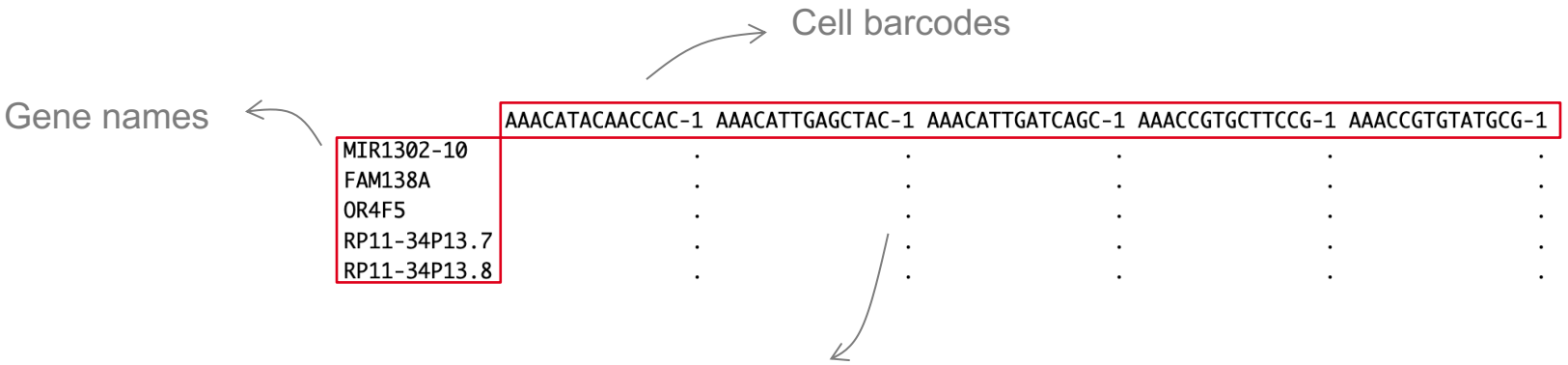
	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0



EMBL-EBI

Unique molecular identifier (UMI) count matrix

- The data used for scRNA-seq data analysis is in the form of a matrix, where rows represent features (gene names) and columns represent samples (cells).



Number of unique molecules for each feature (i.e. gene; row) that are detected in each cell (column)

EMBL-EBI

Tools/Platforms for scRNA-seq data analysis

- Seurat: an R package for integrative scRNA-seq data analysis
- Bioconductor: a R tool collection for the analysis and comprehension of genomic data
- Scanpy: scRNA-seq data analysis package in Python, included in scverse.
- Scverse: a consortium of foundational tools (mostly in Python) for omics data in life sciences.

EMBL-EBI

Seurat for scRNA-seq data analysis

1. Data pre-processing

Quality control

2. Data processing

Data normalization

Feature selection

Data scaling

Dimensional reduction

Cell clustering

Data visualization

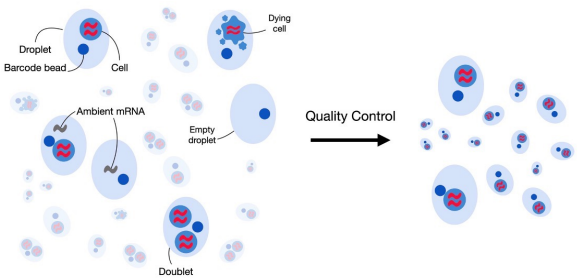
Biomarker Discovery

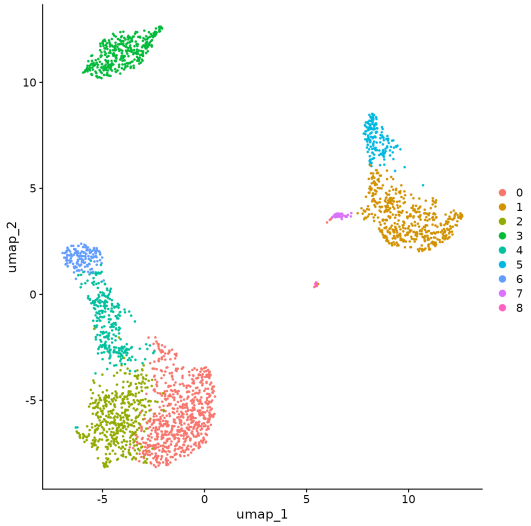
3. Downstream analysis

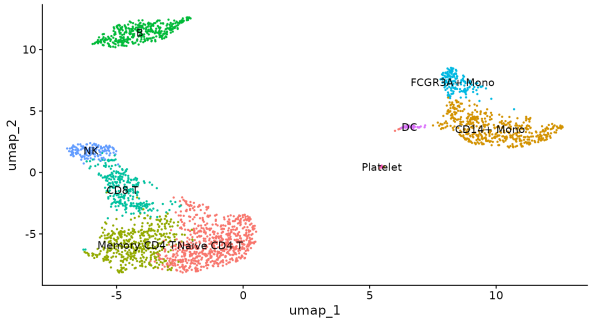
Cell type/state discovery

RNA trajectory analysis

RNA velocity



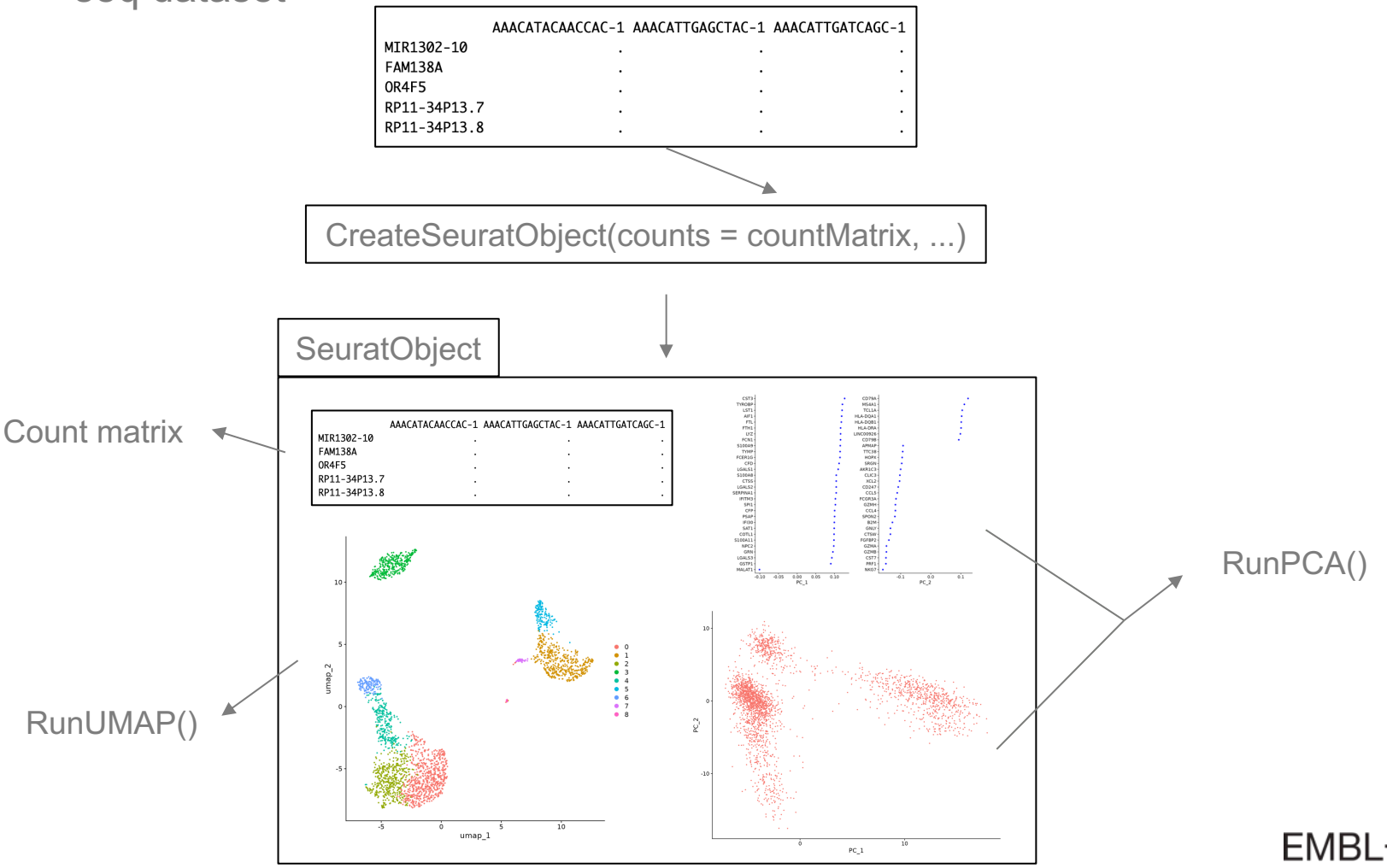




EMBL-EBI

Seurat object

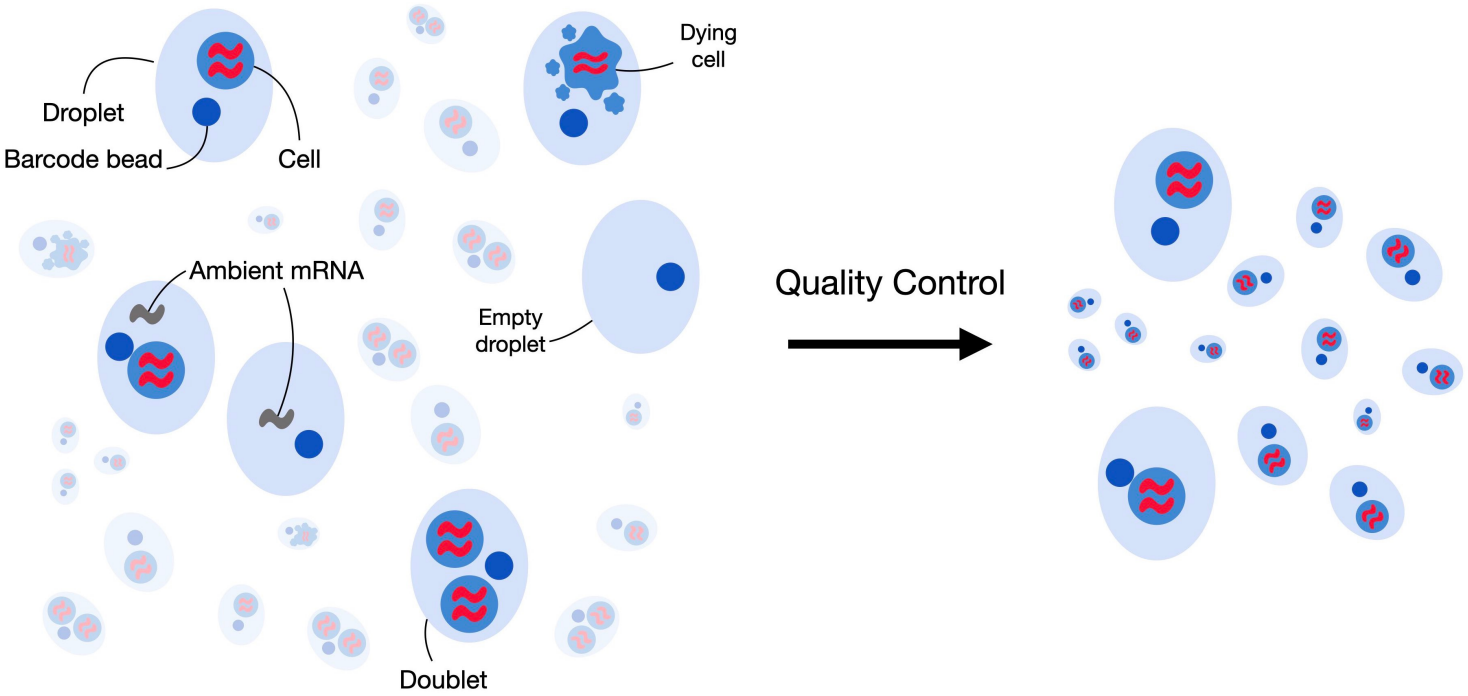
- A container that contains both data (including the UMI count matrix) and analysis (including PCA, UMAP, t-SNE and clustering) results for a scRNA-seq dataset



EMBL-EBI

Quality control

- Single-cell RNA-seq datasets can contain low-quality cells, cell-free RNA, and doublets.
- Quality control aims to remove or correct these issues to ensure a high-quality dataset, where each observation represents an intact single cell.

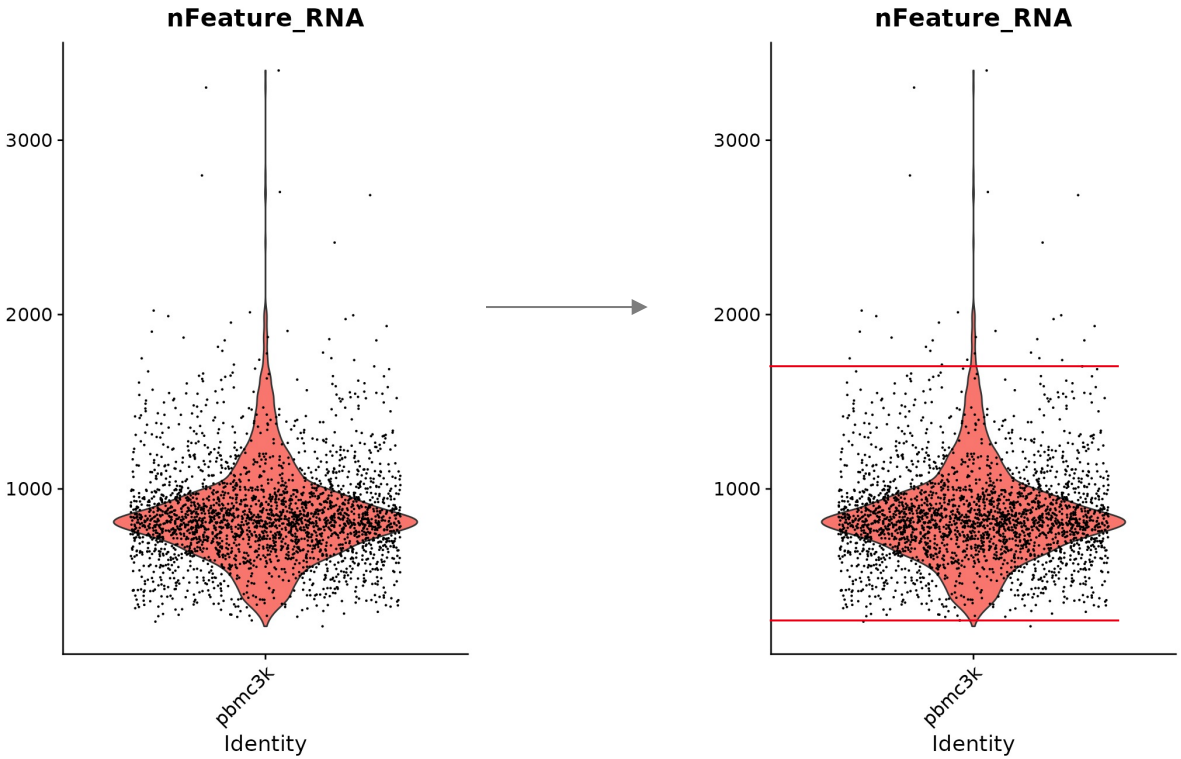


Credit: https://www.sc-best-practices.org/preprocessing_visualization/quality_control.html

EMBL-EBI

Quality control (QC) metrics

- QC and selecting cells for further analysis
  - Metric 1: The number of unique genes detected in each cell (nFeature_RNA)
    - Low-quality cells or empty droplets will often have very few genes
    - Cell doublets or multiplets may exhibit an aberrantly high gene count



Credit: https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

EMBL-EBI

Quality control (QC) metrics

- QC and selecting cells for further analysis
  - Metric 1: The number of unique genes detected in each cell.
    - Low-quality cells or empty droplets will often have very few genes
    - Cell doublets or multiplets may exhibit an aberrantly high gene count
  - Metric 2: The total number of unique molecules detected within a cell (nCount_RNA; which correlates strongly with unique genes)
  - Metric 3*: The percentage of reads that map to the mitochondrial genome
    - Low-quality / dying cells often exhibit extensive mitochondrial contamination

*: Self-defined metric

Credit: https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

Data normalization

- Feature counts for each cell are divided by the total counts for that cell and multiplied by a scale factor
- Relies on an assumption that each cell originally contains the same number of RNA molecules

	AAACATACAACCAC-1	AAACATTGAGCTAC-1	AAACATTGATCAGC-1	AAACCGTGCTTCCG-1	AAACCGTGTATGCG-1
MIR1302-10	.	.	.	.	.
FAM138A	.	.	.	.	.
OR4F5	.	.	.	.	.
RP11-34P13.7	.	.	.	.	.
RP11-34P13.8	.	.	.	.	.

Credit: https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

EMBL-EBI

Identification of highly variable features (feature selection)

- Calculate a subset of features that exhibit high cell-to-cell variation in the dataset (i.e, they are highly expressed in some cells, and lowly expressed in others)
- Focusing on these genes in downstream analysis helps to highlight biological signal in single-cell datasets

	AAACATACAACCAC-1	AAACATTGAGCTAC-1	AAACATTGATCAGC-1	AAACCGTGCTTCCG-1	AAACCGTGTATGCG-1
MIR1302-10	.	.	.	.	.
FAM138A	.	.	.	.	.
OR4F5	.	.	.	.	.
RP11-34P13.7	high	low	low	low	high
RP11-34P13.8	.	.	.	.	.

Highly variable feature

	AAACATACAACCAC-1	AAACATTGAGCTAC-1	AAACATTGATCAGC-1	AAACCGTGCTTCCG-1	AAACCGTGTATGCG-1
MIR1302-10	.	.	.	.	.
FAM138A	high	high	high	high	high
OR4F5	.	.	.	.	.
RP11-34P13.7	.	.	.	.	.
RP11-34P13.8	.	.	.	.	.

Not a highly variable feature

Credit: https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

EMBL-EBI

Data scaling

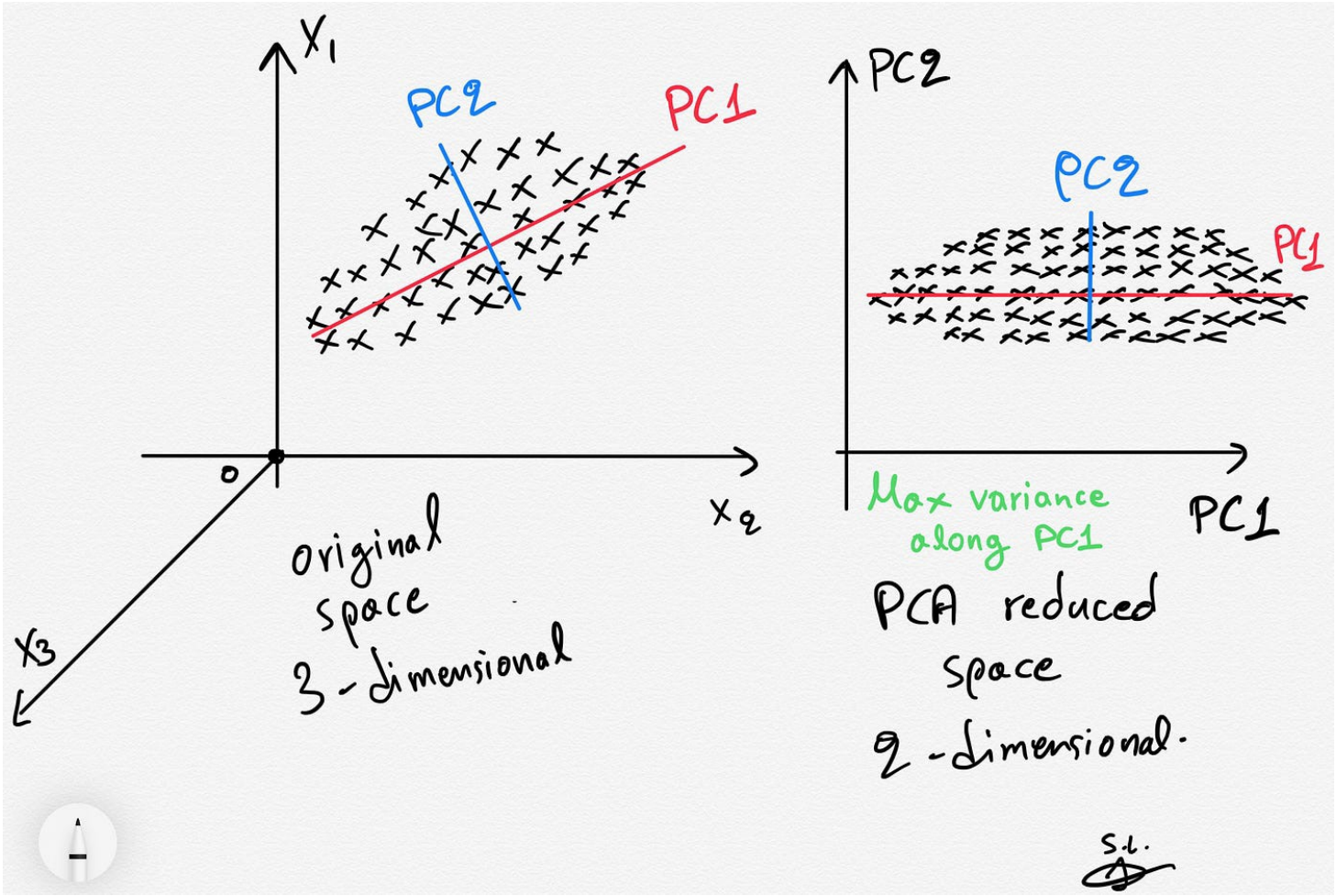
- A standard pre-processing step prior to dimensional reduction techniques like PCA
  - Shifts the expression of each gene, so that the mean expression across cells is 0
  - Scales the expression of each gene, so that the variance across cells is 1. This step gives equal weight in downstream analyses, so that highly-expressed genes do not dominate

	AAACATACAACCAC-1	AAACATTGAGCTAC-1	AAACATTGATCAGC-1	AAACCGTGCTTCCG-1	AAACCGTGTATGCG-1
MIR1302-10	.	.	.	.	.
FAM138A	.	.	.	.	.
OR4F5	.	.	.	.	.
RP11-34P13.7	.	.	.	.	.
RP11-34P13.8	.	.	.	.	.

Credit: https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

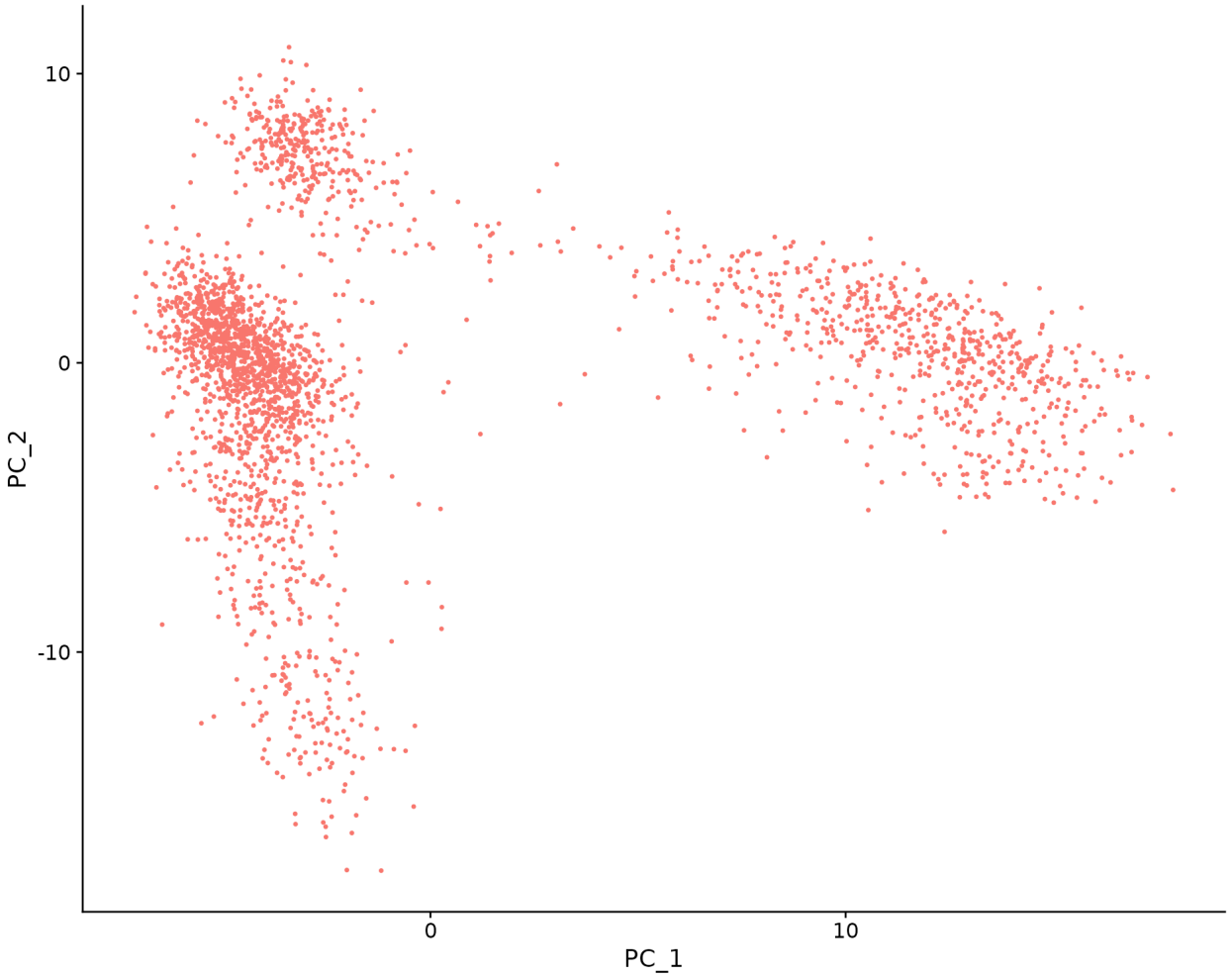
EMBL-EBI

Linear dimensional reduction (PCA)



Credit at: https://towardsdatascience.com/pca-clearly-explained-how-when-why-to-use-it-and-feature-importance-a-guide-in-python-7c274582c37e

Visualization based on PCA

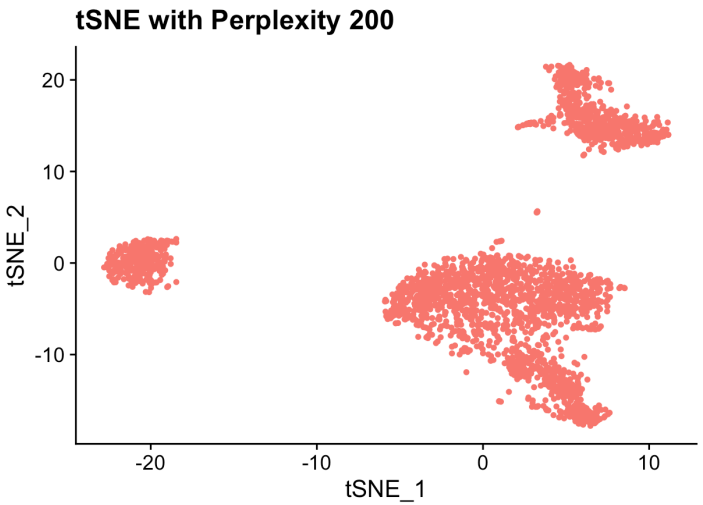


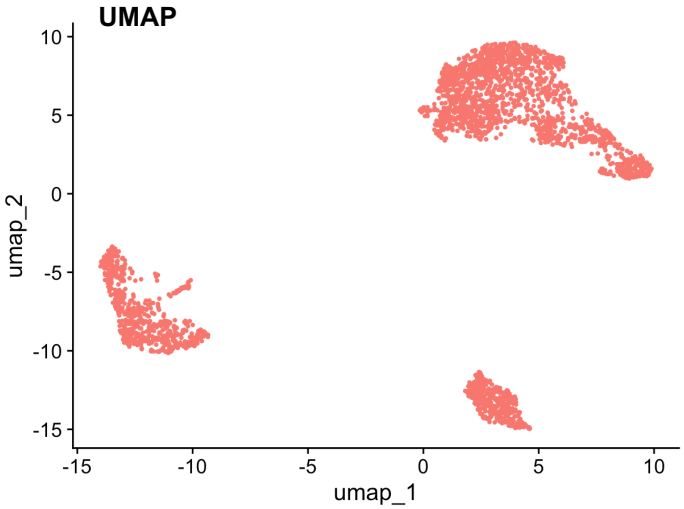
Credit: https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

EMBL-EBI

Non-linear dimensional reduction (UMAP/tSNE)

- These methods aim to preserve local distances in the dataset (i.e. ensuring that cells with very similar gene expression profiles co-localize), but often do not preserve more global relationships.
- Often based on the PCA results



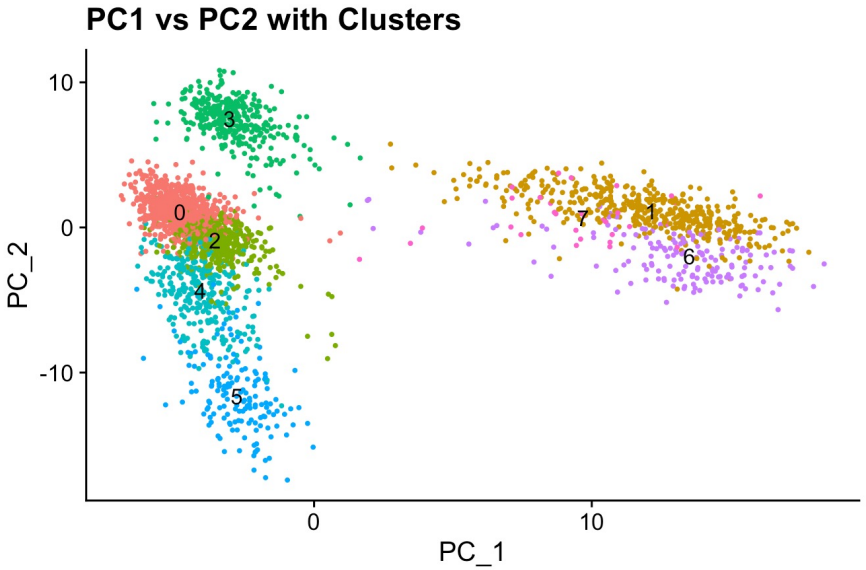


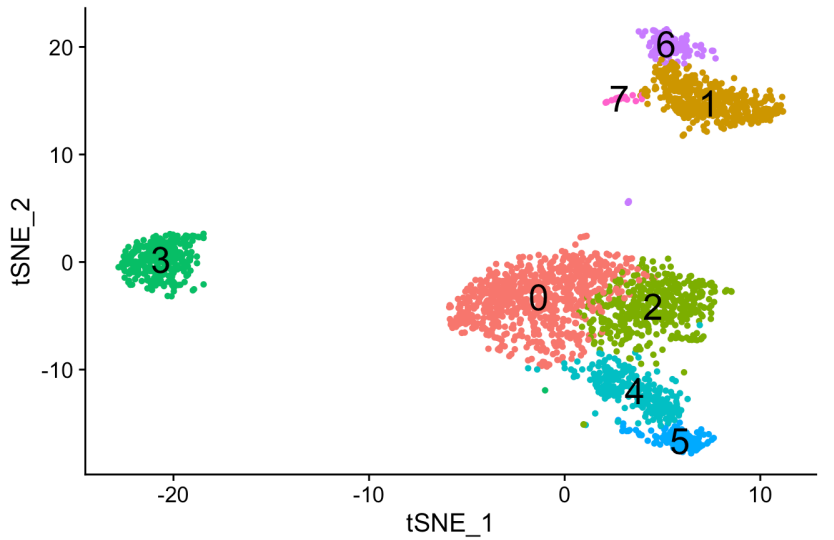
Credit: https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

EMBL-EBI

Cell clustering / Data visualization

- Use a graph-based method to detect clusters
- This finds the ‘k’ nearest neighbours to each cell and makes this into a graph. It then looks for highly inter-connected subgraphs within the graph and uses these to define clusters.
- Often based on the PCA results



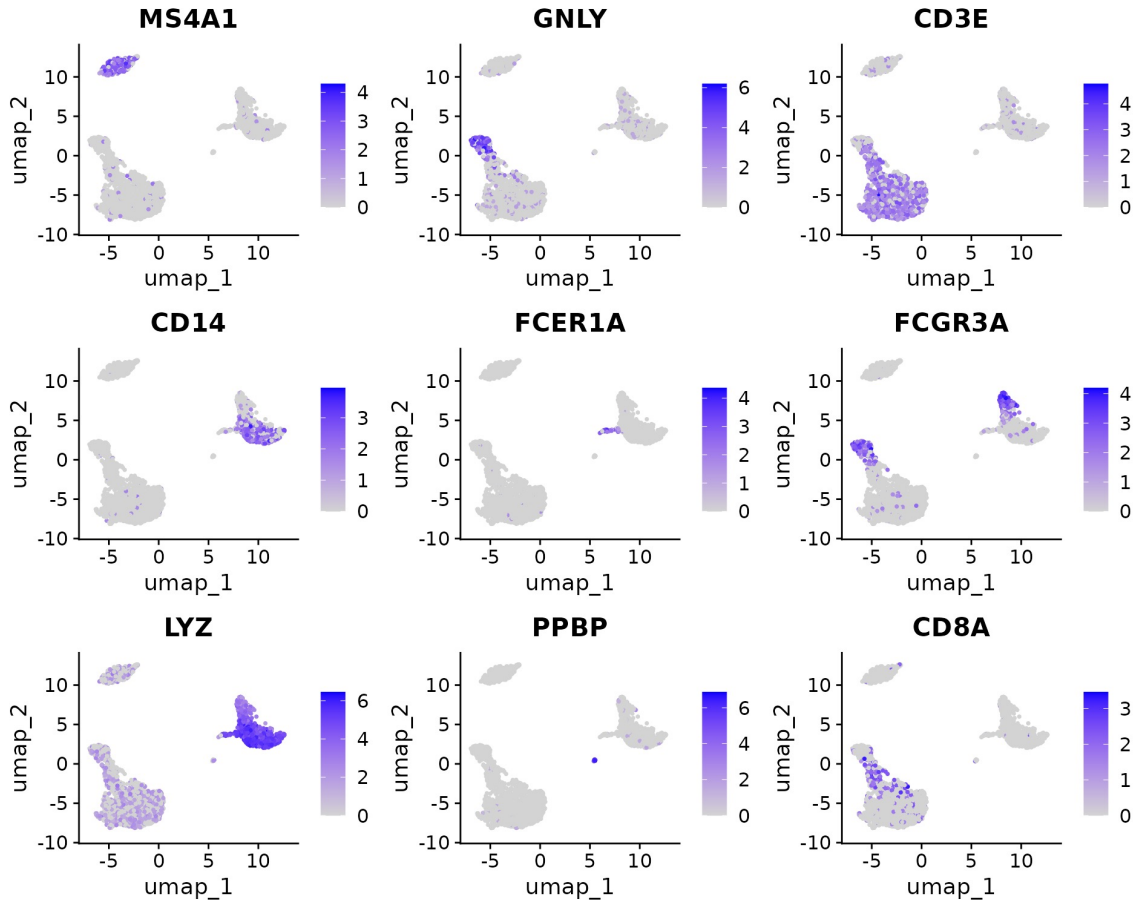


Credit: https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

EMBL-EBI

Marker gene (cluster biomarker) detection

- Identify genes whose expression defines each cluster that has been identified
- Visualization of marker genes



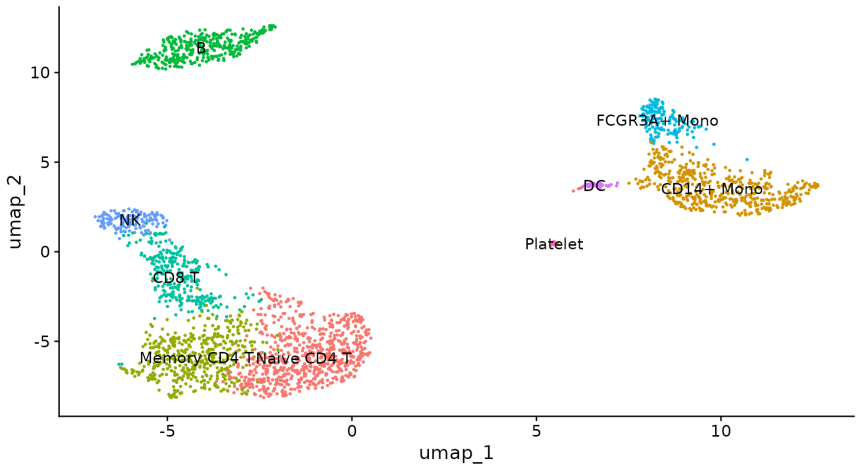
Credit: https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

EMBL-EBI

Cell type/state discovery

- Manual annotation using your biological knowledge with marker genes

Cluster ID	Markers	Cell Type
0	IL7R, CCR7	Naive CD4+ T
1	CD14, LYZ	CD14+ Mono
2	IL7R, S100A4	Memory CD4+
3	MS4A1	B
4	CD8A	CD8+ T
5	FCGR3A, MS4A7	FCGR3A+ Mono
6	GNLY, NKG7	NK
7	FCER1A, CST3	DC
8	PPBP	Platelet



Credit: https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

EMBL-EBI

Cell type/state discovery

- Manual annotation using your biological knowledge with marker genes
- Automatic annotation using tools, e.g., SingleR package from Bioconductor
  - SingleR provides references for human and mouse
  - SingleR allows users to create their own references using any scRNA-seq datasets with labels

Data retrieval	Organism	Samples	Sample types	main labels	fine labels	Cell type focus
`HumanPrimaryCellAtlasData()`	human	713	microarrays of sorted cell populations	37	157	Non-specific
`BlueprintEncodeData()`	human	259	RNA-seq	24	43	Non-specific
`DatabaseImmuneCellExpressionData()`	human	1561	RNA-seq	5	15	Immune
`NovershternHematopoieticData()`	human	211	microarrays of sorted cell populations	17	38	Hematopoietic & Immune
`MonacoImmuneData()`	human	114	RNA-seq	11	29	Immune
`ImmGenData()`	mouse	830	microarrays of sorted cell populations	20	253	Hematopoietic & Immune
`MouseRNAseqData()`	mouse	358	RNA-seq	18	28	Non-specific

Credit: https://bioconductor.statistik.tu-dortmund.de/packages/3.10/bioc/vignettes/SingleR/inst/doc/SingleR.html

Reference

- The content and figures have been credited with links on the relevant slides. These sources are also excellent material for further reading.
- Most of the content, presented in the accompanying demonstration material, originate from Seurat official tutorial and Babraham scRNA-seq analysis tutorial with some of the text copied with a few edits.