Introduction to single-cell RNA-seq analysis

BaRC Hot Topics

Bioinformatics and Research Computing

Whitehead Institute

March 7th 2019

http://barc.wi.mit.edu/hot_topics/

WHITEHEAD INSTITUTE

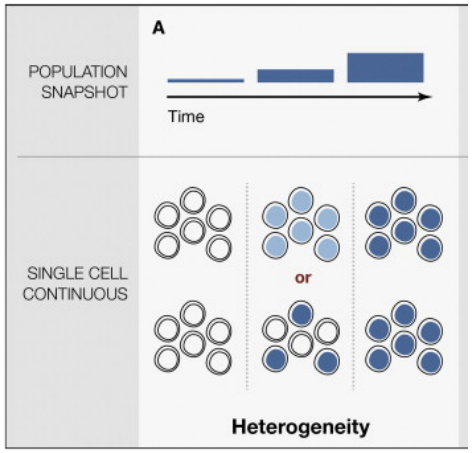
Outline

- Introduction to single-cell RNA-seq data analysis
  - Overview of scRNA-seq technology, cell barcoding, UMIs
  - Experimental design
  - Analysis pipeline
    - Preprocessing and quality control
    - Normalization
    - Dimensionality reduction
    - Clustering of cells
    - Trajectory inference
    - Differential expression and functional annotation

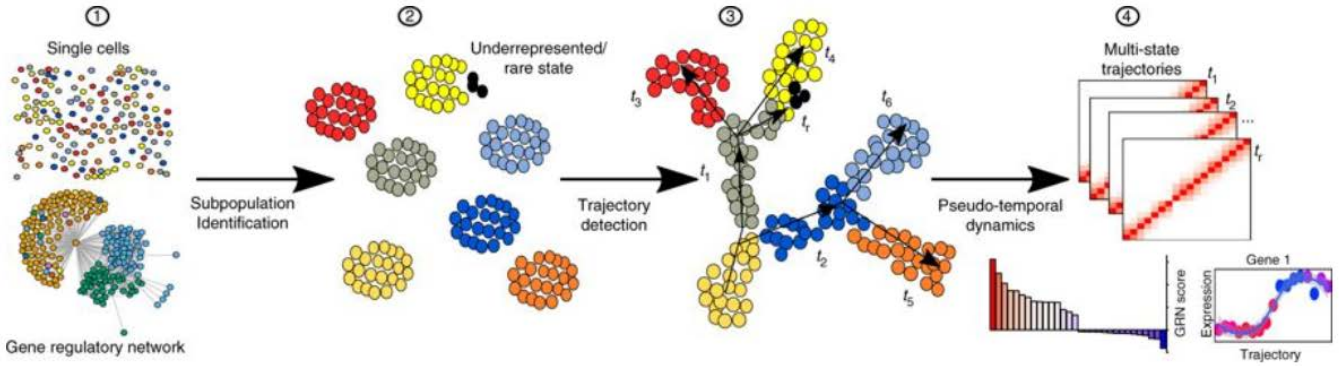
- Hands-on analysis using the package Seurat

Why do single cell RNA-seq?

- Identify expression profiles of individual cells (that may be missed with bulk RNA-seq)
- Discover of new cell states/types
- Order cells within a developmental trajectory

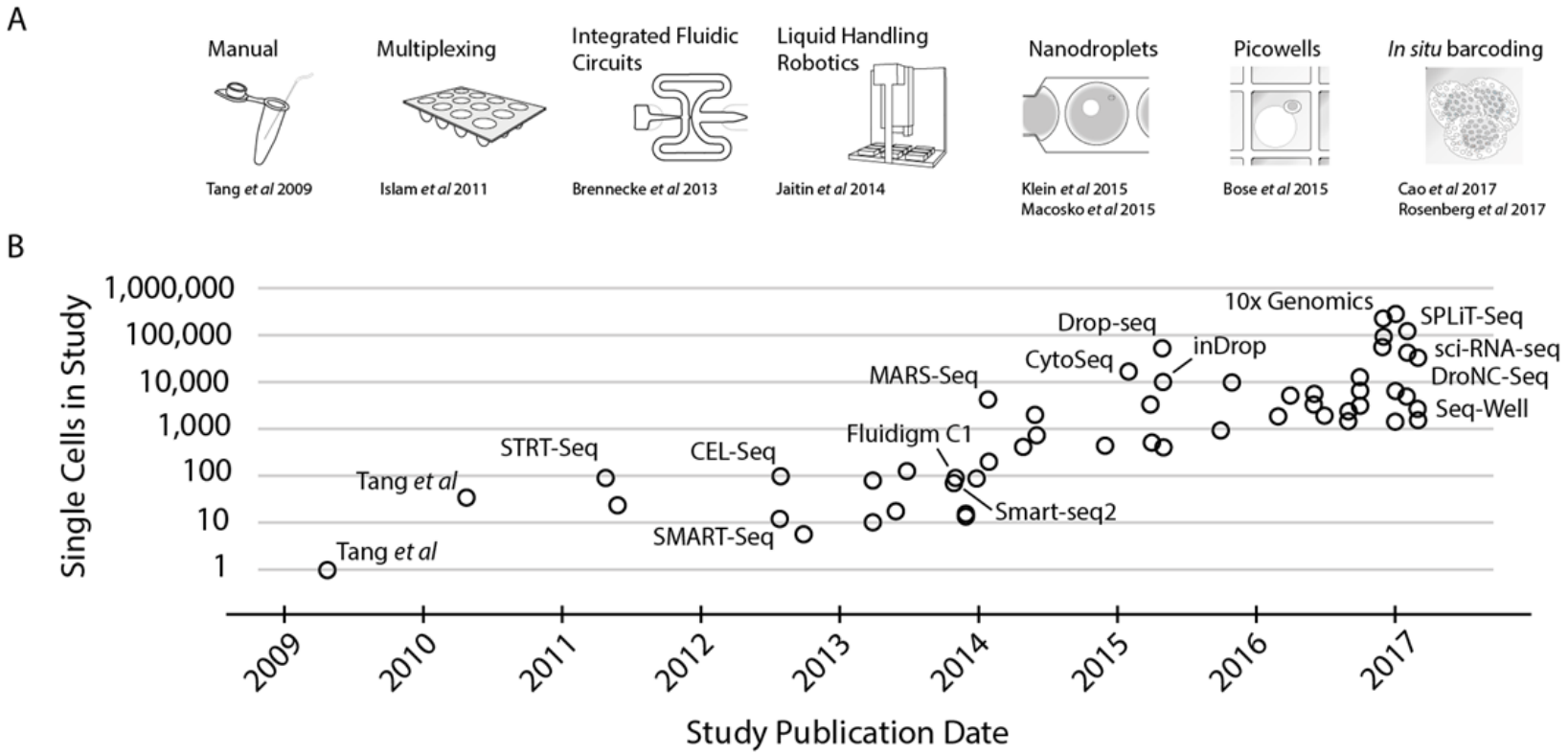


Etzrodt, *Cell Stem Cell* 2014



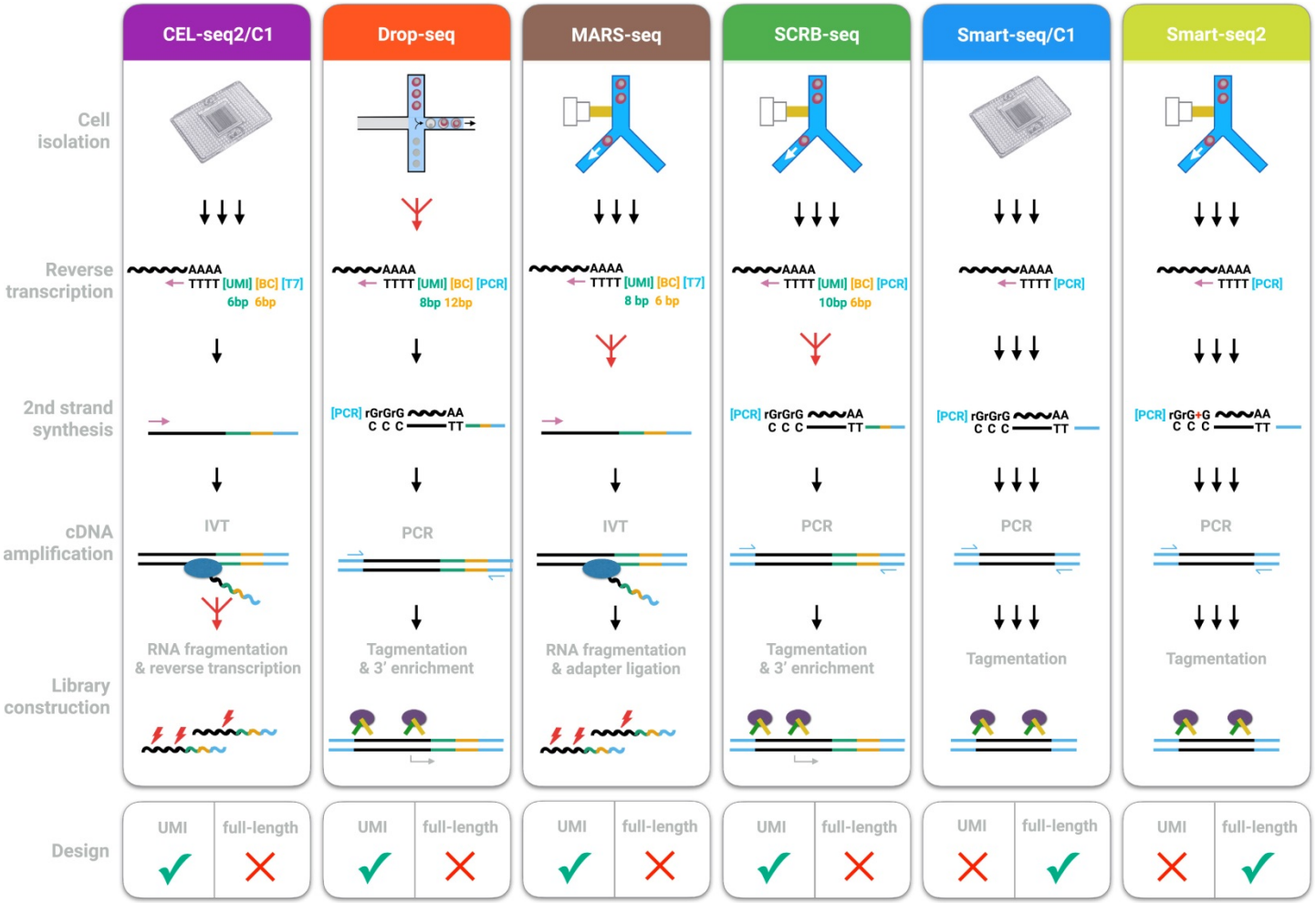
Lummertz da Rocha, *Nature Communications* 2018

Advances on scRNA-seq technology



Svensson, Vento-Tormo, and Teichmann, *arXiv:1704.01379v2*

Library preparation steps



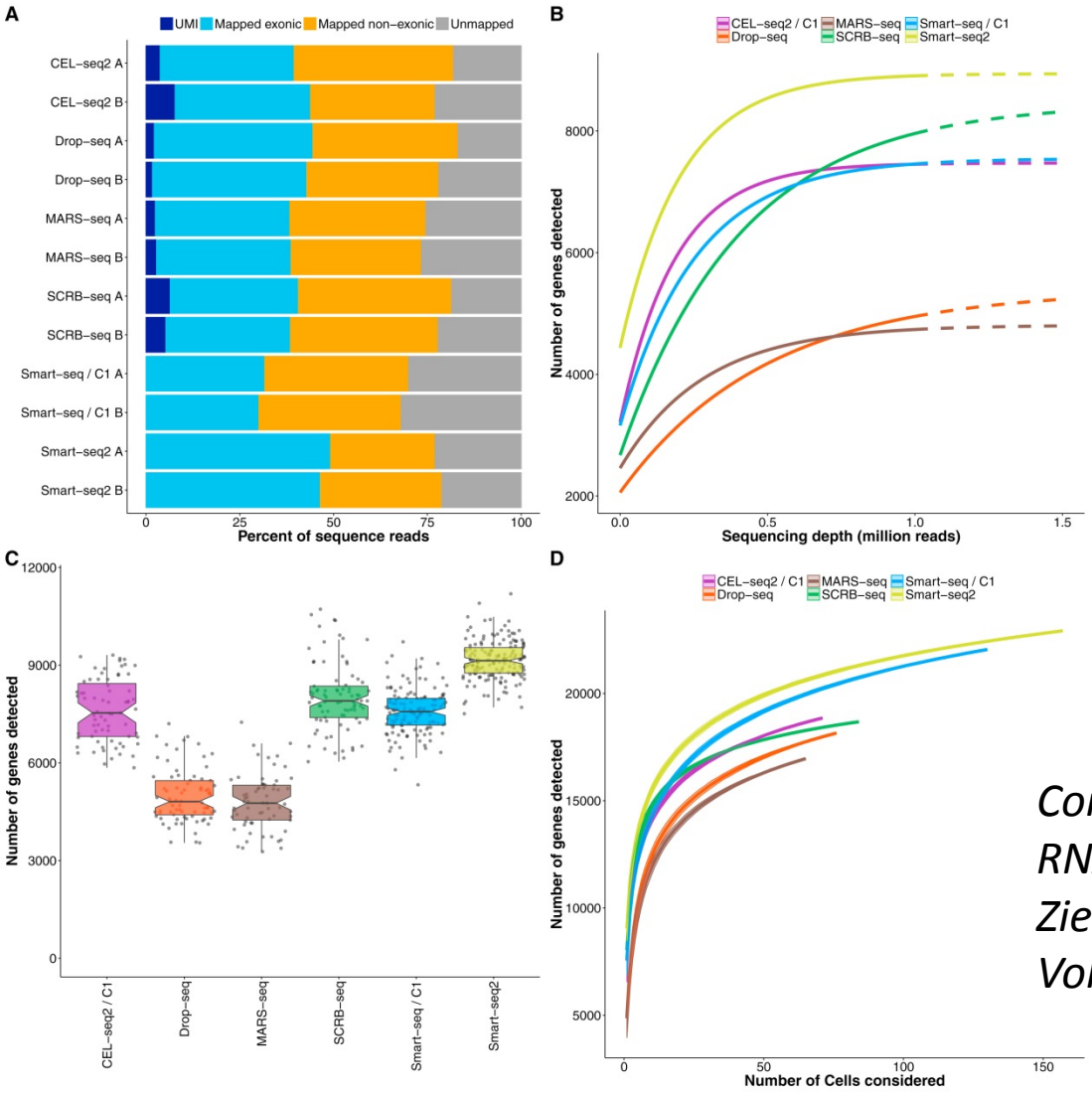
Comparative Analysis of Single-Cell RNA Sequencing Methods
Ziegenhain et. al, Molecular Cell
Volume 65, Issue 4, 16 February 2017,

Features of scRNA-seq methods

Name	Transcript coverage	Strand specificity	Positional bias	UMI possible?
Tang method	Nearly full-length	No	Strongly 3´	No
Smart-seq	Full-length	No	Medium 3´	No
Smart-seq2	Full-length	No	Weakly 3´	No
STRT-seq & STRT/C1	5´-only	Yes	5´-only	Yes
CEL-seq	3´-only	Yes	3´-only	No
CEL-seq2	3´-only	Yes	3´-only	Yes
MARS-seq	3´-only	Yes	3´-only	Yes
CytoSeq	Pre-defined genes only	Yes	3´-only	Yes
Drop-seq/InDrop	3´-only	Yes	3´-only	Yes

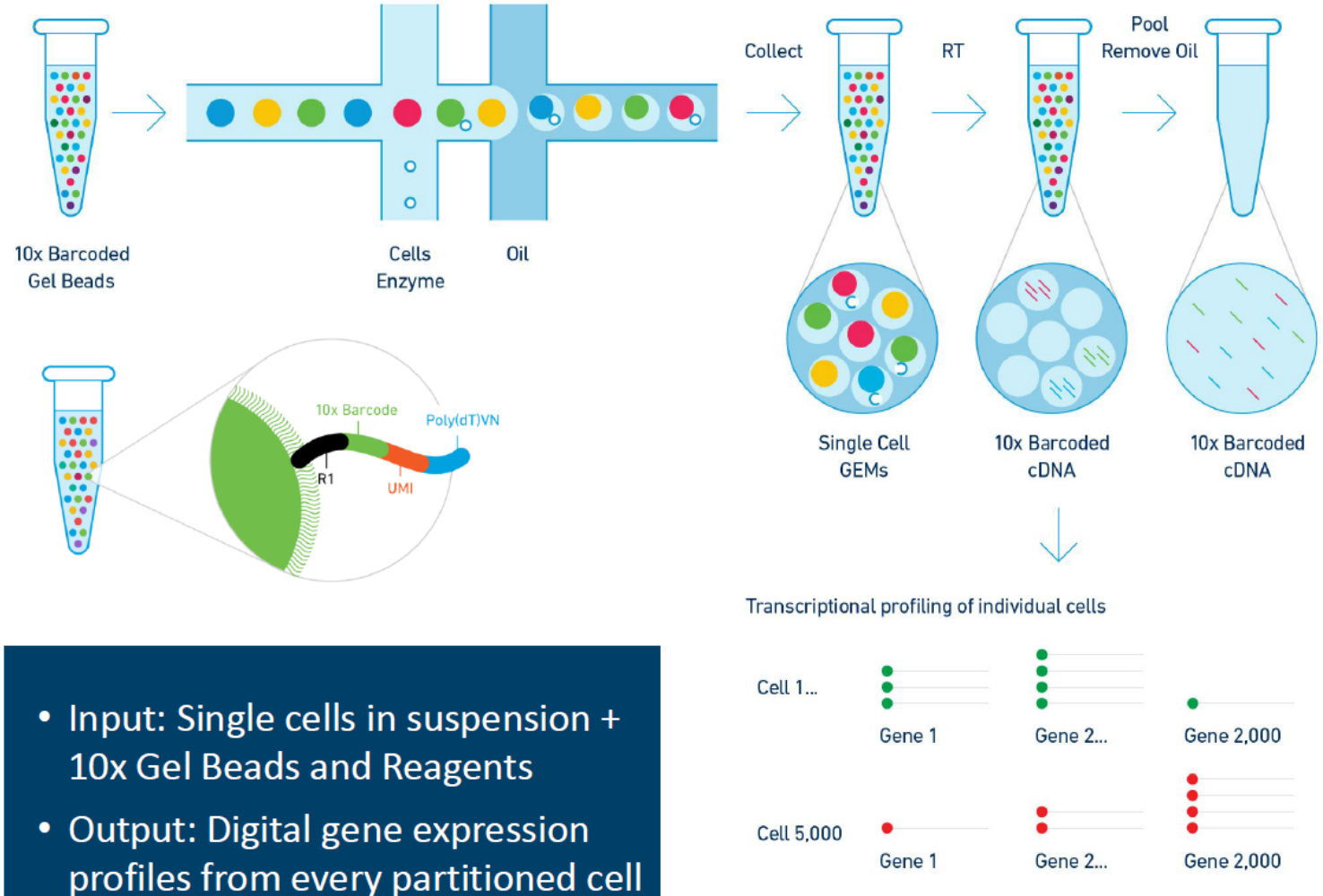
Single-cell RNA-sequencing: The future of genome biology is now
Simone Picelli, RNA Biology, Volume 14, 2017 - Issue 5

Sensitivity of scRNA-seq methods



Comparative Analysis of Single-Cell RNA Sequencing Methods
Ziegenhain et. al, Molecular Cell Volume 65, Issue 4, 16 Feb 2017

Single Cell Digital Gene Expression



- Input: Single cells in suspension + 10x Gel Beads and Reagents
- Output: Digital gene expression profiles from every partitioned cell

https://www.10xgenomics.com/videos/training-modules/

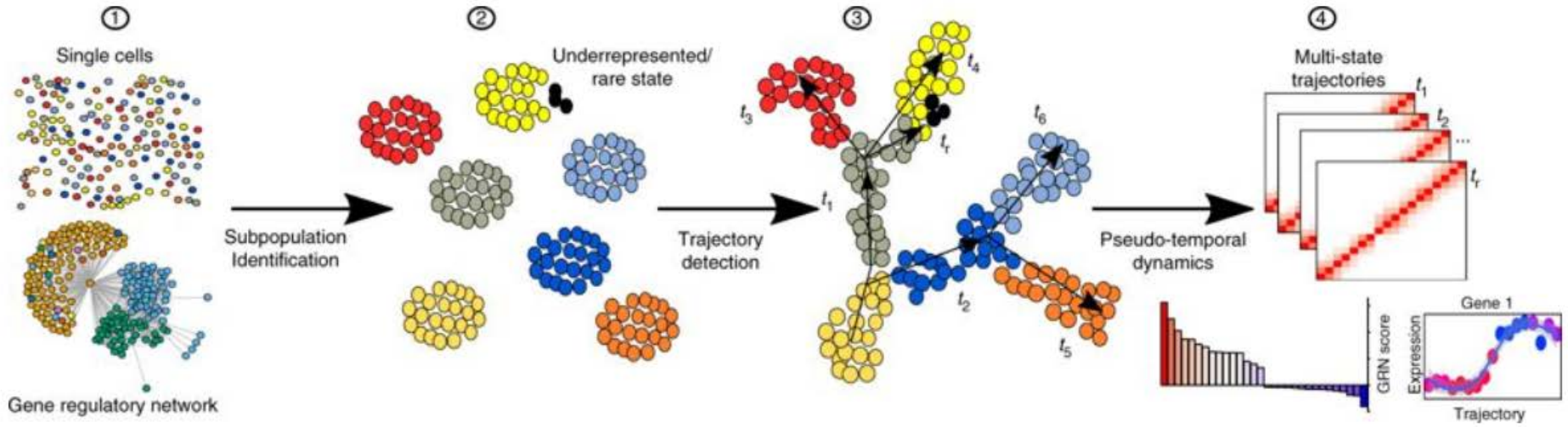
Libraries Compatible with Illumina® Sequencers

Sequencing Read	Description	Length	Notes
Read 1	Cell Barcode+UMI read	26 nt	Extra cycles are unused
i7 Index	Sample Index read	8 nt	Extra cycles are unused
Read 2	RNA read	98nt (recommended)	Length may vary

Validated on: HiSeq 2500 (RR and HO), HiSeq 3000/4000, NextSeq 500/550, MiSeq

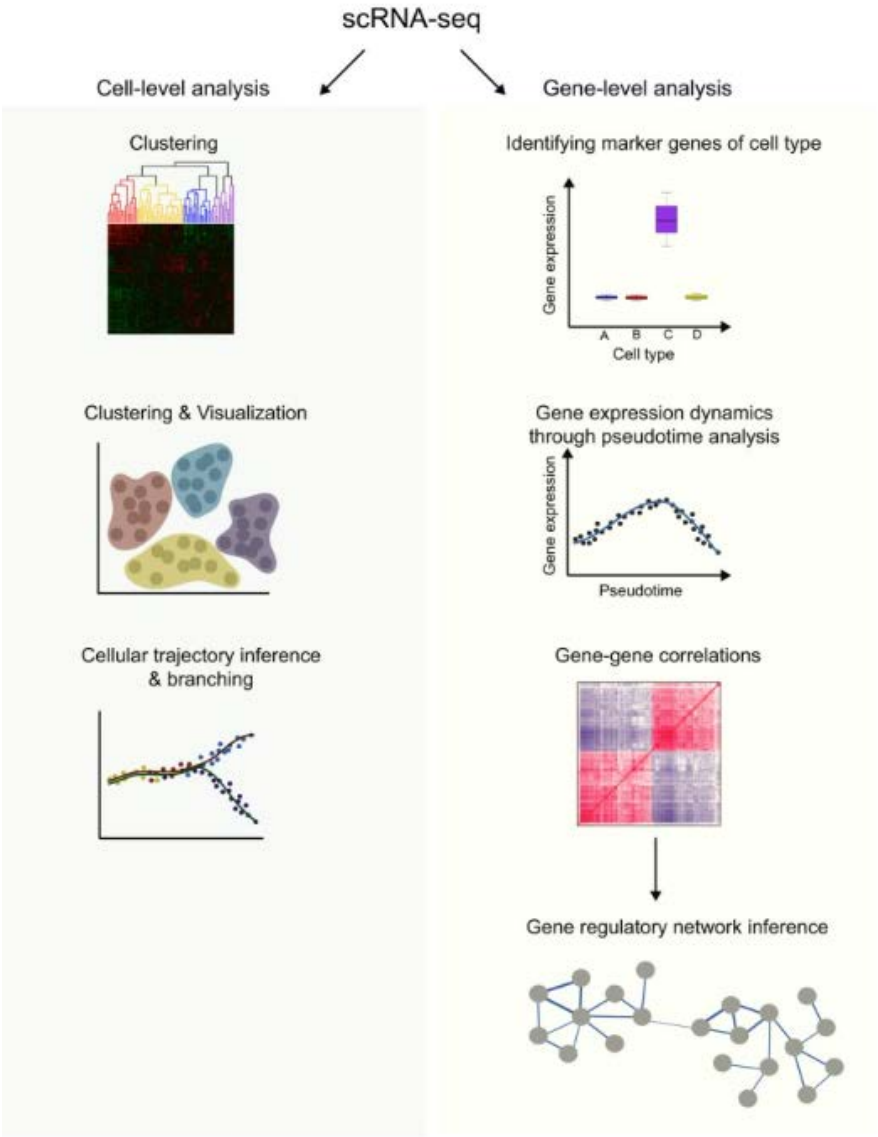
https://www.10xgenomics.com/videos/training-modules/

Goals of scRNA-seq analysis methods



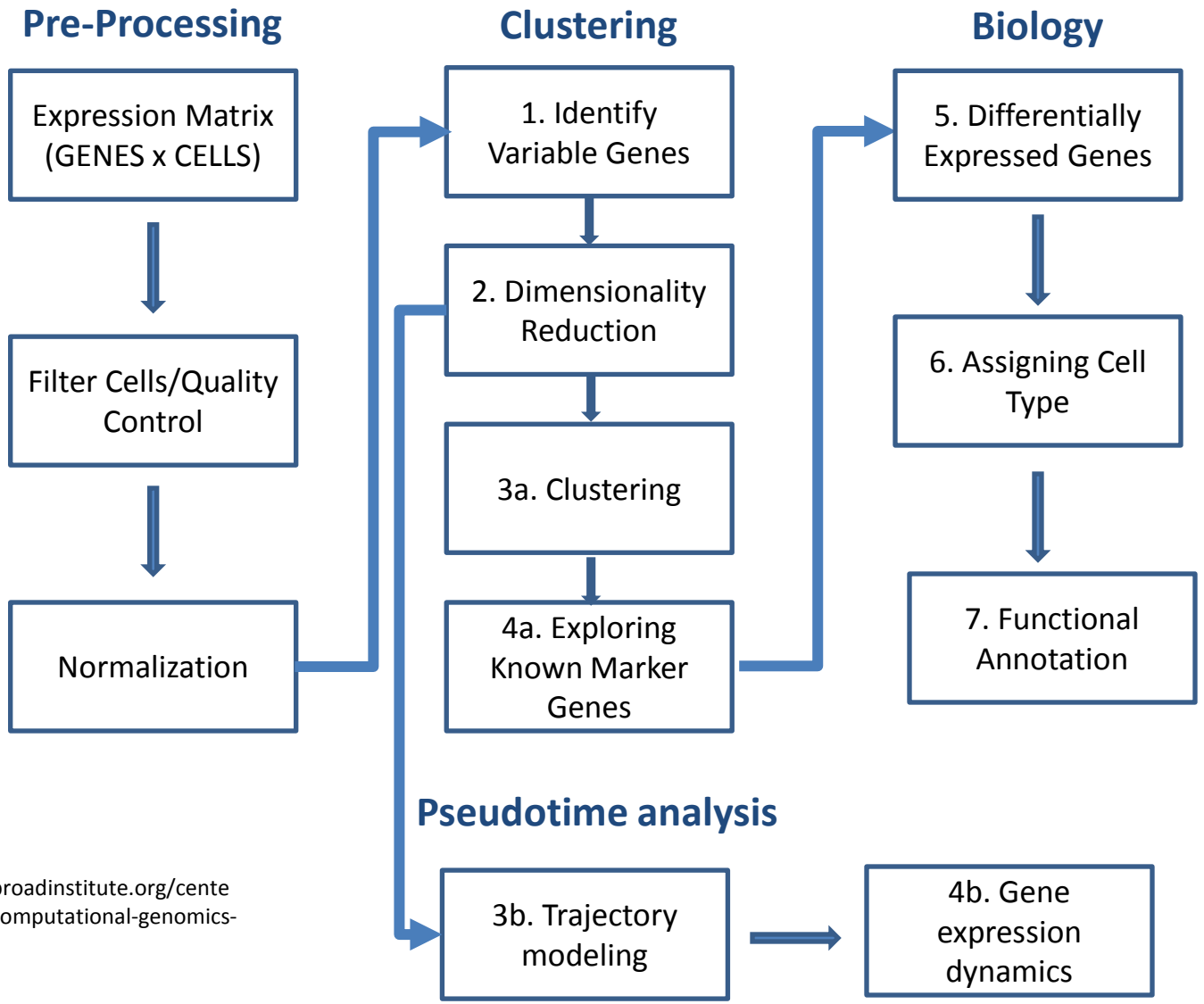
Lummertz da Rocha, *Nature Communications* 2018

Goals of scRNA-seq analysis methods



Computational approaches for interpreting scRNA-seq data, Rostom et al. FEBS Letters, Volume: 591, Issue: 15.

Analysis pipeline



Adapted from https://www.broadinstitute.org/center-cell-circuits-computational-genomics-workshop

Technical challenges

- Data is noisy due to
  - cDNA amplification bias
  - mRNA capture efficiency
  - drop outs: large number of genes with 0 counts due to limiting mRNA. Zero expression doesn't mean the gene isn’t on.
- Cells can change or die during isolation.

Experimental design

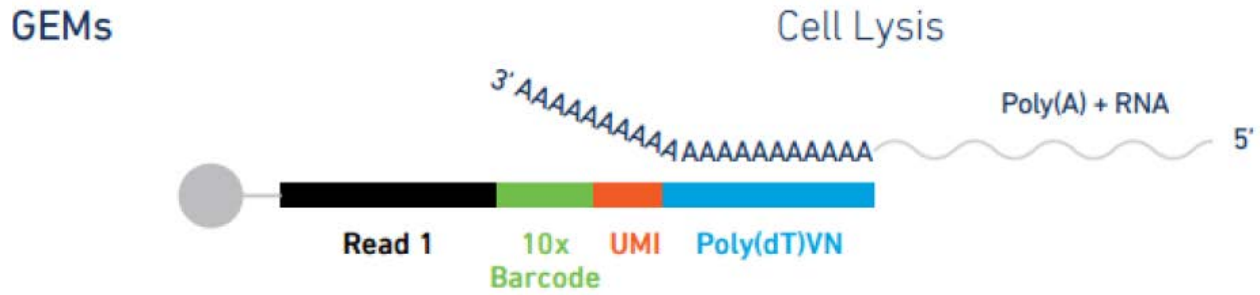
- Process your samples in a way that the condition can not be confounded with a batch effect, like processing date, facility, or reagents used.
  - *i.e.* If you have to process your cells in several batches, each batch should contain an equal number of cells from each condition.
- If you are comparing your data to published data you may have to remove batch effects.
  - R packages like Combat can be used for this (https://www.rdocumentation.org/packages/sva/versions/3.20.0/topics/ComBat)
  - See “Dealing with confounders” section of the "Analysis of single cell RNA-seq data" course (Hemberg Group).

Preprocessing for Smart-seq2

- Demultiplexing: assign all the reads with the same cell barcode to the same cell. Done at the sequencing facility.

- We can check the quality of the reads with *FastQC* and the library composition with *FastQ Screen* as we would do with bulk RNA-seq.

Preprocessing for technologies using Unique Molecular Identifiers (UMIs)



- Demultiplexing: assign all the reads with the same cell barcode to the same cell.
- Remove PCR duplicates: if several reads have the same UMI and map to the same location in the genome, keep only one.
  - Cell range software for 10x data (run by the genome technology core)
  - Drop-seq tools for drop-seq and seq-well data

https://www.10xgenomics.com/videos/training-modules/

Demultiplexing and counting 10x data

Cell Ranger™ Pipelines

Pipeline	Functionality
cellranger mkfastq	Barcode-aware demultiplexing from BCL to FASTQ

cellranger count	• Read-level analysis of a single library
	  • Transcriptome alignment with STAR
	  • Barcode processing
	  • Gene counting
	• Produces gene/cell matrix
	• Produces expression analysis and static visualizations
	• Produces .cloupe file for Loupe™ Cell Browser

https://www.10xgenomics.com/videos/training-modules/

CellRanger web summary

10x GENOMICS™ Cell Ranger

SUMMARY ANALYSIS

Estimated Number of Cells

2,580

Mean Reads per Cell	Median Genes per Cell
97,620	2,559

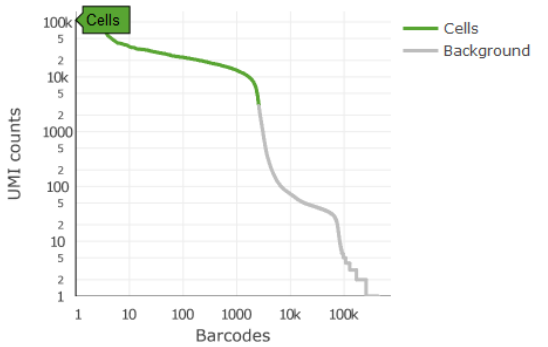
Sequencing

Number of Reads	251,861,835
Valid Barcodes	96.1%
Sequencing Saturation	78.1%
Q30 Bases in Barcode	94.7%
Q30 Bases in RNA Read	66.6%
Q30 Bases in Sample Index	87.9%
Q30 Bases in UMI	94.5%

Mapping

Reads Mapped to Genome	86.4%
Reads Mapped Confidently to Genome	80.5%
Reads Mapped Confidently to Intergenic Regions	2.7%
Reads Mapped Confidently to Intronic Regions	6.7%
Reads Mapped Confidently to Exonic Regions	71.0%
Reads Mapped Confidently to Transcriptome	69.2%
Reads Mapped Antisense to Gene	0.8%

Cells

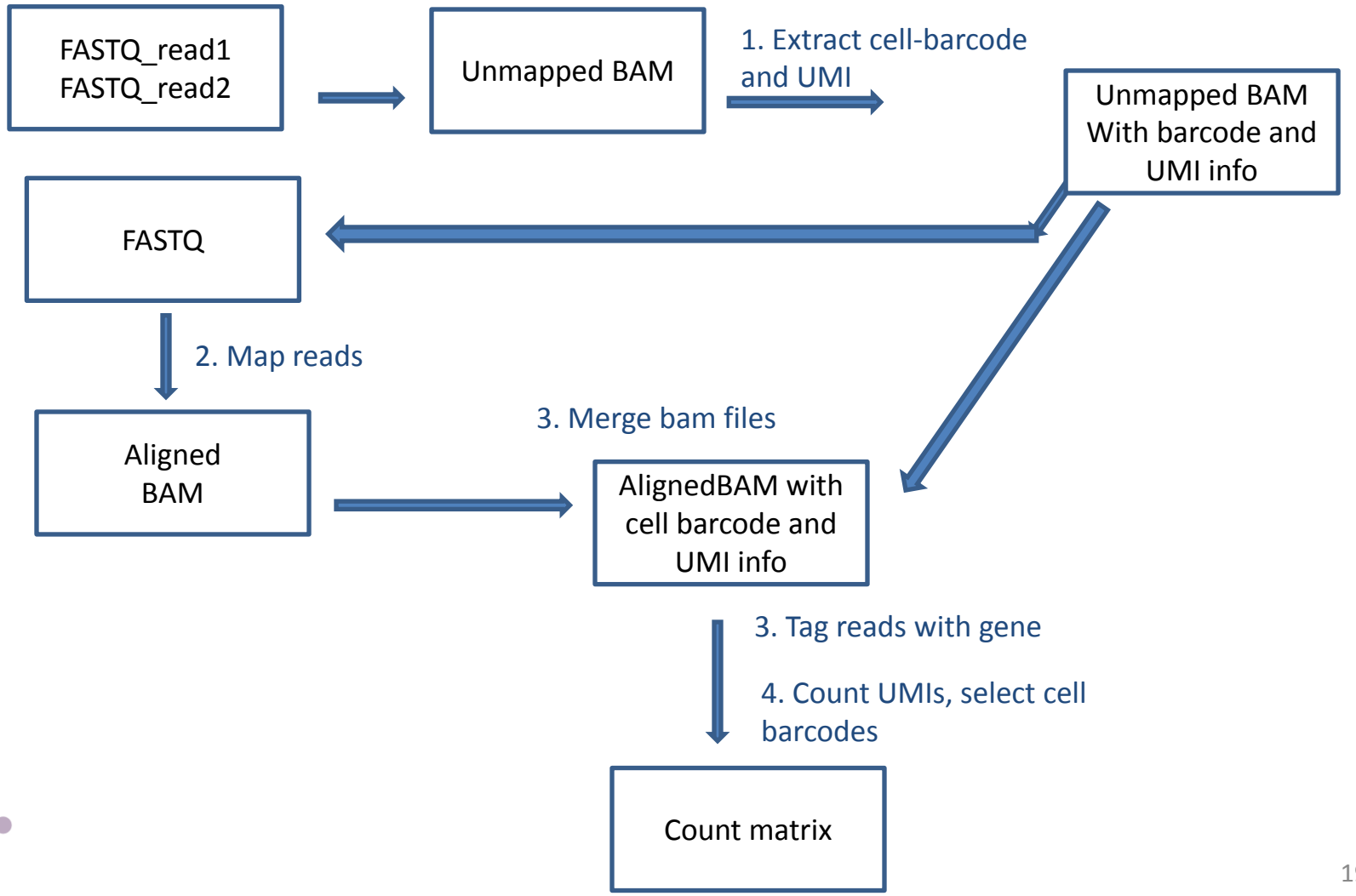


Estimated Number of Cells	2,580
Fraction Reads in Cells	88.1%
Mean Reads per Cell	97,620
Median Genes per Cell	2,559
Total Genes Detected	15,848
Median UMI Counts per Cell	11,729

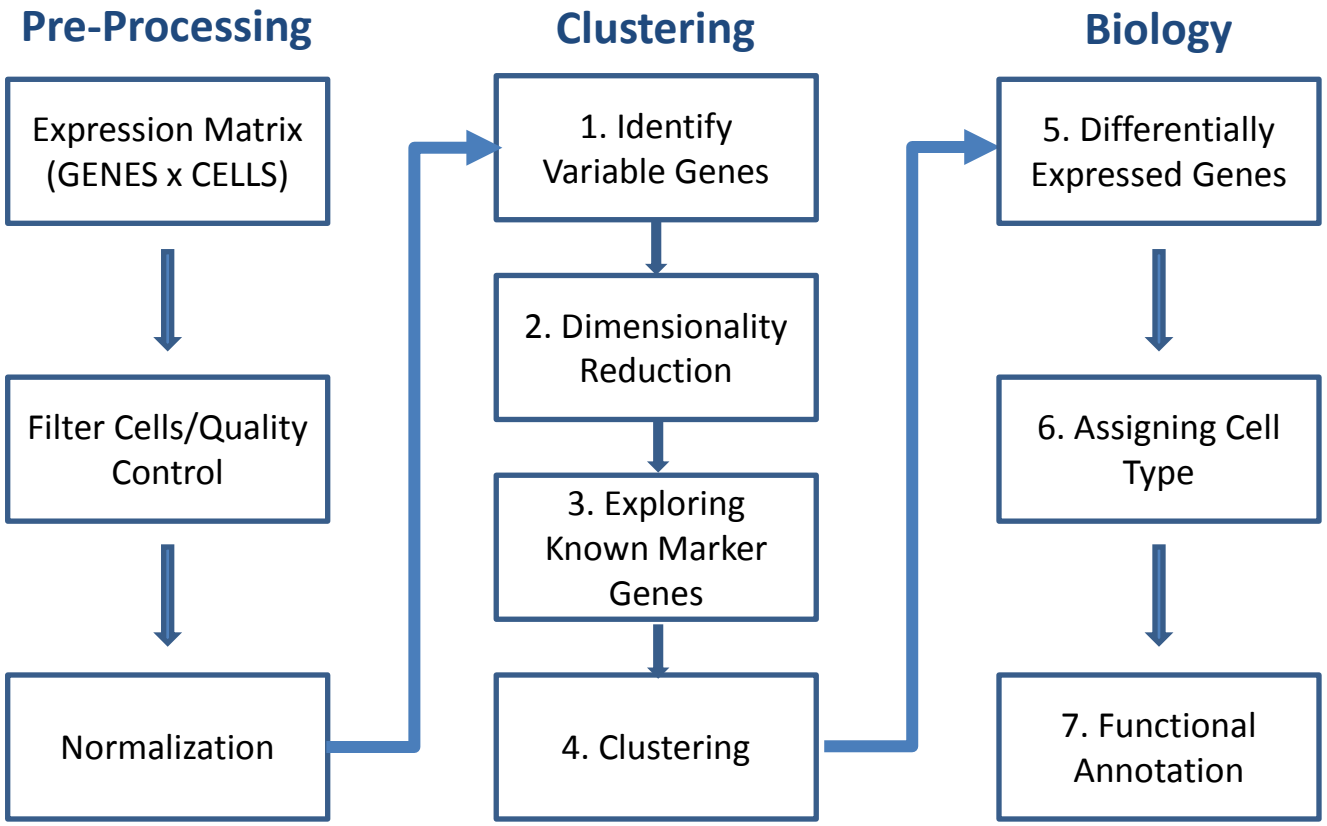
Sample

Name	L21_314
Description	
Transcriptome	mm10
Chemistry	Single Cell 3' v2
Cell Ranger Version	2.1.1

Demultiplexing and counting Drop-seq or Seq-well data



Analysis pipeline



Adapted from
https://www.broadinstitute.org/center-cell-circuits-computational-genomics-workshop

Quality control and filtering

- Quality control
  - Number of reads per cell
  - Number of genes detected per cell
  - Proportion of reads mapping to mitochondrial reads
- Remove cells with poor quality
  - Filter out cells with percentage of mitochondrial reads higher than a cut off
  - Filter out cells with less than a lower threshold on the number of genes or counts per cell
- Remove doublets (two cells captured with one bead in the droplet)
  - Filter out cells with more than an upper threshold on the number of genes or counts per cell in your data
  - More sophisticated way of removing doublets
    - https://github.com/JonathanShor/DoubletDetection
    - https://github.com/AllonKleinLab/scrublet
    - https://www.biorxiv.org/content/early/2018/06/20/352484

Normalization

Correct for sequencing depth (i.e. library size) of each cell so we can compare across cells

1. Normalize gene expression for each cell by total expression
2. Multiply by a scale factor (*i.e.* 10,000).
3. Log transform the scaled counts

This is the log normalization implemented in Seurat

Clustering and Biology: What do you want to learn from the experiment?

- Classify cells and discover new cell populations
- Compare gene expression between different cell populations
- Reconstruct developmental 'trajectories' to reveal cell fate decisions of distinct cell subpopulations

Lots of software available to analyze single-cell RNA-seq data

- Seurat
- Monocle
- ScanPy
- Destiny
- See

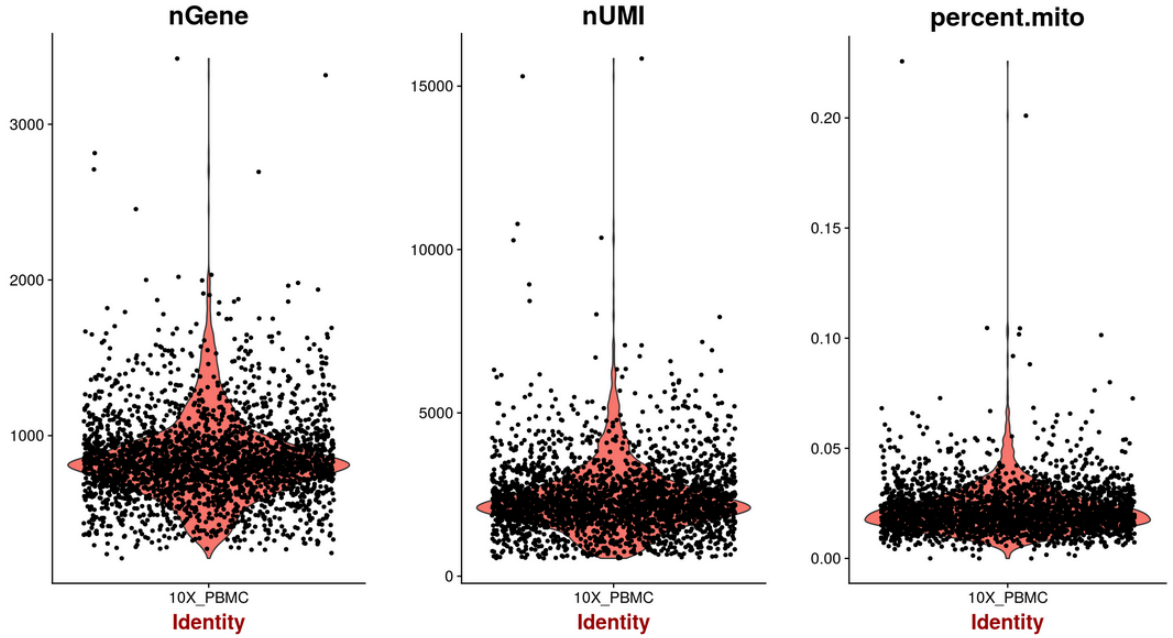
  https://github.com/seandavi/awesome-single-cell

Seurat
https://satijalab.org/seurat/

- Seurat is an R package designed for QC, analysis, and exploration of single cell RNA-seq data.
- Developed and by the Satija Lab at the New York Genome Center.
- It is well maintained and well documented.
- It has a built in function to read 10x Genomics data.
- It has implemented most of the steps needed in common analyses.

Read data and explore QC metrics plots

- Read data
  - **`Read10X()`**
  - **`read.table()`**
- Create Seurat object: **`CreateSeuratObject()`**
- Calculate the % mitochondrial genes
- Plot nUMI, nGenes and % mito to decide on cut offs



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

Select cells, normalize and scale data.

- Filter cells based on number of genes detected and percent of mitochondrial genes

```
SObj <- FilterCells(object = SObj,
    subset.names    = c("nGene","percent.mito"),
    low.thresholds  = c(4000, -Inf),
    high.thresholds = c(11000, 0.06))
```

- Normalize counts

```
SObj <- NormalizeData(object = SObj,
    normalization.method = "LogNormalize",
    scale.factor      = 1e4)
```

- Scaling the data and removing unwanted sources of variation

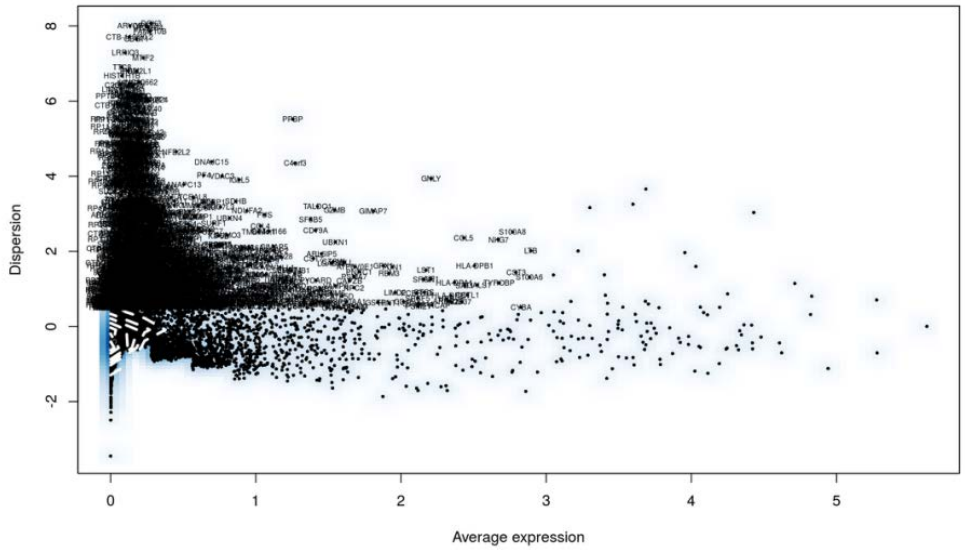
SObj <- ScaleData(object = SObj) # just scale genes across samples

SObj <- ScaleData(object = SObj, vars.to.regress = c("batch"))# remove cell-cell variation in gene expression driven by the batch/day samples were processed.

Select variable genes that will be used for dimensionality reduction

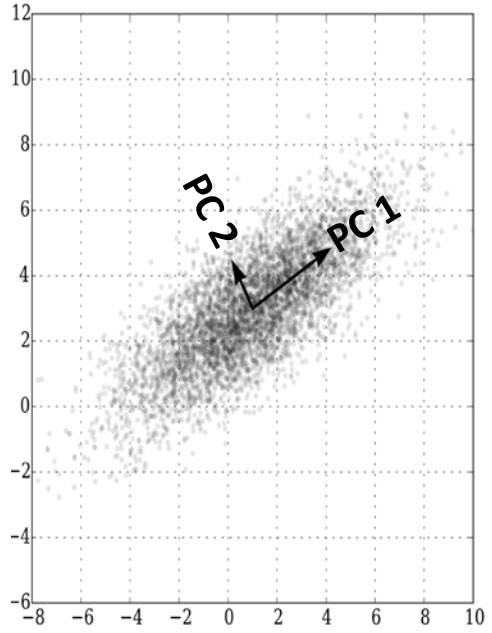
"FindVariableGenes" calculates the average expression and dispersion for each gene, places these genes into bins, and then calculates a z-score for dispersion within each bin. This helps control for the relationship between variability and average expression.

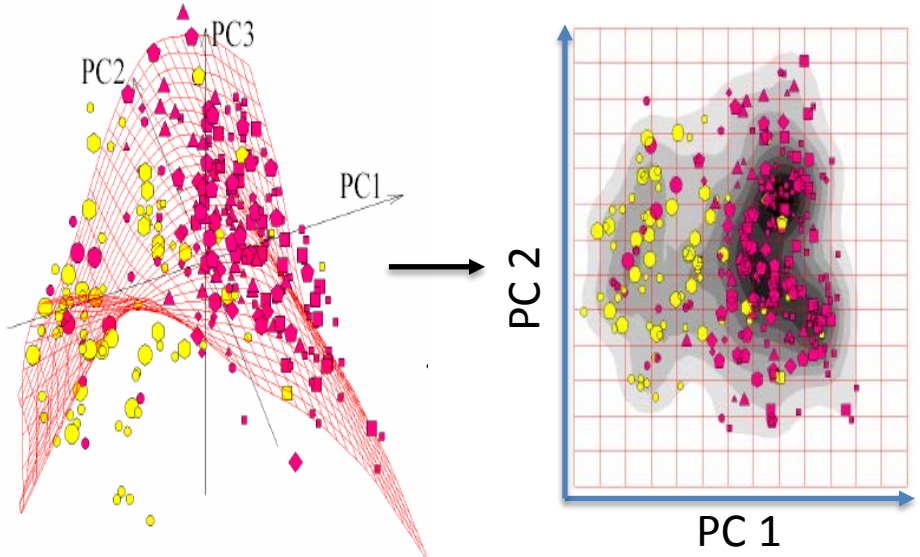
```
pbmc <- FindVariableGenes(object = pbmc, mean.function =
ExpMean, dispersion.function = LogVMR, x.low.cutoff = 0.0125,
x.high.cutoff = 3, y.cutoff = 0.5)
```




```
length(x = pbmc@var.genes)
## gives you the number of genes selected, 1838 in this example
```

Principal component analysis





Cells in 20000 (genes) dimensional space

PCA →

Cells in 10-50 principal components space

Some genes have low expression
Many genes are co-regulated

Wikipedia and adapted from Hojun Li

Other dimensionality reduction methods

Cells in 20000 (genes) dimensional space → PCA → Cells in 10-50 principal components space

How can we further summarize these multiple PCAs into just 2 dimensions?

Cells in 10-50 principal components space → tSNE, UMAP, other → Cells in 2D space

t-Distributed Stochastic Neighbor Embedding (tSNE)

- Takes a set of points in a high-dimensional space and finds a faithful representation of those points in a lower-dimensional space, typically the 2D plane.
- The algorithm is **non-linear** and adapts to the underlying data, performing different transformations on different regions.
- The t-SNE algorithm adapts its notion of “distance” to regional density variations in the data set. As a result, it naturally expands dense clusters, and contracts sparse ones, evening out cluster sizes.
- Distances between clusters might not mean anything.

https://distill.pub/2016/misread-tsne/

UMAP

Uniform manifold approximation and projection

- It is a non linear dimensionality reduction algorithm.
- Preserves the local structure but also the global structure and the continuity of the cell subsets better.
- See PMID: 30531897 for comparison of Seurat and UMAP.

Dimensionality reduction and clustering

- Linear dimensionality reduction: PCA

```
pbmc <- RunPCA(object = pbmc, pc.genes = pbmc@var.genes, do.print =
TRUE, pcs.print = 1:5, genes.print = 5)
```

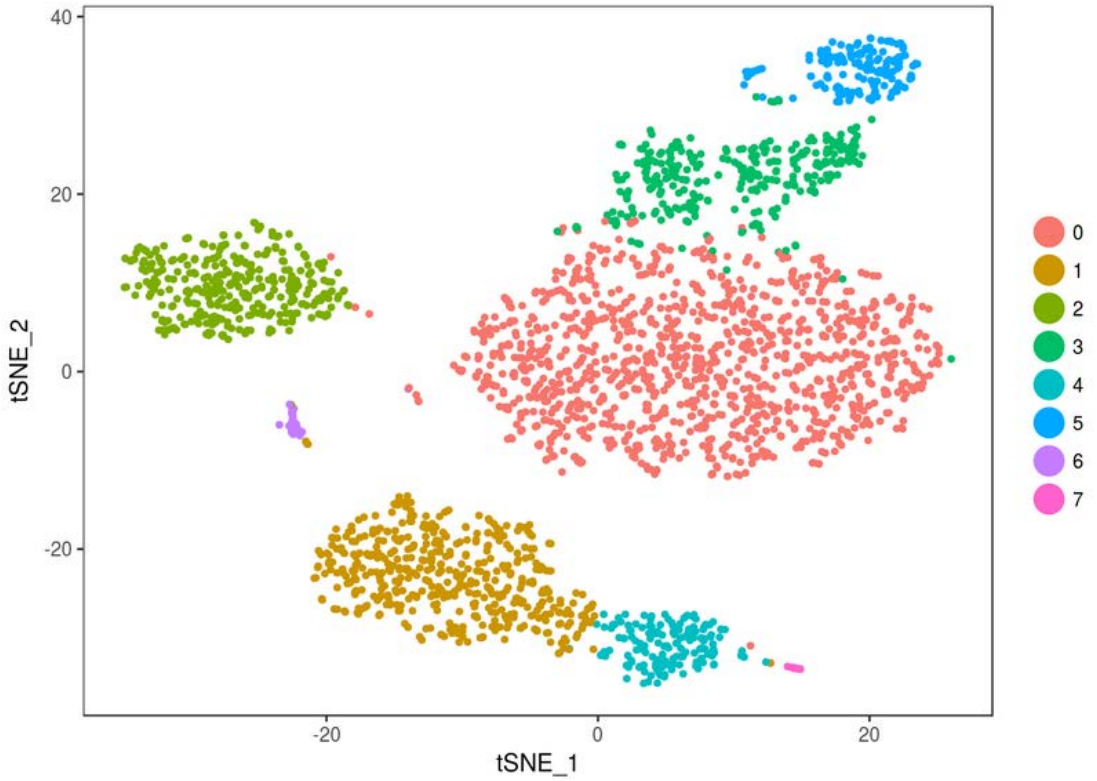
- Cluster the cells and run non-linear dimensional reduction (tSNE)

```
pbmc <- FindClusters(object = pbmc, reduction.type = "pca", dims.use =
1:10, resolution = 0.6, print.output = 0, save.SNN = TRUE)

pbmc <- RunTSNE(object = pbmc, dims.use = 1:10, do.fast = TRUE)
```

Visualize the tSNE plot

```
TSNEPlot(object = pbmc)
```



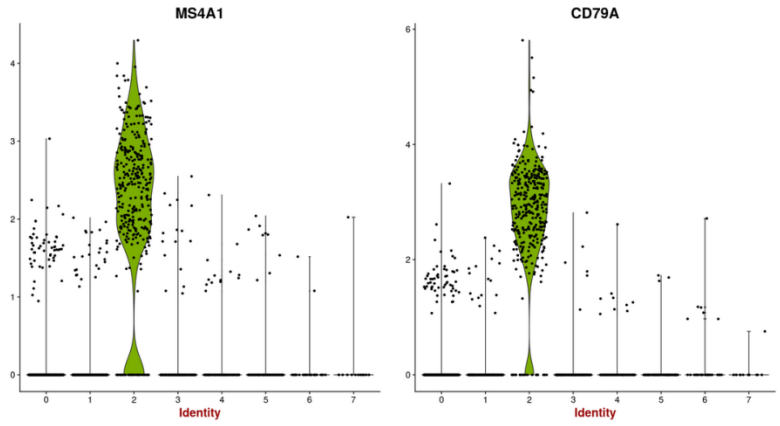
Location of cells on the plot is coming from the tSNE plot, color is coming from the "**`FindClusters`**" the algorithm.

Differential expression and visualization

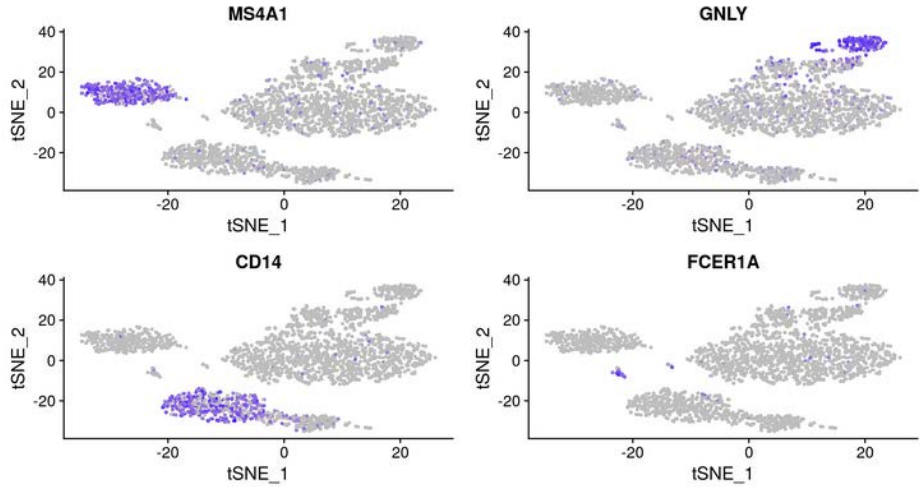
- Finding differentially expressed genes (cluster biomarkers)

```
# find all markers distinguishing cluster 5 from clusters 0 and 3
cluster5.markers <- FindMarkers(object = pbmc, ident.1 = 5, ident.2 = c(0, 3), min.pct = 0.25)
```

- Visualize DE genes



VlnPlot(object = pbmc, features.plot = c("MS4A1", "CD79A"))

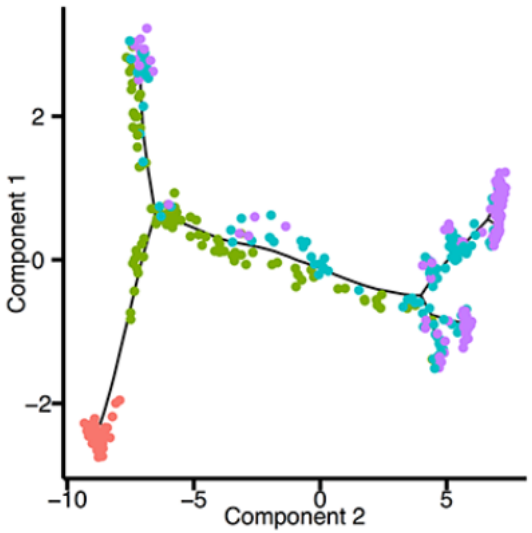


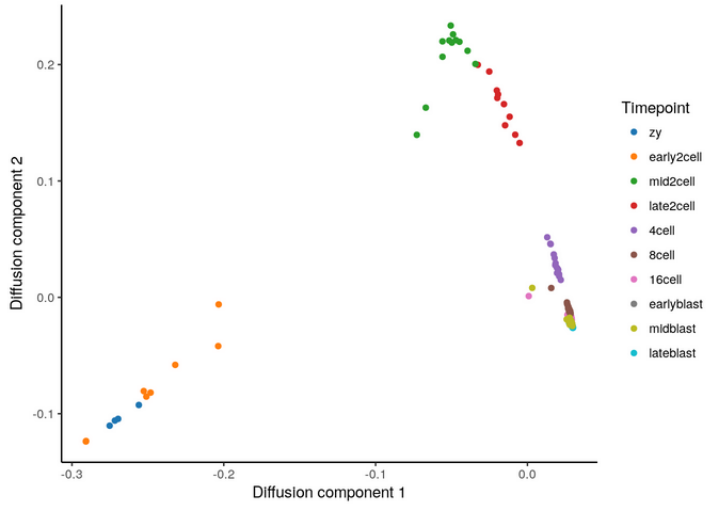
FeaturePlot(object = pbmc, features.plot = c("MS4A1", "GNLY", "CD14", "FCER1A"), cols.use = c("grey", "blue"), reduction.use = "tsne")

Reconstructing 'trajectories‘ Pseudotime analysis

Applicable when studying a process where cells change continuously. For example cell differentiation during development, or cell response to a stimulus.

- Monocle
- TSCAN
- Slicer
- Slingshot
- Diffusion map
  - Scanpy,
  - Seurat
  - density





References and resources

- A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. PMID: 28821273
- "Analysis of single cell RNA-seq data" course (Hemberg Group).
- Single cell RNA sequencing - NGS Analysis - NYU
- 2017/2018 Single Cell RNA Sequencing Analysis Workshop (UCD,UCB,UCSF)
- seandavi/awesome-single-cell
- Broad Institute single cell portal
- Tabula Muris (https://tabula-muris.ds.czbiohub.org/)

Exercises

- Goal:
  - To walk you through an example analysis of scRNA-seq data.
    - Exploring the data
    - Performing quality control
    - Identifying cell type subsets.
  - To introduce you to scRNA-seq analysis using the Seurat package.
- We will be analyzing the a dataset of Non-Small Cell Lung Cancer Cells (NSCLC) freely available from 10X Genomics (https://support.10xgenomics.com/single-cell-vdj/datasets/2.2.0/vdj_v1_hs_nsclc_5gex)