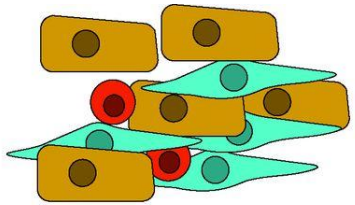
מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

Single Cell Analysis (Seurat)

An Introduction to deep-sequencing analysis for biologists 2021

Dena Leshkowitz
Bioinformatics Unit

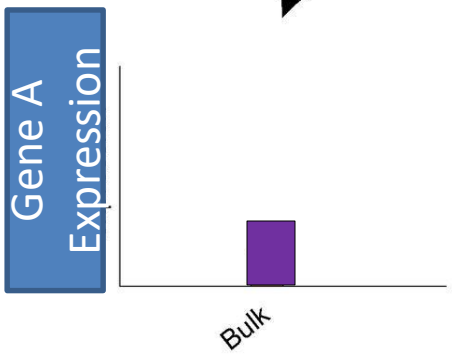
Why Perform Single Cell Analysis ?

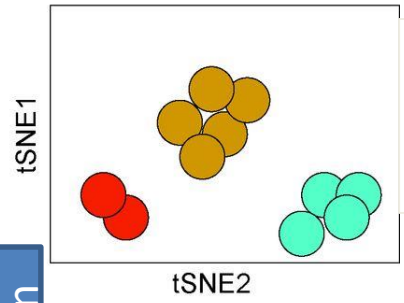


Bulk RNA Sequencing
(RNA-Seq)

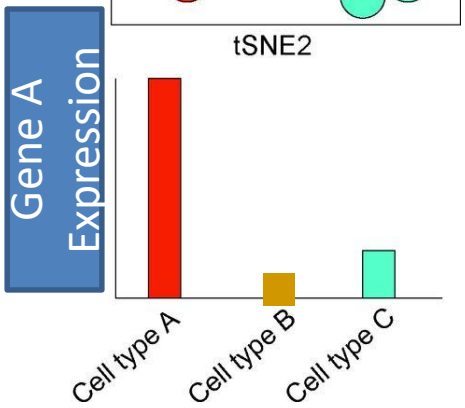
Single Cell RNA Sequencing
(scRNA-Seq)

Result:
Gene expression
is an average
across **all cell types**





Identify cells with a similar expression profile



Result: Gene expression is an average for a certain cell type

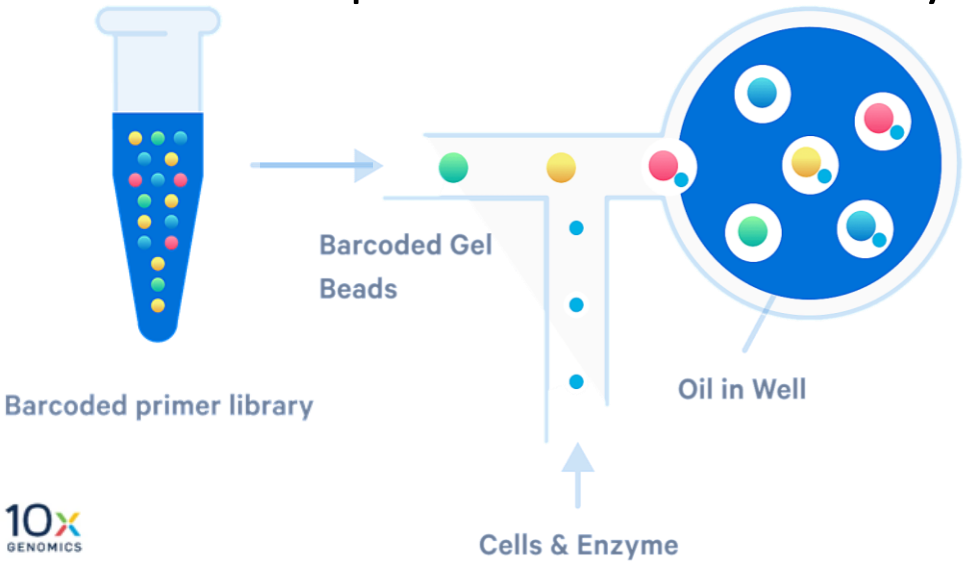
scRNA-Seq allows us to reveal previously unknown cell types or cell states in a complex tissue

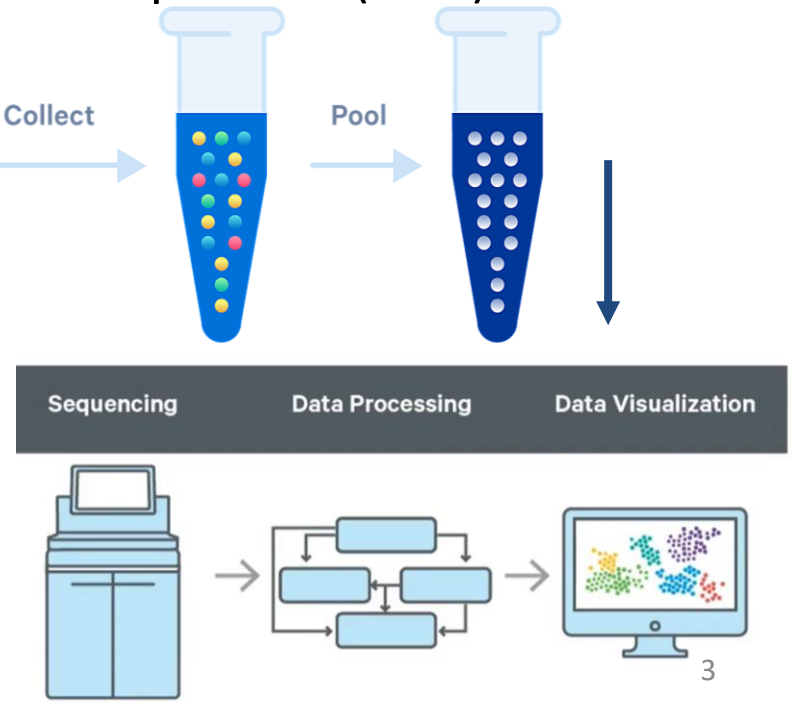
Modified - Evan Der et al. Lupus Sci Med 2019;6:e000329

How is scRNA-Seq done?

10x Genomics single-cell RNA Sequencing (scRNA-seq) technology:

- Microfluidics partitioning to **capture** single cells in droplets
- In the droplet the cell undergoes lysis and reverse transcription, in a process that tags the cDNA with a **barcode**.
  - All transcripts from the same cell will get the same barcode
  - Each transcript is also tagged with a UMI
- All cDNA is pooled to create a library and is sequenced (NGS)

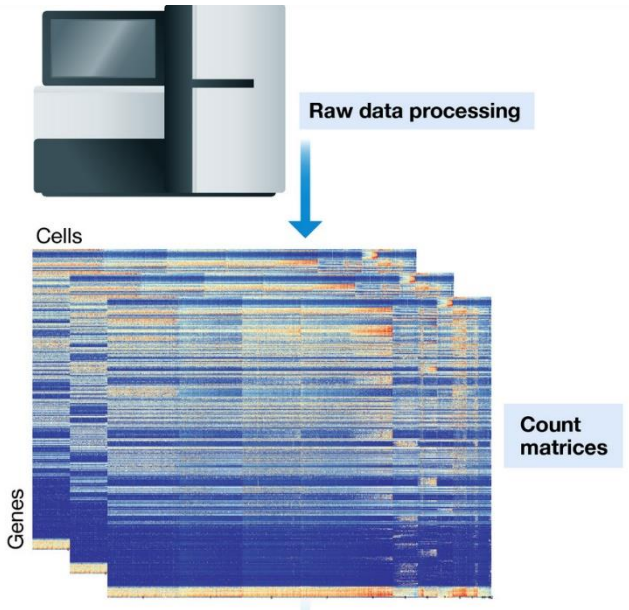


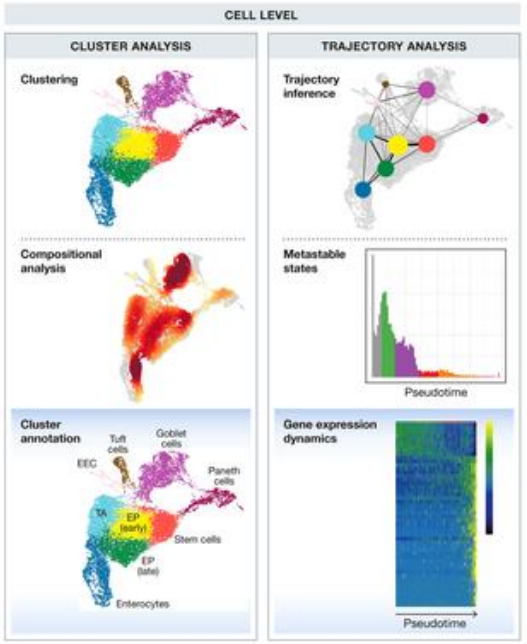


Lecture Outline

- scRNA analysis: From count matrix to biological knowledge
- Multiomics analysis
  - CITE-Seq
  - Multiome – study example

From the Count Matrix to Biological Knowledge





Mol Syst Biol, Volume: 15, Issue: 6, First published: 19 June 2019, DOI: (10.15252/msb.20188746)

scRNA-Seq Count Matrix
UMI counts per cell

- The count matrix after running CellRanger consists of the cell barcodes we consider as “real cells”
- The counts in CellRanger output are the UMI counts per gene
- The count matrix is large and sparse

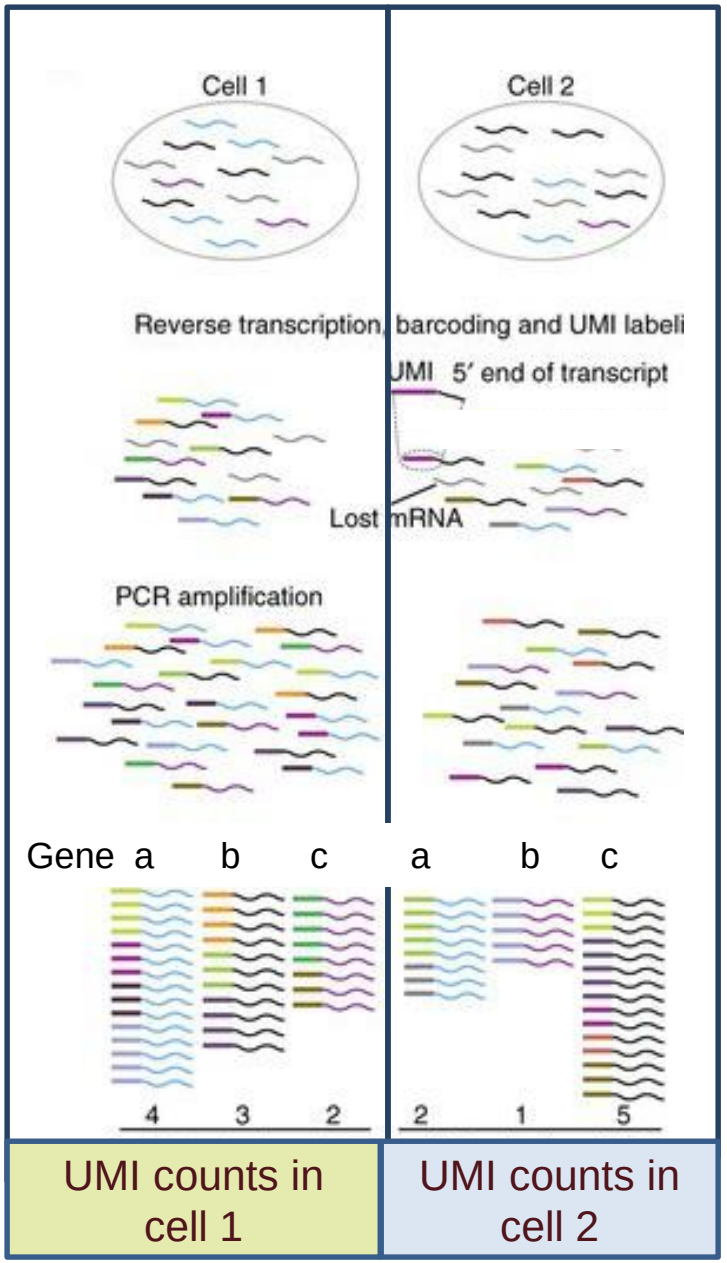
	Cell1	Cell2	…	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
…	.	.	.	.
…	.	.	.	.
…	.	.	.	.
GeneM	25	0	.	0

UMI Counts

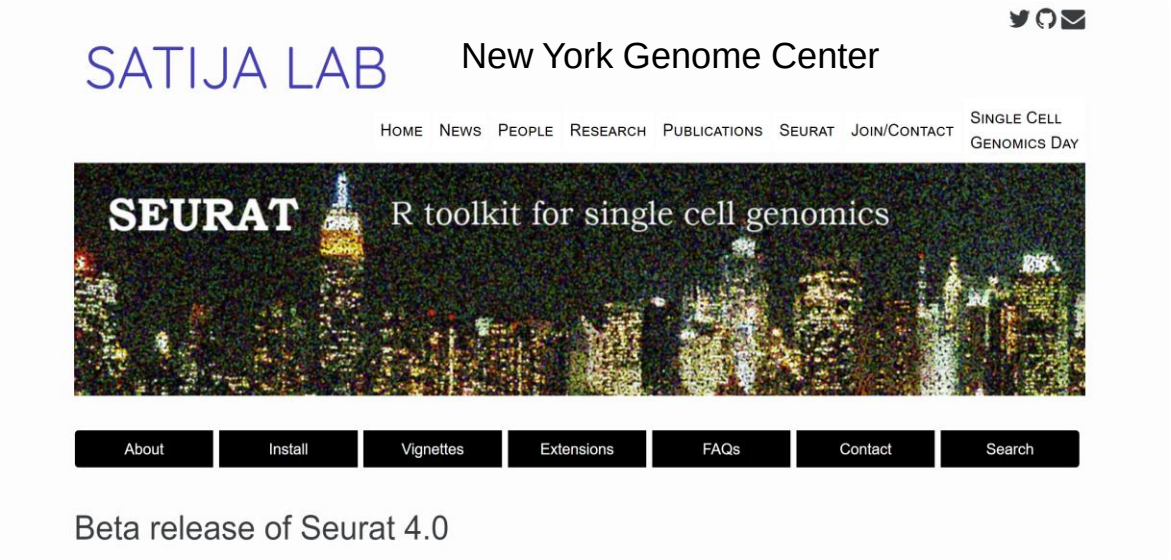
We count UMI (Unique Molecular Identifier) in order to remove PCR amplification

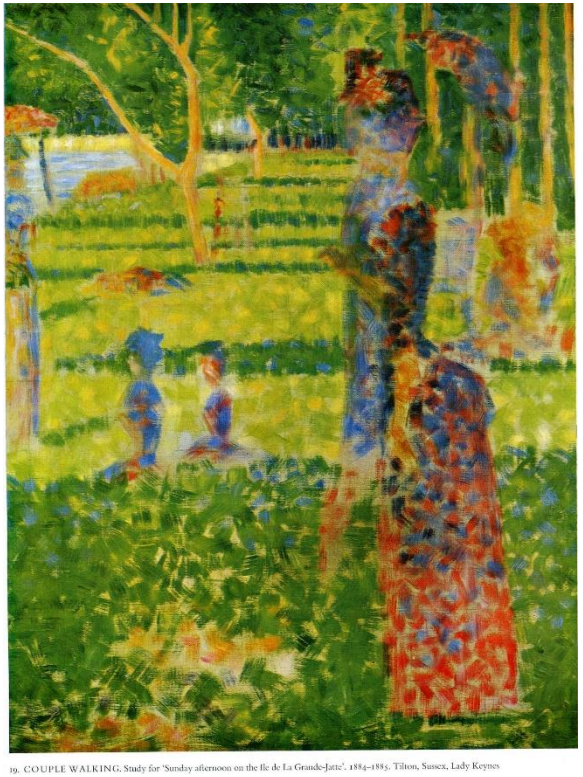
- Reads are considered duplicated, if they map to the same gene and have the same UMI
- Instead of counting reads- sequences we will count number of unique UMIs per gene per cell.

This figure is adapted from Islam et al (2014)



Analysis with Seurat Package





19. COUPLE WALKING. Study for 'Sunday afternoon on the Ile de La Grande-Jatte', 1884–1885. Tilton, Sussex, Lady Keynes

The Couple

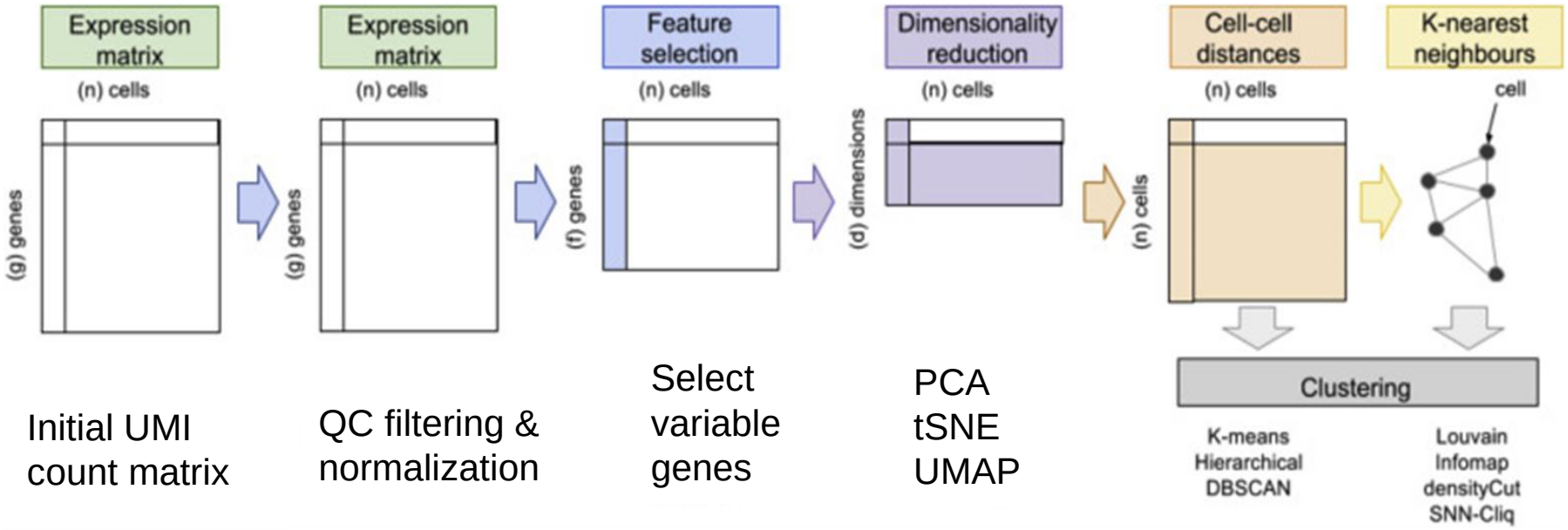
Georges Seurat

Date: 1884; France

Style: Pointillism, Neo-Impressionism

https://www.wikiart.org/en/georges-seurat/the-couple-1884

Analysis Workflow



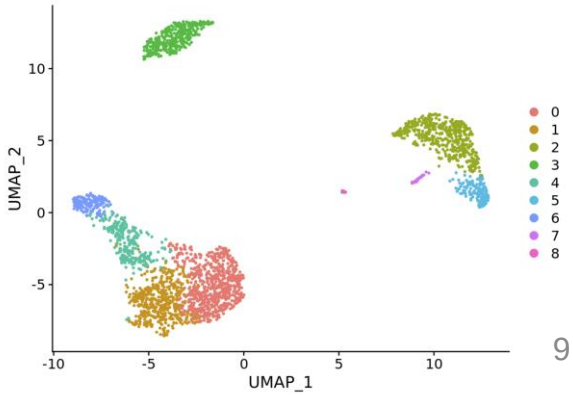
Initial UMI count matrix

QC filtering & normalization

Select variable genes

PCA
tSNE
UMAP

Modified plot from-
Andrews et al. Molecular Aspects of Medicine
Volume 59, February 2018, Pages 114-122

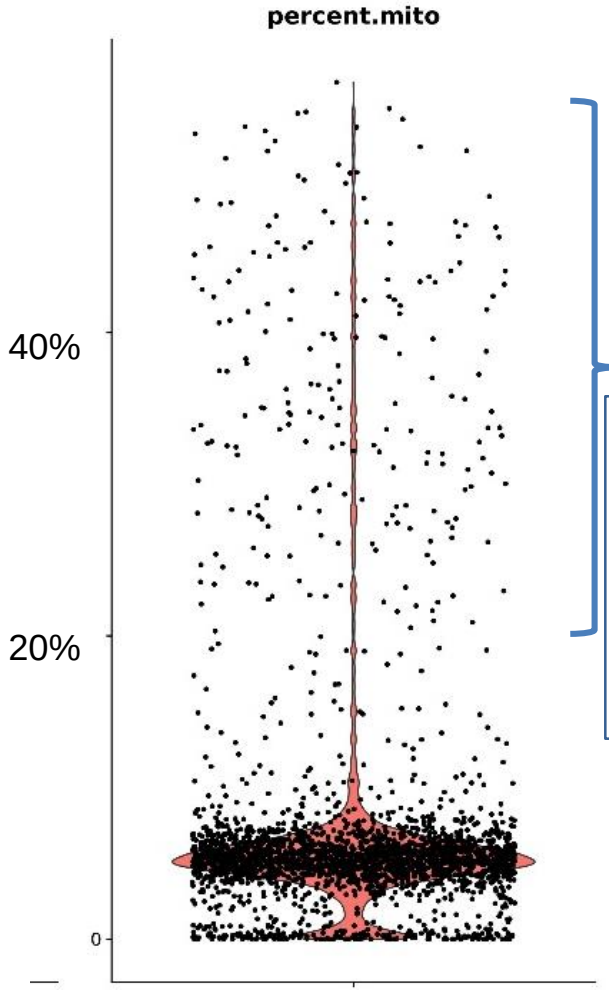


Cell Quality Control

- Removing damaged cells
  - Remove cells whose cytoplasmic mRNA has leaked out through a broken membrane, these cells can still maintain the mRNA located in the mitochondria.
- These cells will have:
  - High percent of mitochondrial gene counts (out of total UMI counts)
  - Low UMI count depth
  - Few detected genes

Violin plot

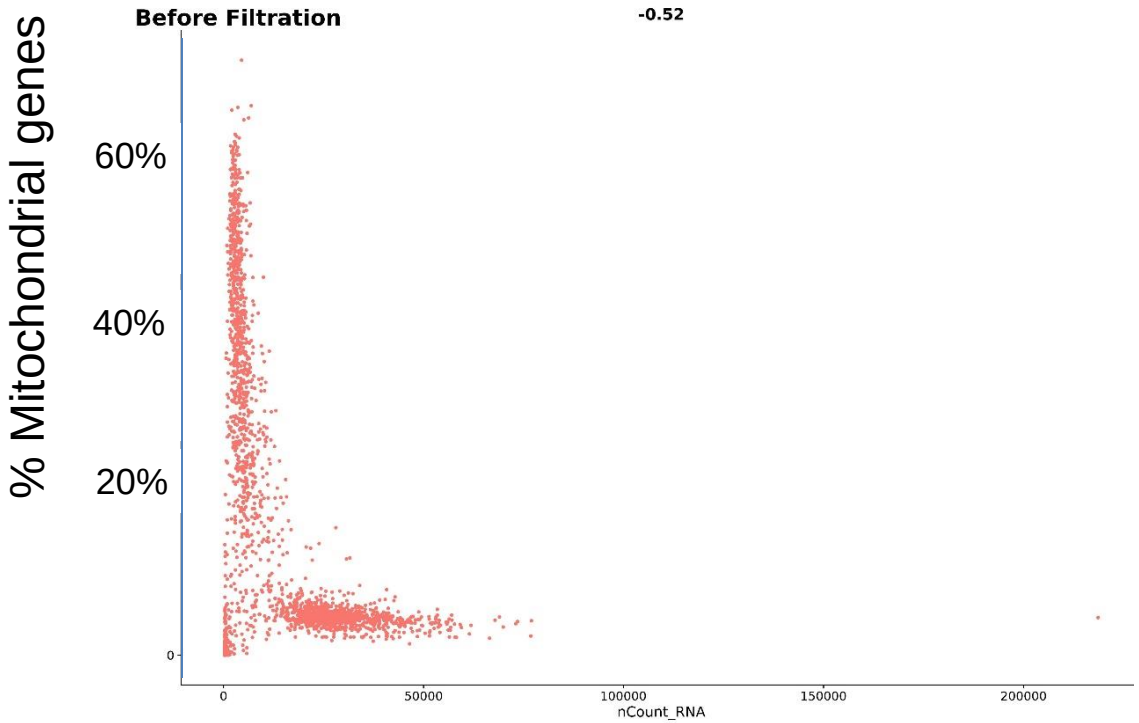
- We use violin plots to view the distribution of counts
- A violin plot is similar to a box plot, with the addition of a rotated kernel density plot on each side
- Each point represents a cell



Remove cells that have above 20% mitochondrial genes

Mitochondrial genes

The % of mitochondrial genes is anti-correlated with the expression of cellular genes



Number of total genes detected per cell

Cell Quality Control (QC)

Cell QC is commonly performed based on three criteria:

- The total number of UMI counts per cell (transcription depth)
- The total number of genes per cell
- The percent counts from mitochondrial genes

Aim: We would like to filter out the cells which are outliers, bad quality.

Question

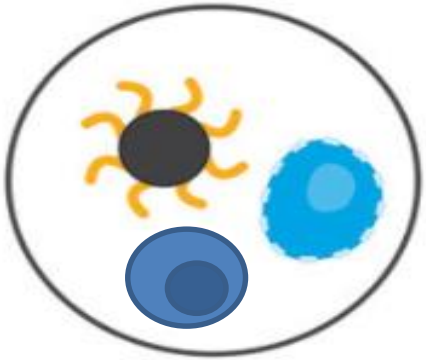
Is their an additional reason to filter out barcodes-cells with low gene or UMI counts?

Yes , these can be barcodes with ambient RNA, capturing free mRNA that leaked from damaged cells

Question

Is their a reason to filter out barcodes-cells with high gene or UMI counts?

Cell Quality Control



Removing doublets

Need to remove cells-barcodes representing a droplet that captures more than a single cell, can be identified by :

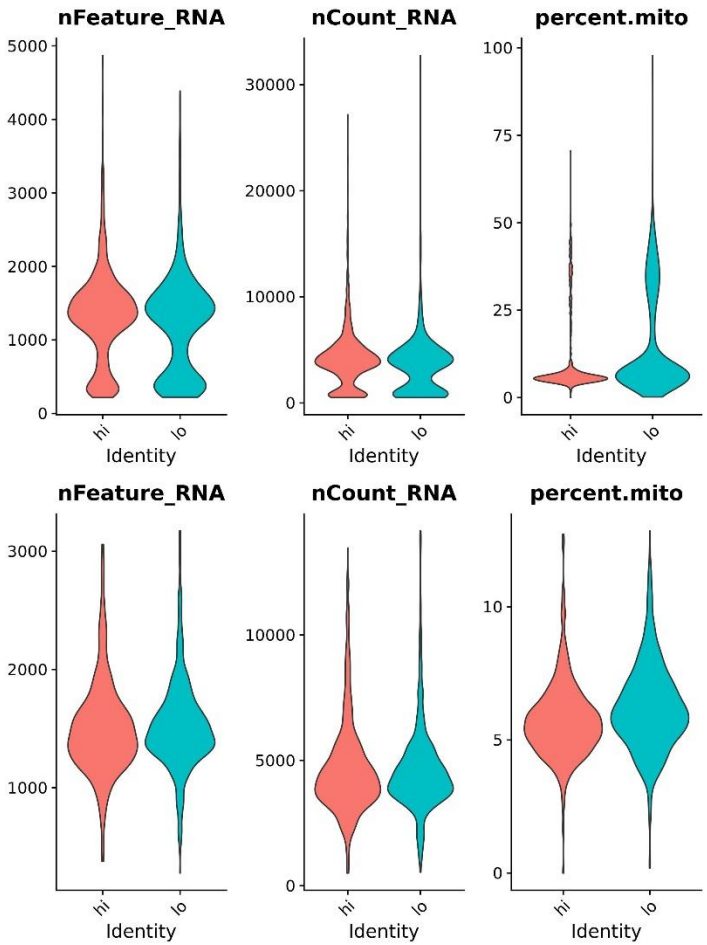
- High # of total detected genes
- High # of total UMI counts

Summary of filtering

Total # of genes Total # of UMIs % Mitochondrial

2. Doublet

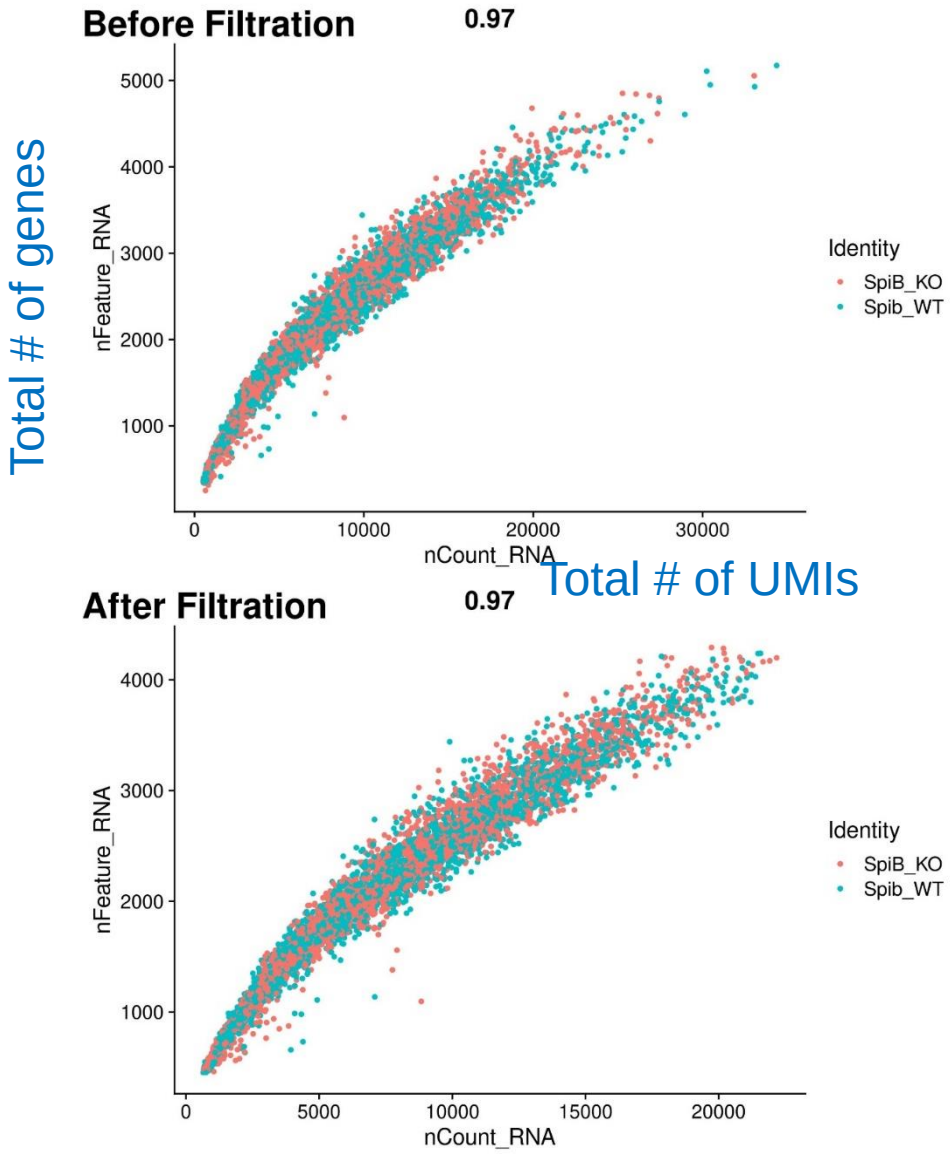
3. Ambient



1. Damaged cell

- High and low expression of total UMI counts or genes (can use upper an lower percentile threshold)
- High percentage of mitochondrial gene expression

UMI counts and Gene counts



- Total number of UMI counts & genes counts are highly correlated
- The counts can vary significantly between cells, spanning more than one order of magnitude

Question

What is required in order to perform comparisons of gene expression between the cells?

Normalization

- Our goal - is to remove the influence of technical effects in the underlying counts, while preserving true biological variation.
- The use of UMIs in scRNA-seq removes technical variation associated with PCR
- Yet, there are many other sources for technical variation:
  - Cell lysis efficiency
  - Reverse transcription efficiency
  - Stochastic molecular sampling during sequencing

Normalization

Seurat normalizes the UMI counts measurements for each cell and gene, by:

- Dividing by the total counts for that cell
- Multiply by a scaling factor (10,000 by default)
- Log transformation

Alternatively there is new procedure - sctransform

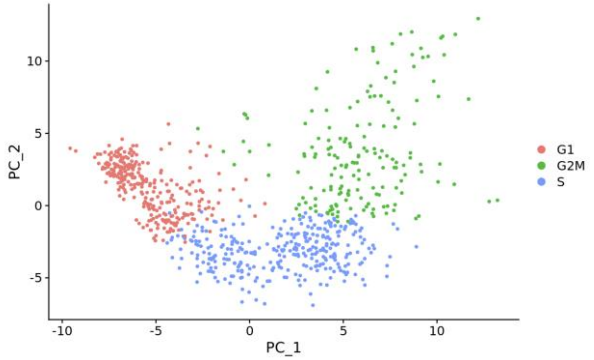
This method constructs a generalized linear model (GLM) for each gene with UMI counts as the response and sequencing depth as the explanatory variable.

Regress-out

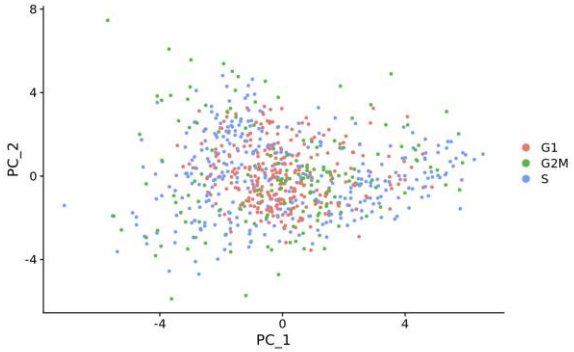
Seurat attempts to subtract or 'regress out' heterogeneity derived from either biological or technical sources, such as:

- Cell cycle scores
  - We assign each cell a score, based on its expression of G2/M and S phase gene markers
- Mitochondrial gene expression (also an indication of cell stress)
- Remark- we need to consider our goals when selecting the sources to regress-out

Before – PCA only cell cycle genes



After regression – PCA only cell cycle genes

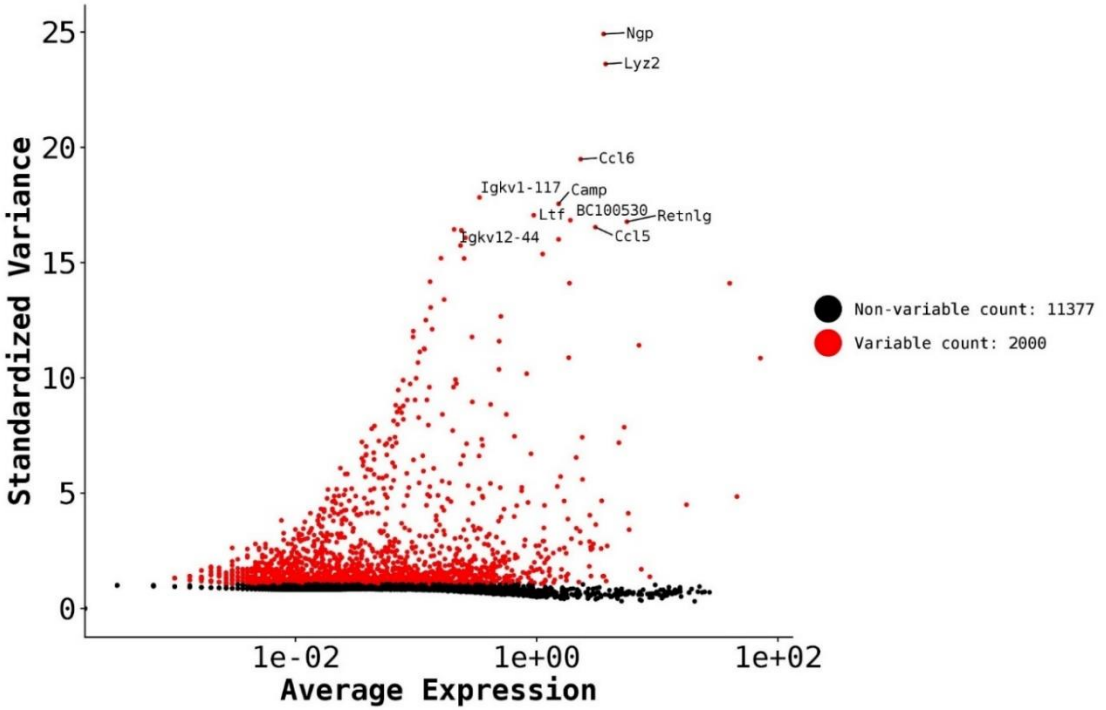


Scaling the Data

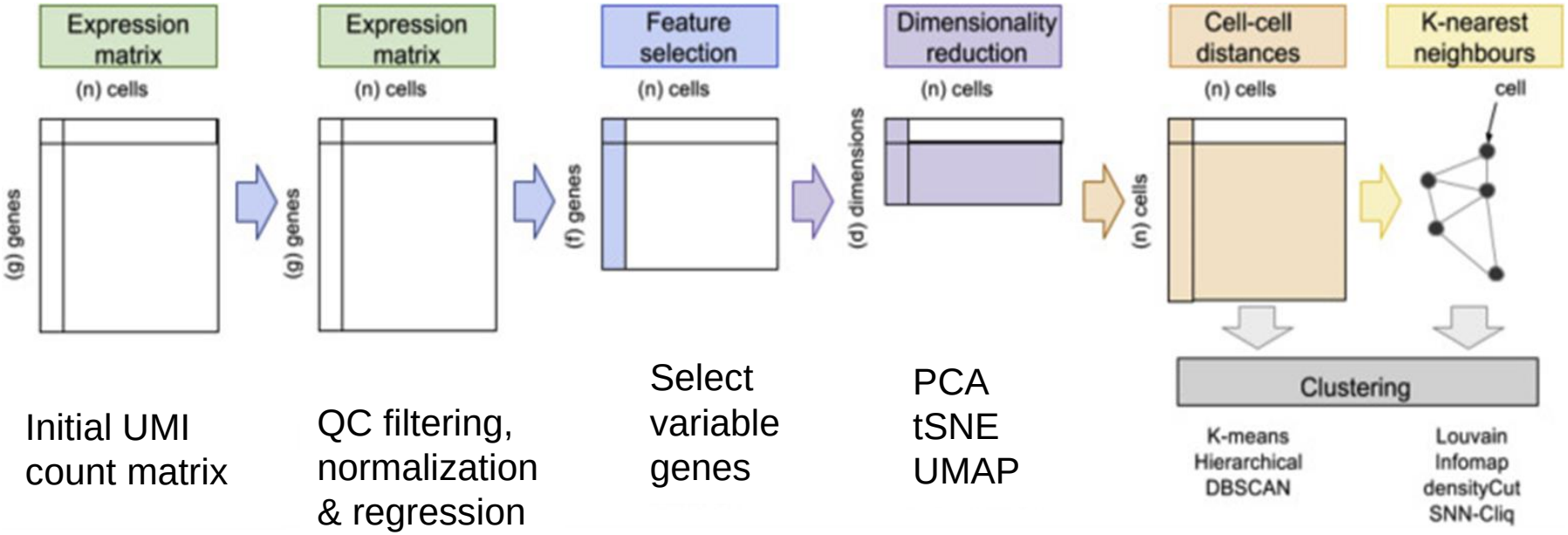
- We scale the data by linear transformation, i.e. shift the expression of each gene, so that:
  - Mean expression across cells is 0
  - Variance across cells is 1
- By preforming scaling we prevent highly expressed genes from dominating the downstream analysis (highly expressed genes might also have the highest variability)

Selection of Genes with High Variability

- Not all the genes are used for downstream analysis
- We calculate a subset of genes that exhibit high cell-to-cell variation
- Selection is done per bin of gene expression
- We select around 3000 genes



Analysis Workflow



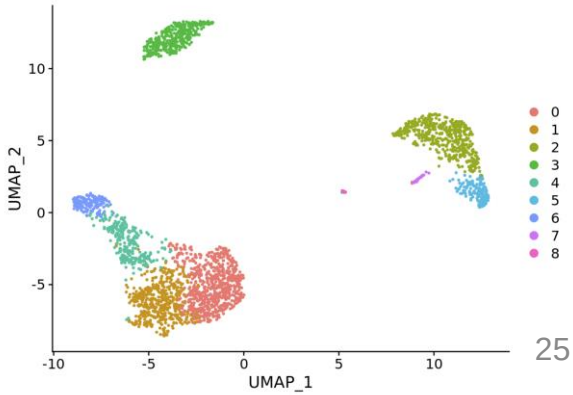
Initial UMI count matrix

QC filtering, normalization & regression

Select variable genes

PCA
tSNE
UMAP

Modified plot from-
Andrews et al. Molecular Aspects of Medicine
Volume 59, February 2018, Pages 114-122



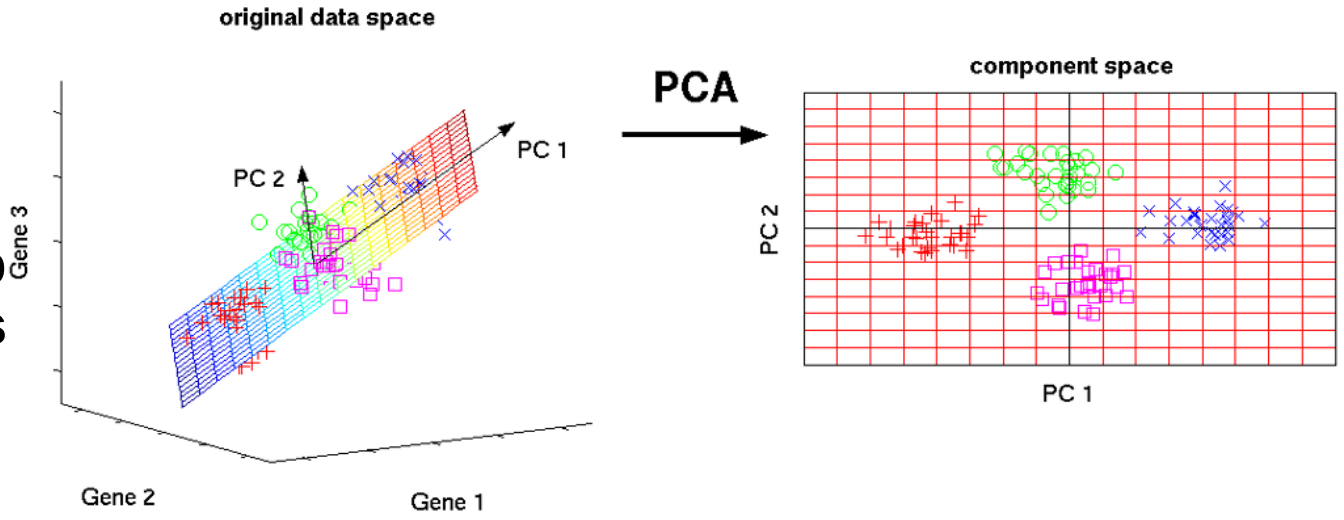
PCA – Principle Component Analysis

Principle Component Analysis (PCA) is a multivariate technique for analyzing quantitative data.

The goal of PCA is to reduce dimensionality, noise, and extract important information (features / attributes) from large amount of data.

We perform PCA on the scaled data

In PCA analysis we find a linear projection of high dimensional data so that **the variance is maximized**



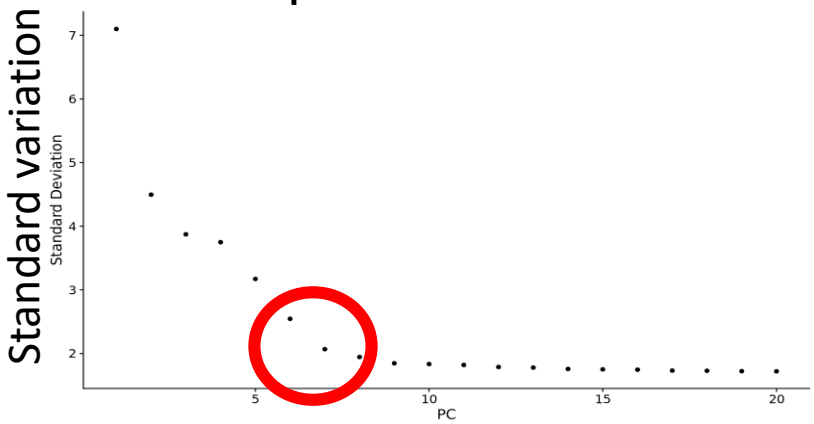
http://www.nlpca.org/pca_principal_component_analysis.html
https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigen

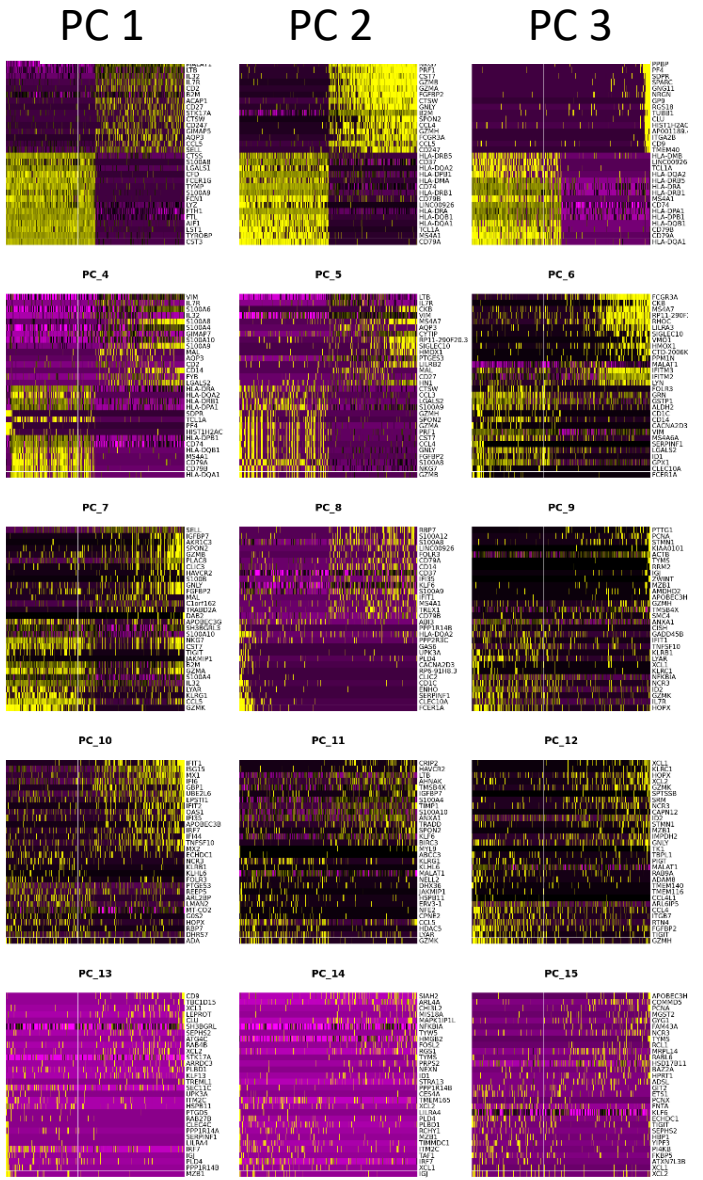
How Many PCs to Choose?

PC heatmaps

- Genes (rows)
- Cells (columns)
- Both cells and features are ordered according to their PCA scores
- Only the 'extreme' cells on both ends of the spectrum are plotted
- In this example we can consider using 7 PCs

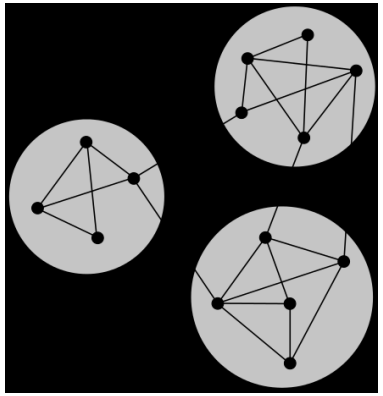
Elbow plot





Clustering

https://en.wikipedia.org/wiki/Community_structure



- Grouping cells into groups in the PC space
- Very briefly - Seurat uses a graph-based clustering approach, which embeds cells in a graph structure, using a K-nearest neighbor (KNN) graph, with edges drawn between cells with similar gene expression patterns. Then, we partition this graph into highly interconnected ‘communities’.

Clustering

We can control the number of clusters by two parameters:

- The number PCs
  - We select the top PCs (since the PCs are sorted by the amount of variance they explain)
  - Selecting too many PCs can introduce noise
- Resolution
  - a parameter which sets the 'granularity' of the clusters. Increased values leading to a greater number of clusters.
- We do not know a priori what parameters to select

Question

Given a dataset is there a true - definite number of clusters?

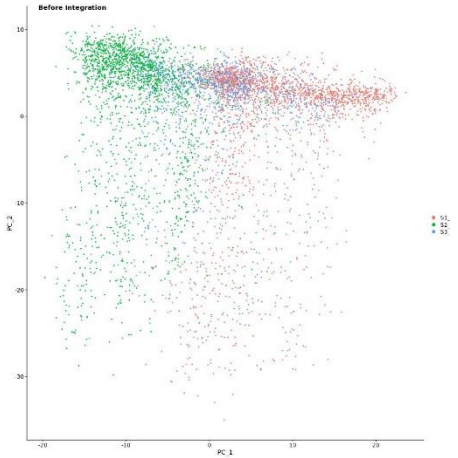
NO

Data Visualization

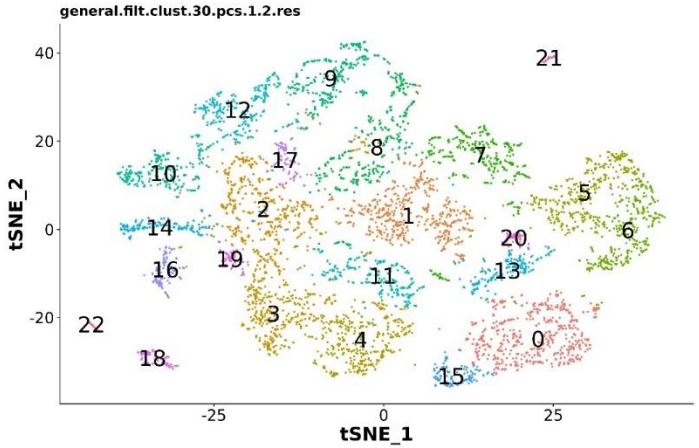
In order to view the data we further reduce the PCs space to 2-3 dimensions

- tSNE preserves local structure in the data
- UMAP preserves both local and global structure

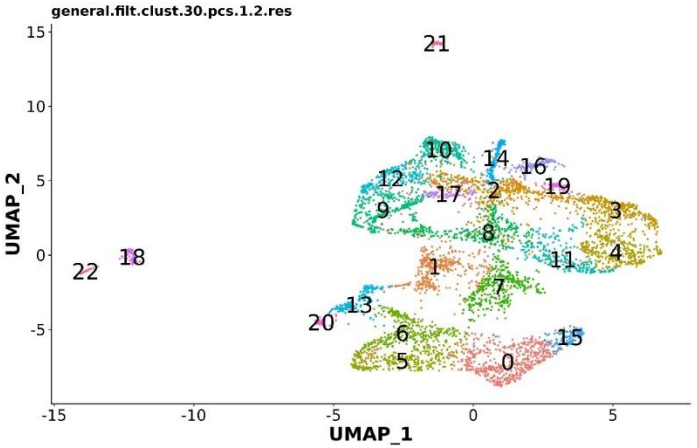
PCA (2 PCs)



tSNE (30 PCs)



UMAP (30 PCs)



Look at the location of cluster 16 in respect to 18 & 22

Finding Differentially Expressed Genes (DEGs)

Find cluster marker genes:

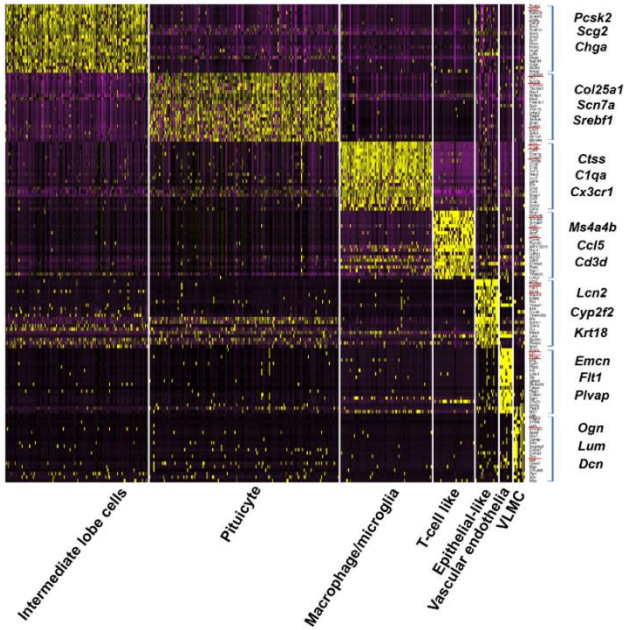
- Identify positive (and negative) DEGs of a single cluster compared to all other cells (Wilcoxon Rank Sum - default)

Table

gene	cluster	avg_logFC	p_val	p_val_adj	pct.1	pct.2
Wfdc17	0	3.985889	0	0	0.975	0.091
Ifitm1	0	3.363648	0	0	0.971	0.074
Lrg1	0	2.987844	0	0	0.984	0.08
Igfbp4	1	1.274831	8.86E-98	1.19E-93	0.661	0.166
Cd8b1	2	2.439779	0	0	0.931	0.029
Cd8a	2	1.700122	0	0	0.672	0.012
Ctsw	2	1.04197	2.42E-154	3.24E-150	0.64	0.082
Igfbp4	3	1.263992	7.48E-81	1.00E-76	0.621	0.173
H2-Eb1	4	1.604532	2.75E-162	3.68E-158	1	0.207
H2-DMb2	4	1.494226	4.19E-162	5.60E-158	0.979	0.191
H2-Aa	4	1.688614	2.70E-159	3.62E-155	1	0.229

pct.1 = fraction of cells in the cluster that express the gene
pct.2 = fraction of cells in all other cells that express the gene

Heatmap



The Seurat Object

The Seurat object serves as a container for:

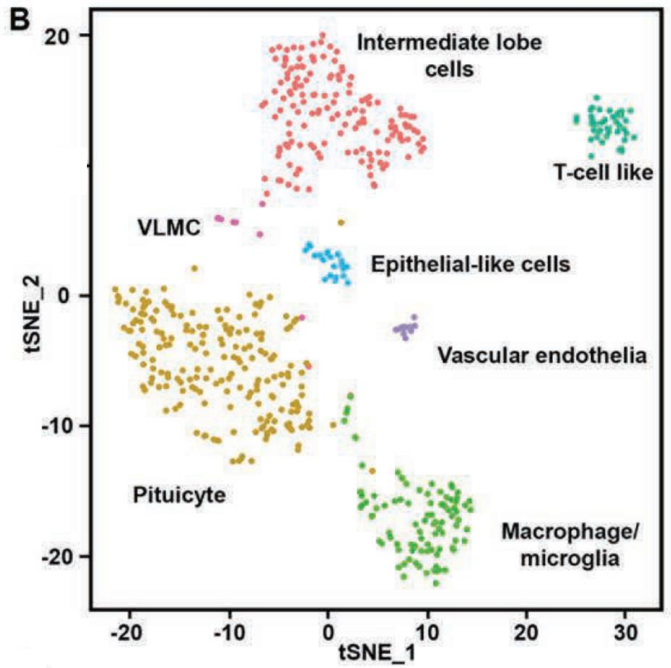
- Data - like the count matrix
- Analysis - like clustering results

seurat_clusters_30PC_1_2...	S4 [15435 x 142] (SeuratObject::S	S4 object of class Seurat
assays	list [2]	List of length 2
RNA	S4 [16474 x 142] (SeuratObject::A	S4 object of class Assay
counts	S4 [16474 x 142] (Matrix::dgCMa	S4 object of class dgCMatrix
data	S4 [16474 x 142] (Matrix::dgCMa	S4 object of class dgCMatrix
scale.data	double [0 x 0]	
key	character [1]	'rna_'
assay.orig	NULL	Pairlist of length 0
var.features	character [0]	
meta.features	list [16474 x 0] (S3: data.frame)	A data.frame with 16474 rows and 0 columns
misc	list [0]	List of length 0
SCT	S4 [15435 x 142] (Seurat::SCTAss	S4 object of class SCTAssay
meta.data	list [142 x 13] (S3: data.frame)	A data.frame with 142 rows and 13 columns
orig.ident	character [142]	'Thy1' 'Thy1' 'Thy1' 'Thy1' 'Thy1' 'CD3_TCRb' ...
nCount_RNA	double [142]	1181 1350 1345 1833 1634 13175 ...
nFeature_RNA	integer [142]	596 497 703 898 735 3101 ...
percent.mito	double [142]	0.000 0.000 0.223 0.164 0.367 2.239 ...
nCount_SCT	double [142]	6064 6364 5772 6124 6055 8338 ...
nFeature_SCT	integer [142]	1221 1077 1260 1260 1183 2994 ...
S.Score	double [142]	0.0456 -0.0608 -0.0923 -0.0617 -0.0468 -0.2075 ...
G2M.Score	double [142]	0.0525 -0.1426 -0.1532 -0.1733 -0.1692 -0.3946 ...
Phase	character [142]	'G2M' 'G1' 'G1' 'G1' 'G1' 'G1' ...
old.ident	factor	Factor with 4 levels: "CD3_TCRb", "HD", "LD", "Thy1"
SCT_snn_res.1.2	factor	Factor with 26 levels: "0", "1", "2", "3", "4", "5", ...
seurat_clusters	factor	Factor with 26 levels: "0", "1", "2", "3", "4", "5", ...

Cluster Annotation

- Clusters can be annotated by enrichment tests comparing cluster marker genes to marker genes from an annotated reference database (hypergeometric test)
- Source for reference - Panglaodb https://panglaodb.se/, contains a comprehensive collection of single cell data

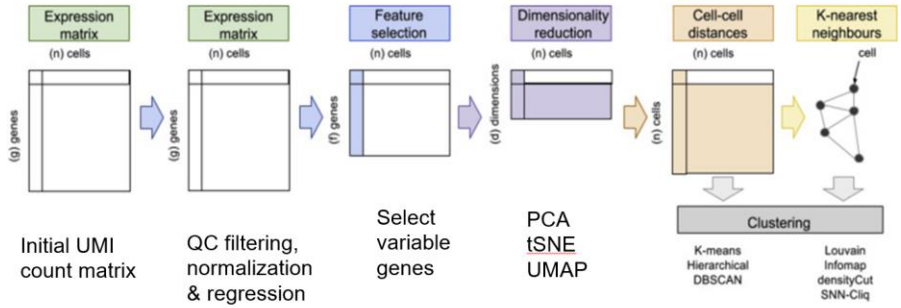
	Mus musculus	*Homo sapiens*
Samples	1063	305
Tissues	184	74
Cells	4,459,768	1,126,580
Clusters	8,651	1,748



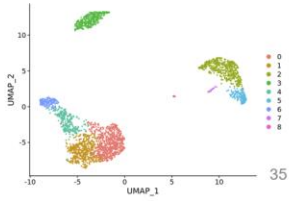
Chen et al. eNeuro 8 January 2020

Lecture Outline

- scRNA analysis: From count matrix to biological knowledge



Modified plot from-
Andrews et al. Molecular Aspects of Medicine
Volume 59, February 2018, Pages 114-122



- Multiomics analysis
  - CITE-Seq
  - Multiome – study example

Single Cell Multiomics Analysis

nature methods

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature methods › editorials › article

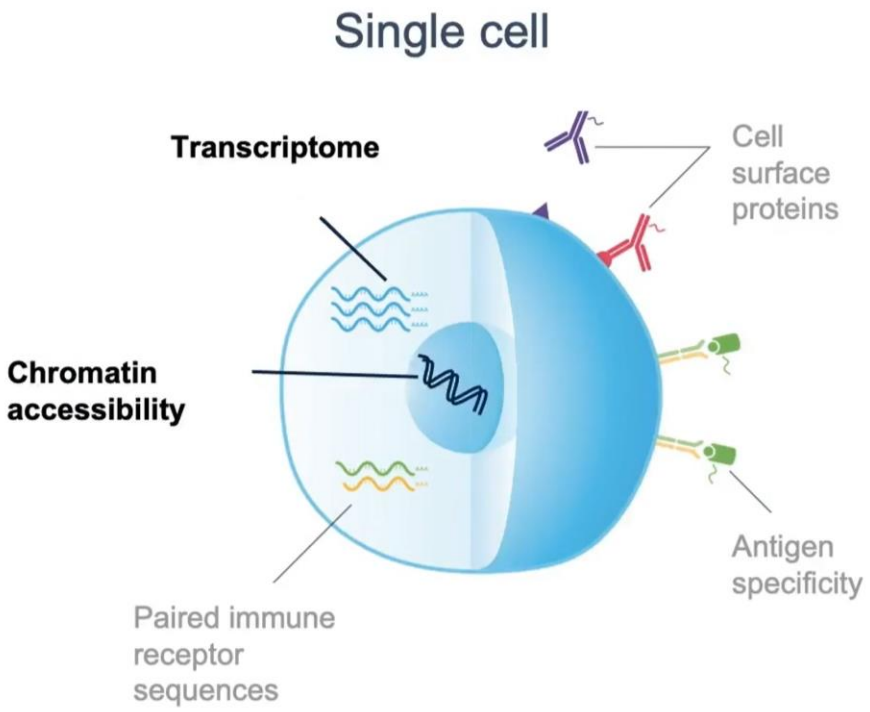
Editorial | Published: 06 January 2020

Method of the Year 2019: Single-cell multimodal omics

Nature Methods **17**, 1 (2020) | Cite this article

34k Accesses | **30** Citations | **129** Altmetric | Metrics

What is Single Cell Multiomics?

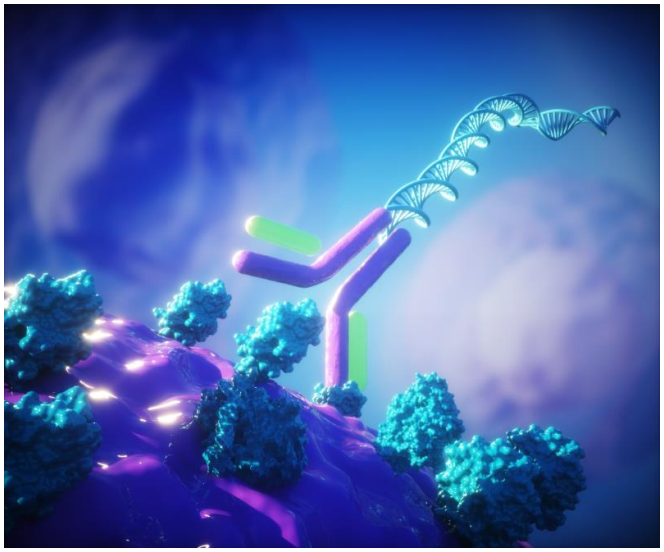


Simultaneously assay multiple data types from the same cells.

For instance we can simultaneous profile the transcriptome and chromatin accessibility.

10X GENOMICS

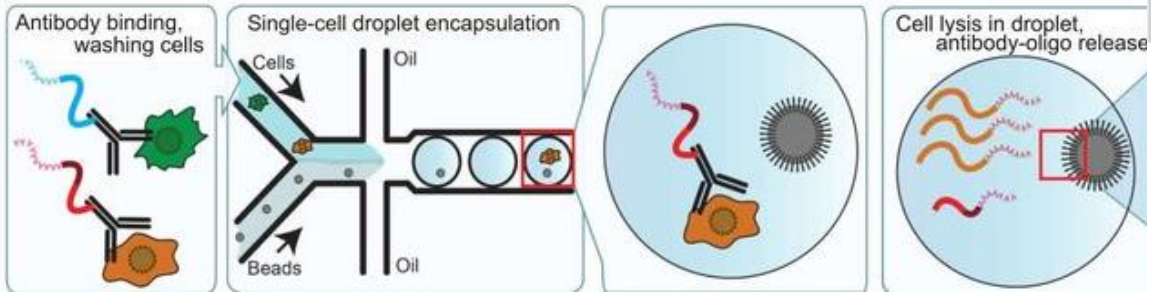
Multi-omics: CITE-seq



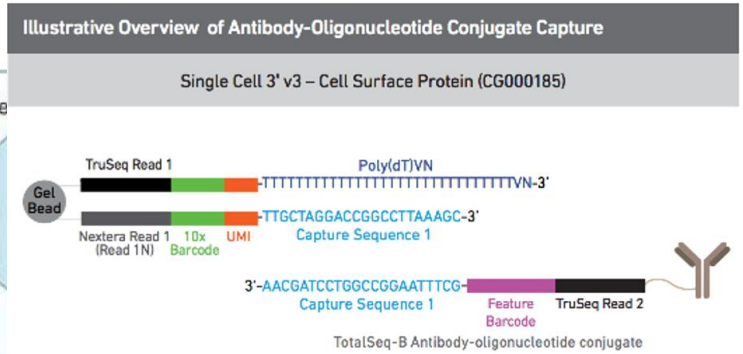
https://www.nature.com/articles/d42473-020-00052-9

Cellular Indexing of Transcriptomes and Epitopes by sequencing

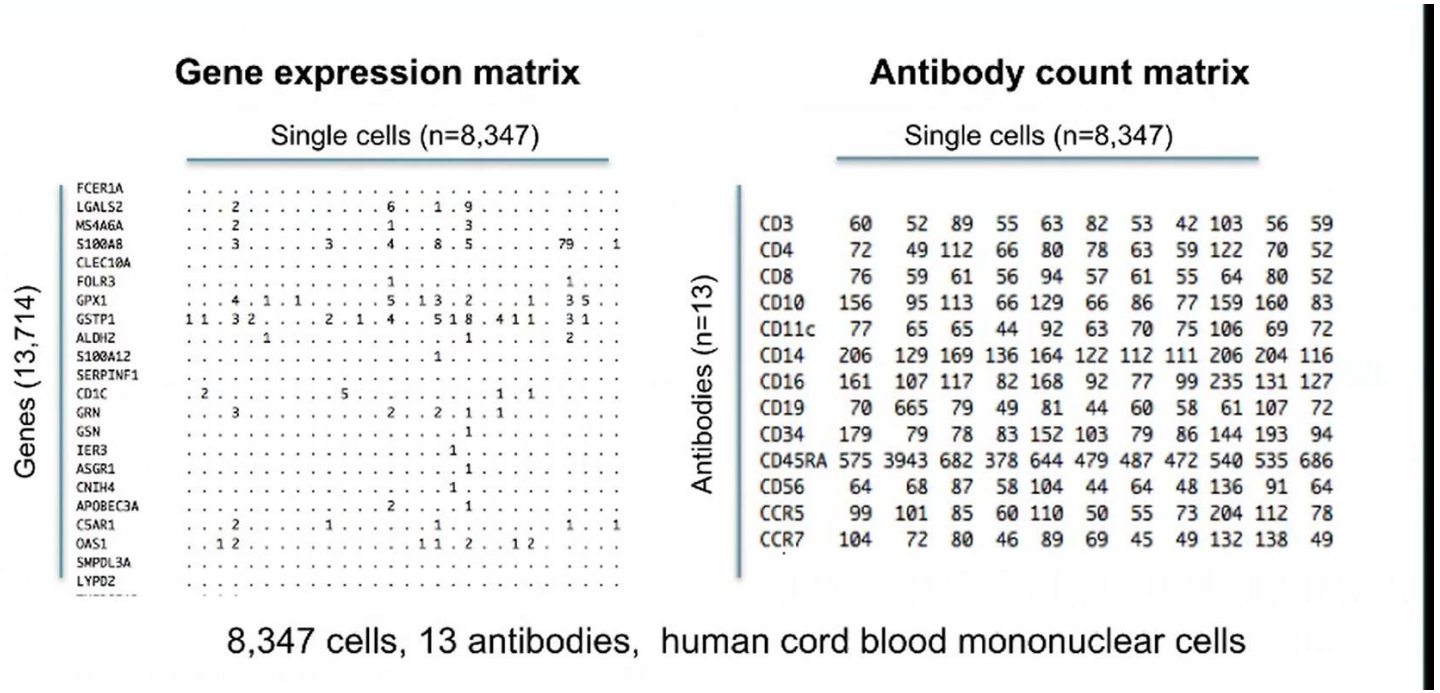
CITE-seq enables simultaneous detection of single-cell transcriptomes and protein markers



Modified DOI: 10.1126/sciadv.aax8978



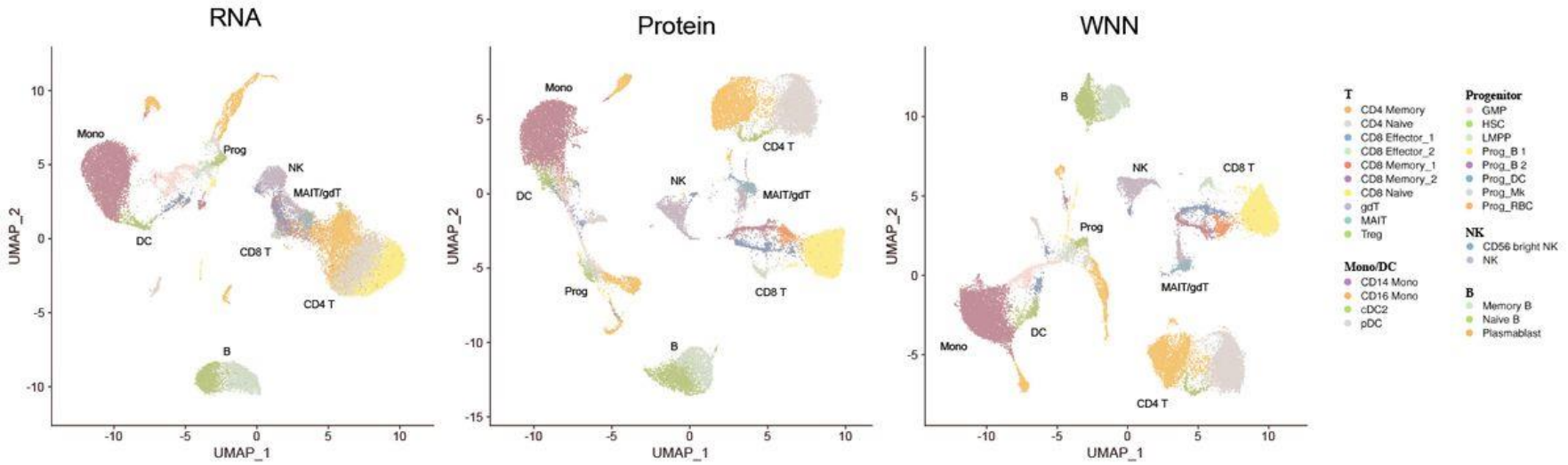
Count matrix of CITE-Seq



Taken from a presentation of Rahul Satija

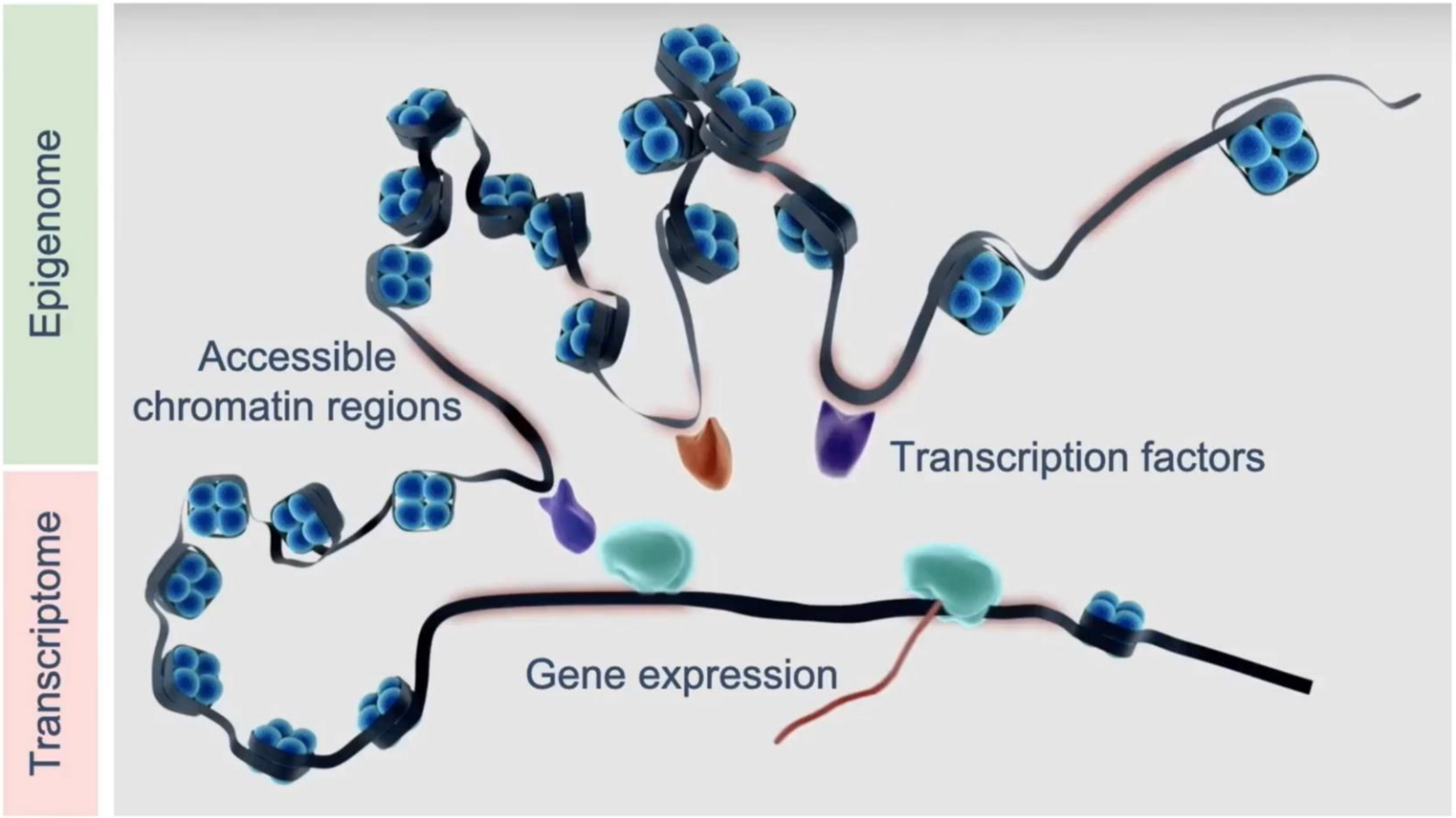
Integrated Analysis of Multimodal Single-Cell Data

- Hai et al. Satija Introduce ‘weighted-nearest neighbor’ (WNN) analysis
- They demonstrate that WNN analysis substantially improves the ability to define cellular states



https://doi.org/10.1016/j.cell.2021.04.048

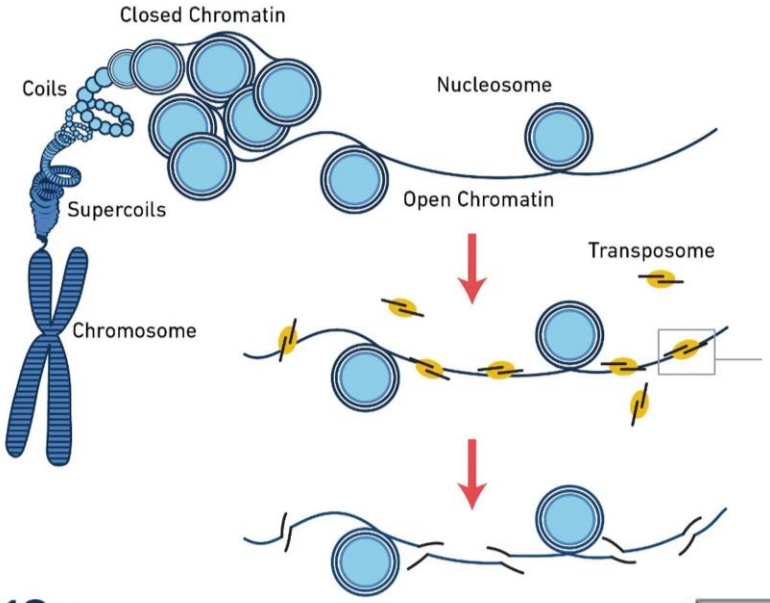
Co-assay Chromatin Accessibility & Transcriptome in Single Cells



10x GENOMICS

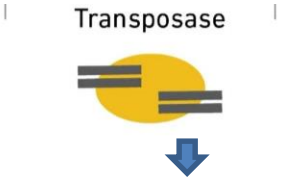
Chromatin Accessibility

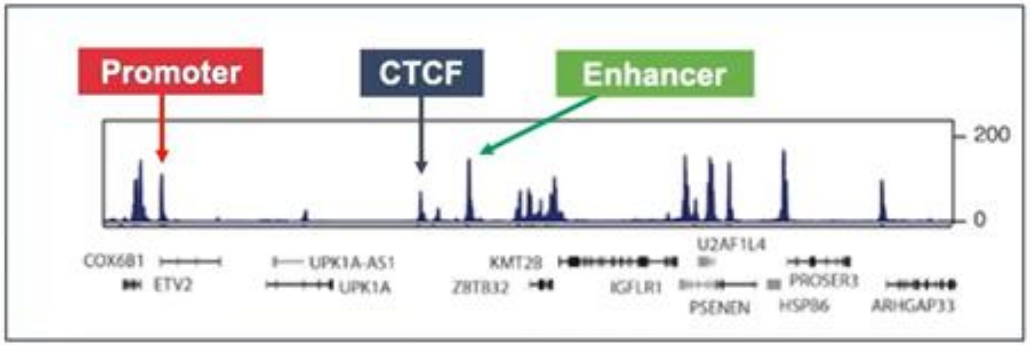
scATAC-Seq : Single Cell Assay for Transposase-Accessible Chromatin using sequencing



10X GENOMICS

- Nuclei are incubated with Tn5 transposase and sequencing adapters
- The transposase cleaves open regions in the DNA and inserts the sequencing adapters.





Open chromatin represented as "peaks"

Multiomics Analysis Workflow

scRNA-Seq

Cell Ranger (10x)

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

Seurat and Signac (Satija Lab)

Clustering

Integration

scATAC-Seq

	Cell1	Cell2	...	CellN
Peak1	3	2	.	13
Peak2	2	3	.	1
Peak3	1	14	.	18
	.	.	.	.
	.	.	.	.
	.	.	.	.
PeakM	25	0	.	0

Clustering

Study Example (ongoing work):
Single Cell Multiome ATAC & Gene Expression
Thymic Epithelial Cells (TEC) Heterogeneity

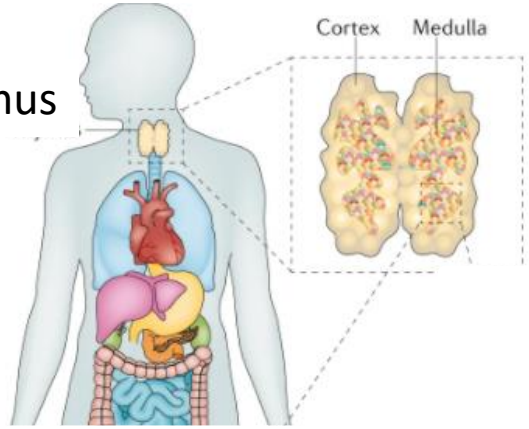
Dr. Hadas Keren-Shaul

Dr. Merav Kedmi

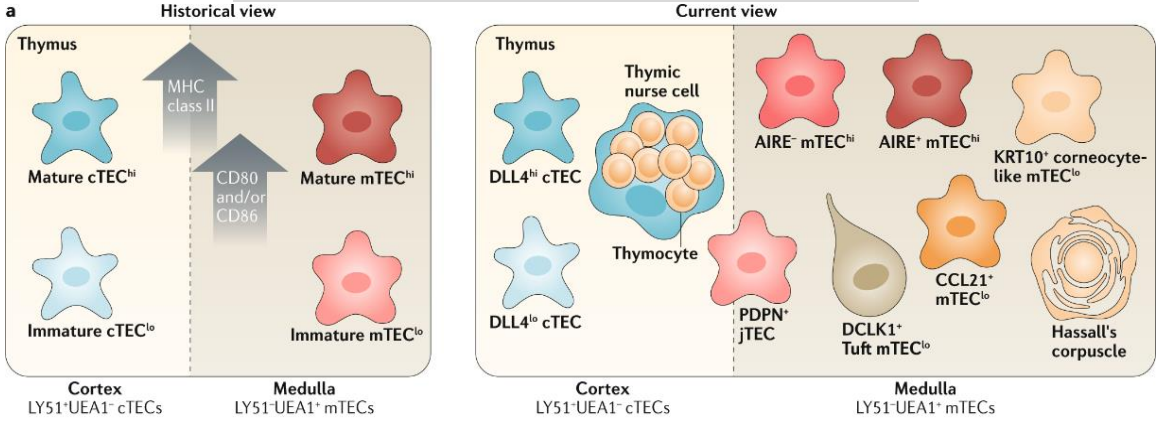
Genomics, Sandbox

Dr. Jakub Abramson

Dr. Yael Goldfarb



Thymic epithelial heterogeneity



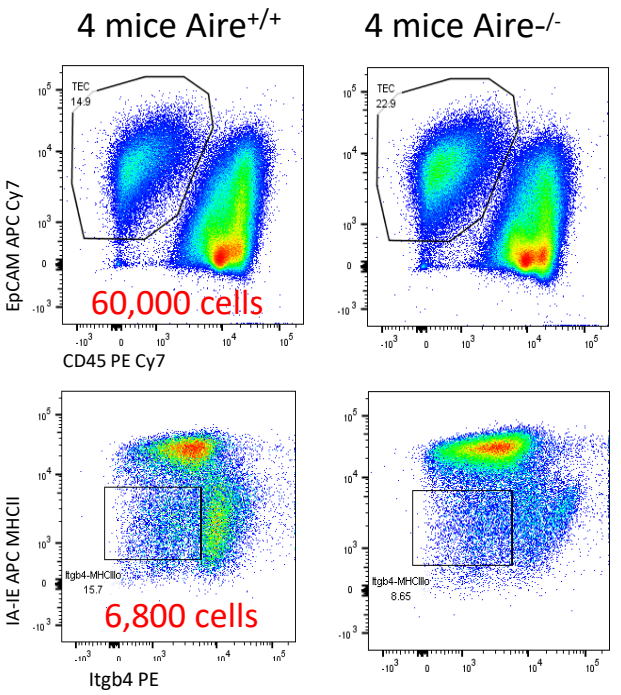
Nature Reviews Immunology volume 20, pages239–253 (2020)

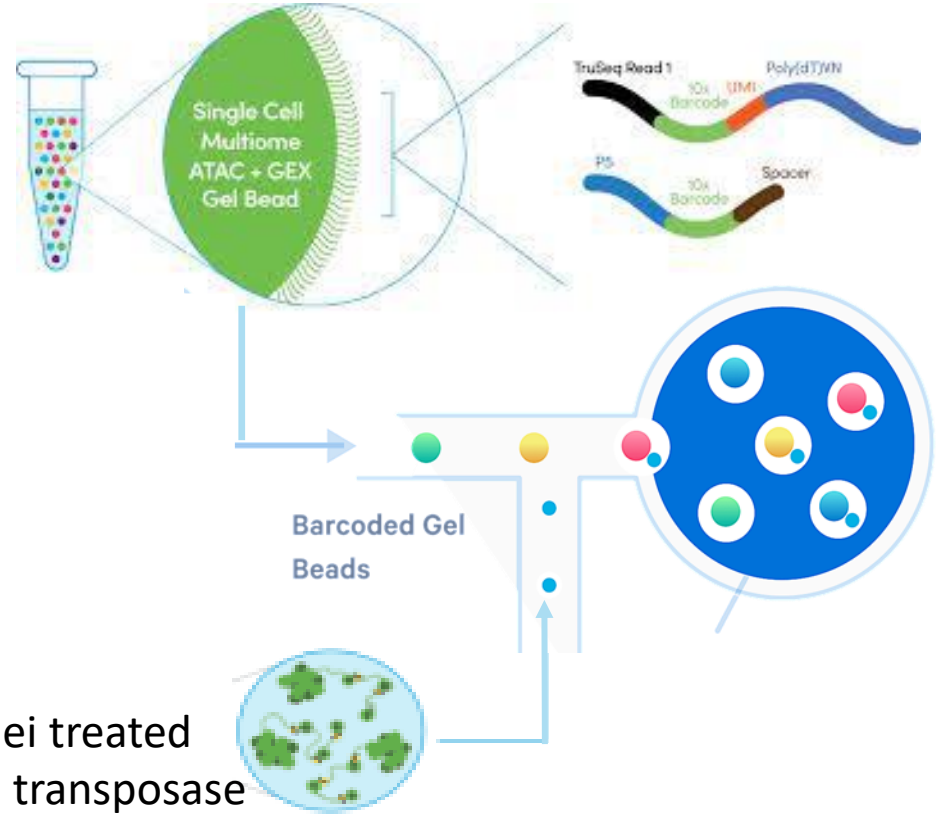
T cell education is done mainly by the Thymic Epithelial Cells

- Cortex - initial stages of development
- Medulla - express and present tissue-restricted antigens (TRA) enabling elimination of T cells with self-reactivity

Experimental Details

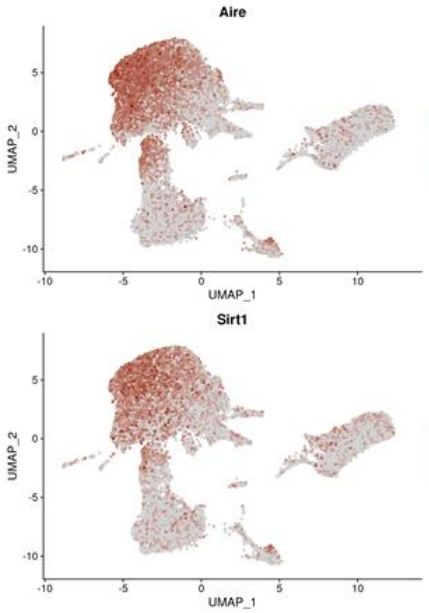
- FACS sorting thymic cells to enrich for rare epithelial cell types
- Aire gene is expressed in mTEChi and regulates many genes among them Tissue Restricted Antigens (TRA)

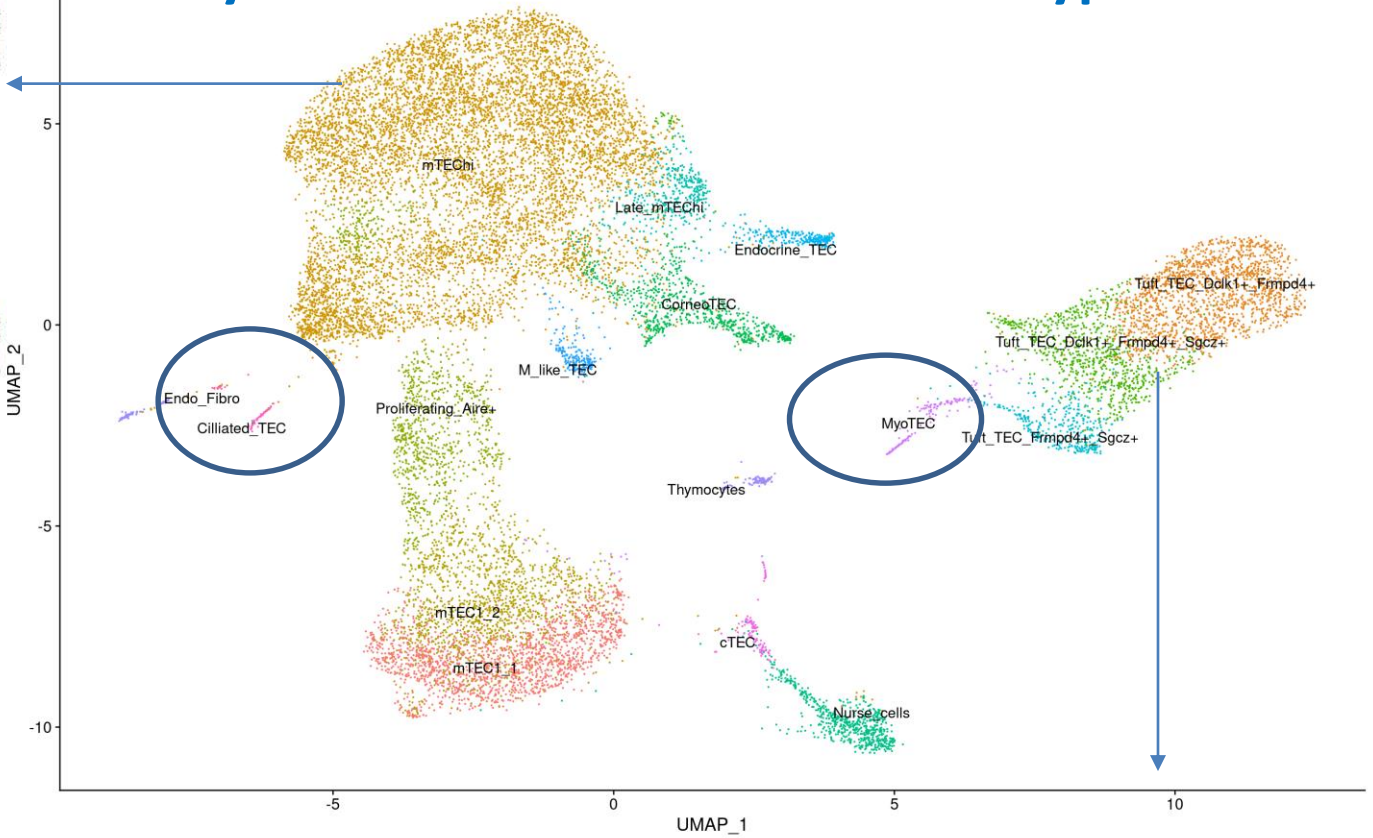


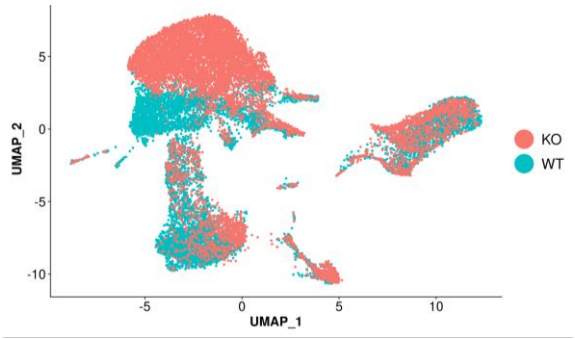


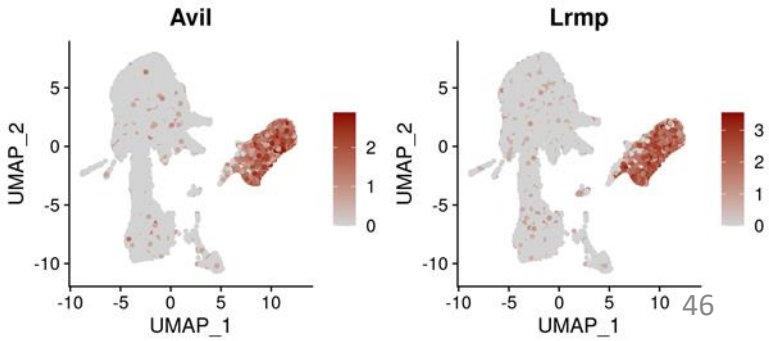
scRNA-Seq Analysis

Identify clusters and annotate to cell types

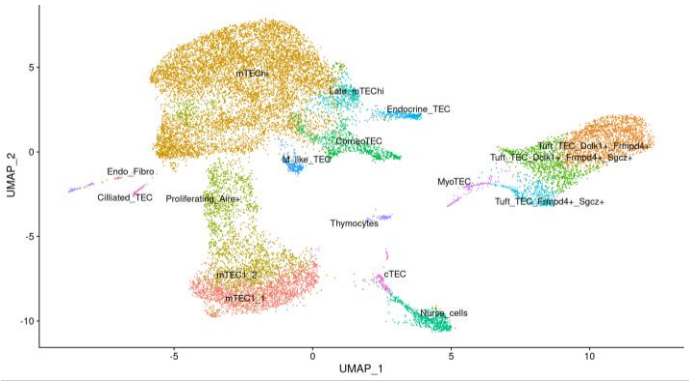




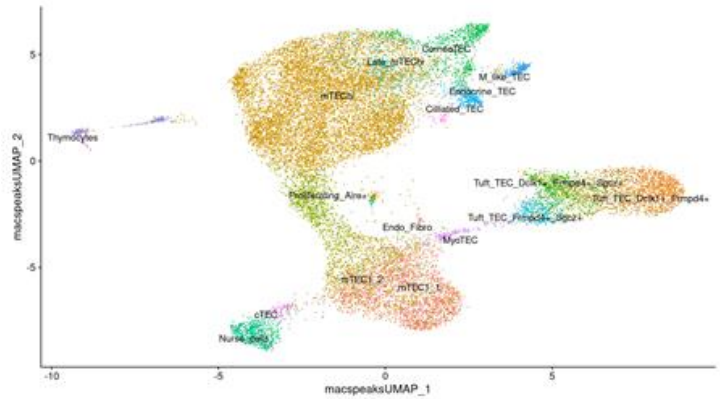




Multiome - Integration

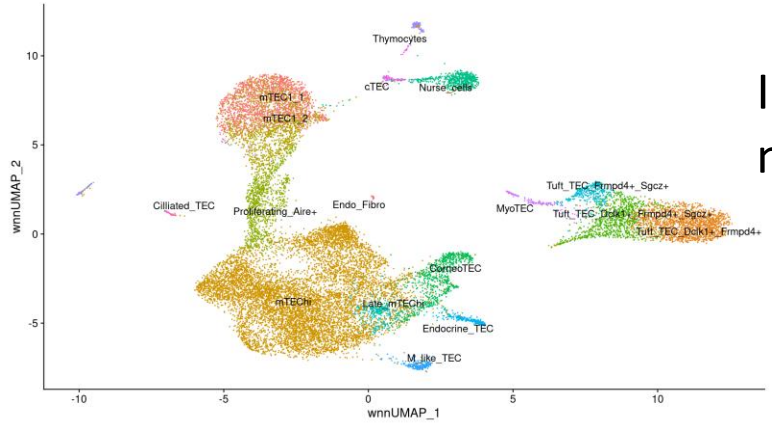


scRNA-Seq



scATAC-Seq

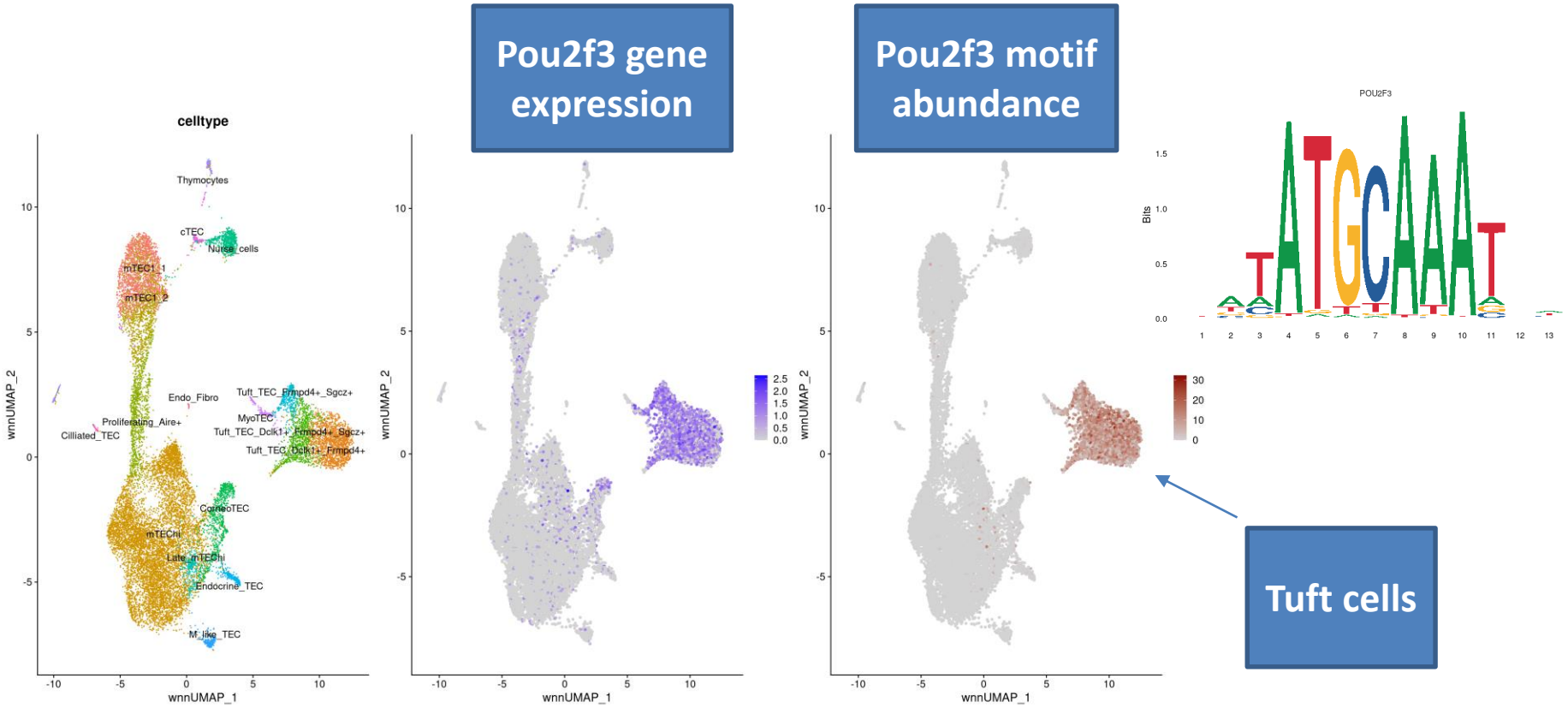
Identification of peaks for each cell type separately (macs2)



Integrated Analysis – Weighted nearest neighbor (WNN)

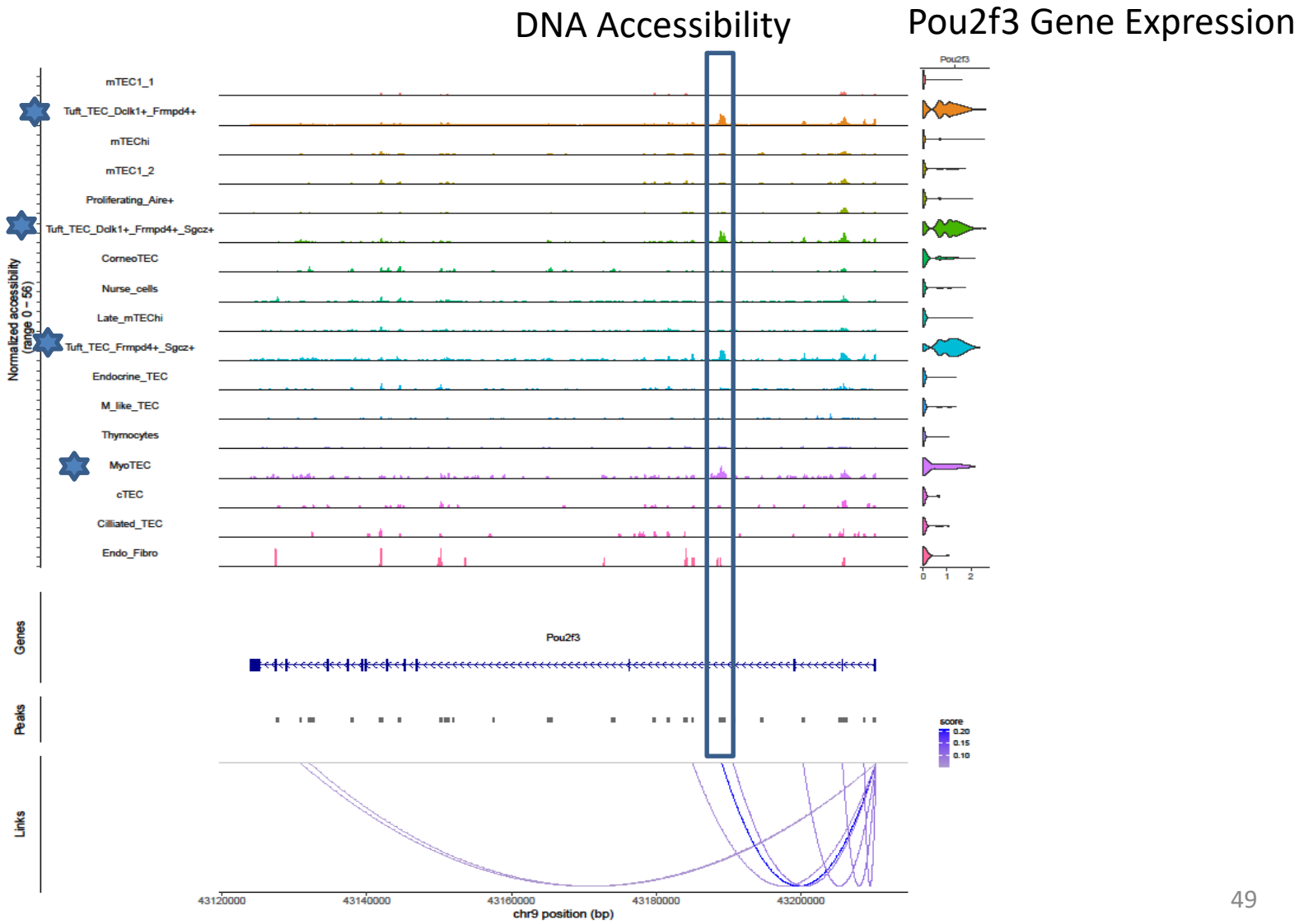
Multiome Advantages

- Identify regulatory mechanisms using both modalities:
  - scRNA : Detect Differentially Expressed Transcription Factors (TFs)
  - scATAC : Overrepresented TF DNA motifs

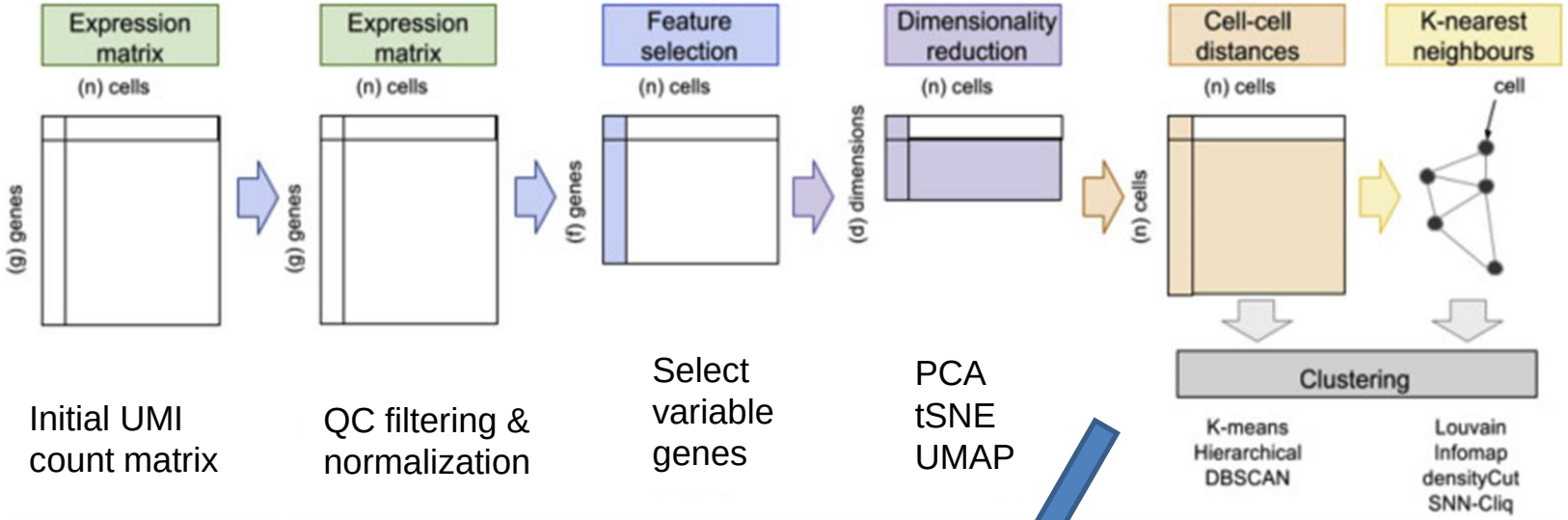


Identify Peaks that Regulate Gene Expression

Linking peaks to genes by computing correlation between gene expression and accessibility of a nearby peak



Summary: Analysis Workflow



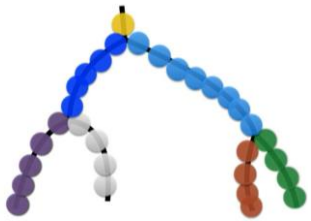
Initial UMI count matrix

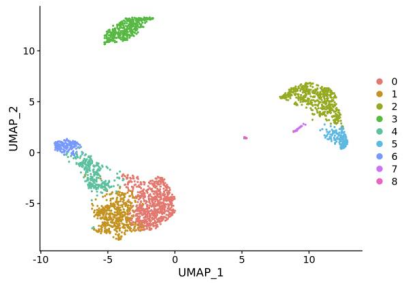
QC filtering & normalization

Select variable genes

PCA
tSNE
UMAP

Trajectory





Modified plot from-

Andrews et al. Molecular Aspects of Medicine
Volume 59, February 2018, Pages 114-122

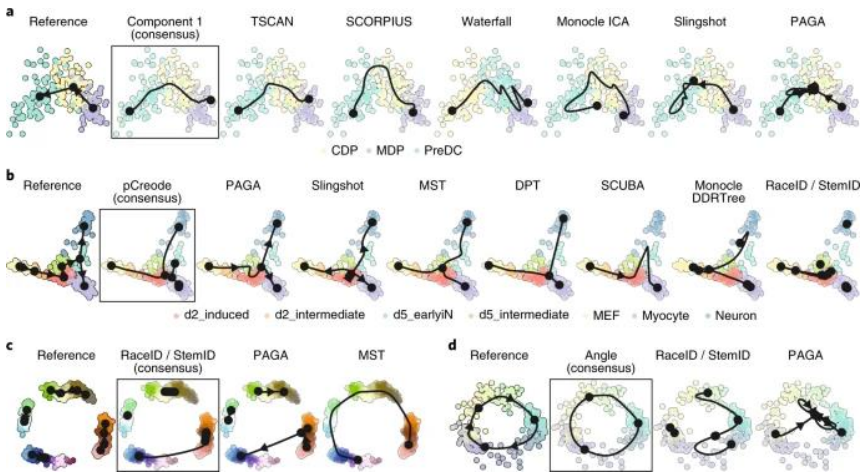
Trajectory Inference Analysis

- Clustering is a discrete classification approach and therefore lacks in the ability to capture:
  - Transitions between cell identities
  - Branching differentiation processes
- Trajectory inference methods interpret single-cell data as a snapshot of a continuous process
- Interpretation of a trajectory requires additional data sources

A comparison of single-cell trajectory inference methods

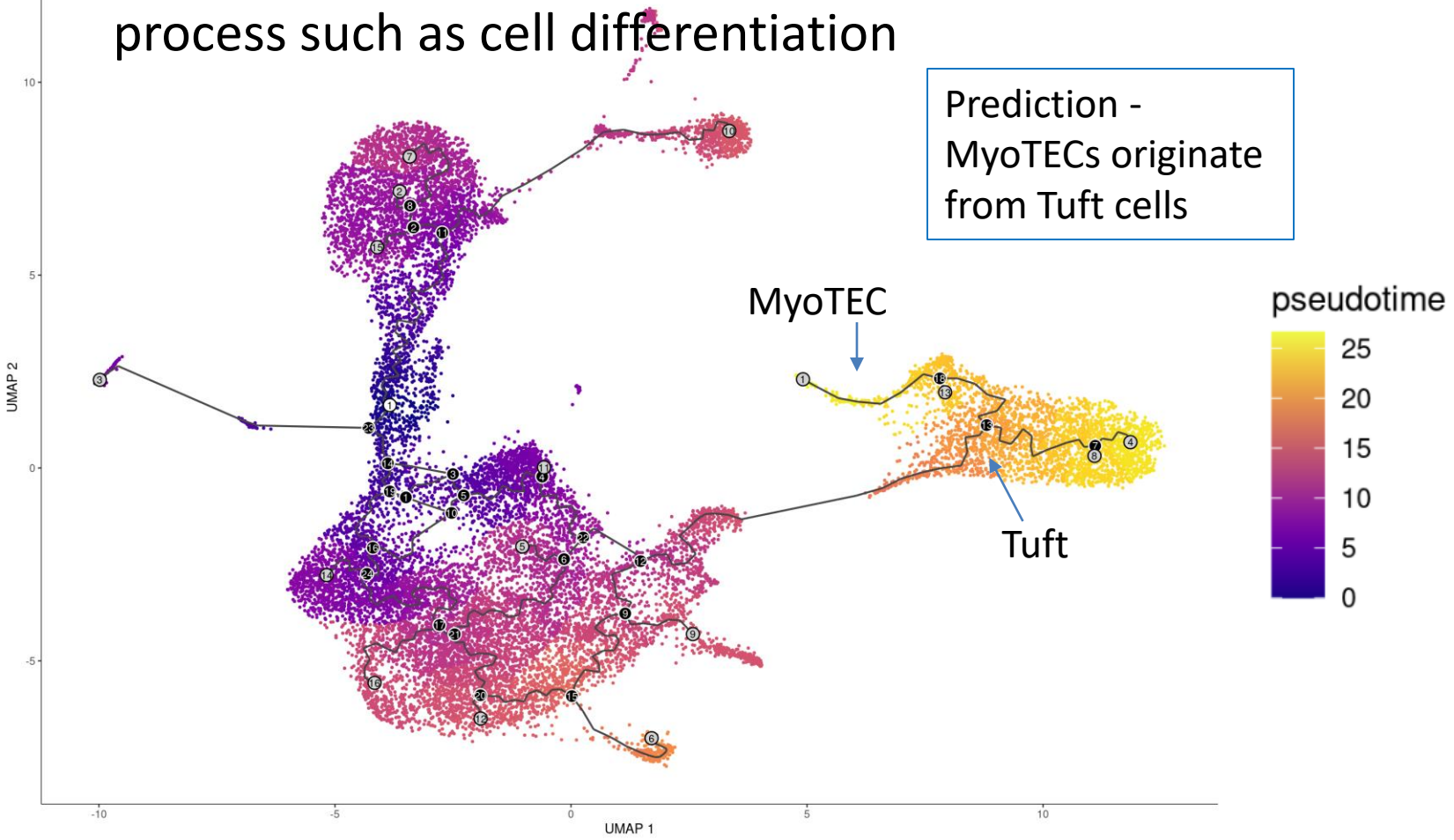
Wouter Saelens, Robrecht Cannoodt, Helena Todorov & Yvan Saeys

Nature Biotechnology **37**, 547–554(2019) | Cite this article



Trajectory Inference Analysis

- Trajectory inference methods interpret single-cell data as a continuous process (monocle)
- Cells are placed along a trajectory corresponding to a biological process such as cell differentiation



References

Current best practices in single-cell RNA-seq analysis: a tutorial, Luecken et al. Mol Syst Biol. (2019) 15: e8746

satijalab.org/publications/

Apps | UTAP - NGS PIPELI... | From samples to an... | Mouse mm10 chr1... | Box | Login | Everything : Panopto | Be Awesome in ggp... | Results | R cheat-sheet | rmarkdown-cheats... | Terra | EndN

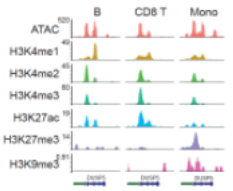
SATIJA LAB

News People Research Publications Seurat Join/Contact Single Cell Genomics Day

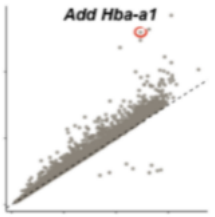
2021

Stuart T, Srivastava A, Madad S, Lareau C, Satija R
Single-cell chromatin state analysis with Signac
Nature Methods. 2021

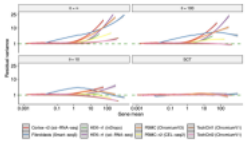
Zhang B*, Srivastava A*, Mimitou E, Stuart T, Raimondi I, Hao Y, Smibert P, Satija R
Characterizing cellular heterogeneity in chromatin state with scCUT&Tag-pro
bioRxiv. 2021



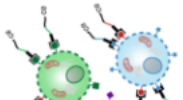
Missarova A, Jain J, Butler A, Ghazanfar S, Stuart T, Brusko M, Wasserfall C, Nick H, Brusko T, Atkinson M, Satija R*, Marioni J*
geneBasis: an iterative approach for unsupervised selection of targeted gene panels from scRNA-seq
bioRxiv. 2021



Choudhary S, Satija R
Comparison and evaluation of statistical error models for scRNA-seq
bioRxiv. 2021



Mimitou E, Lareau C, Chen K, Zorzetto-Fernandes A, Hao Y, Takeshima Y, Luo W, Huang T, Yeung B, Papalexi E, Thakore P, Kibayashi T, Wing J, Hata M, Satija R, Nazor K, Sakaguchi S, Ludwig L, Sankaran V, Regev A, Smibert P
Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells
Nature Biotechnology. 2021



The End
Thanks for listening

Questions?