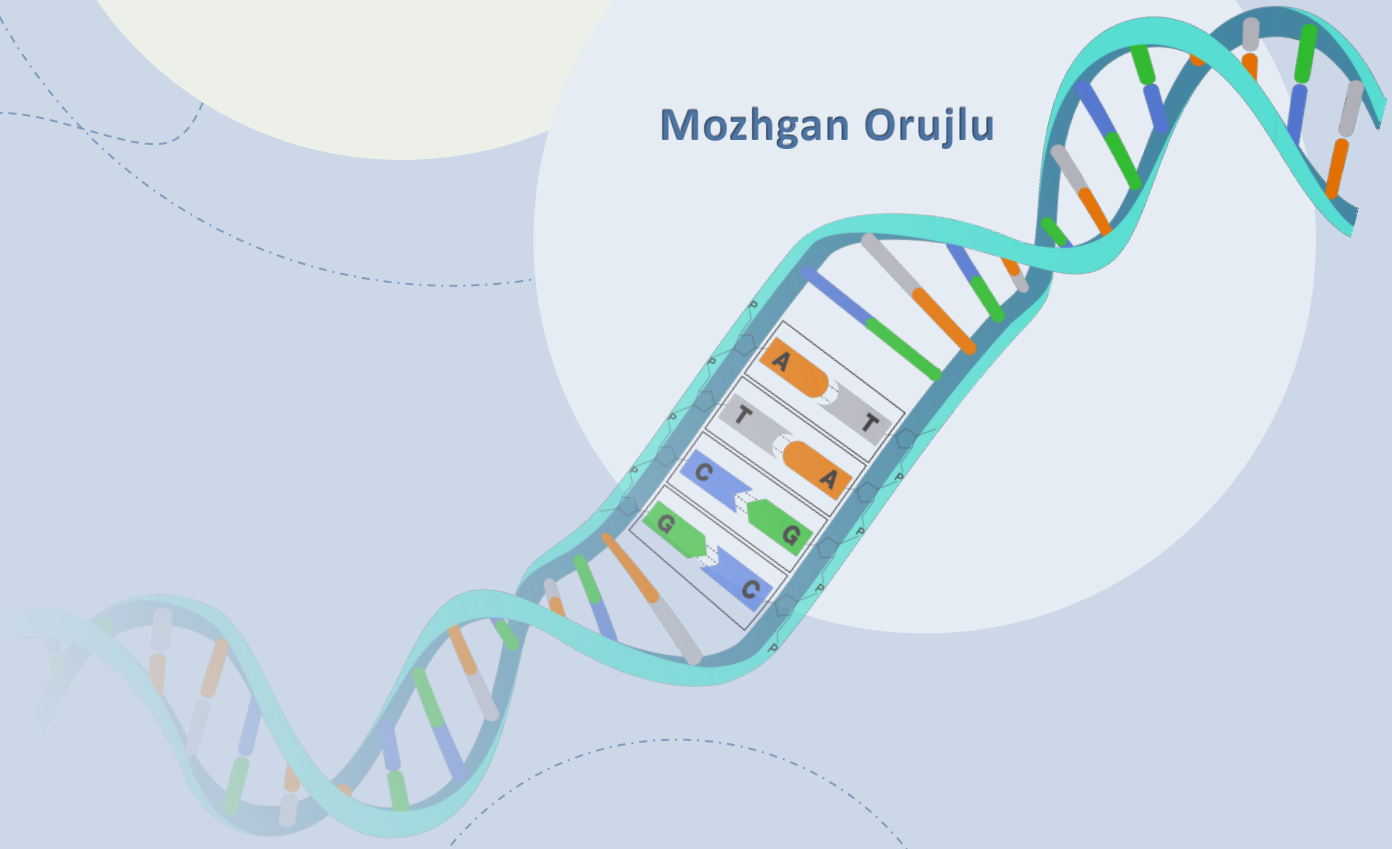
An Introduction to

Single-Cell RNA Sequencing Data

The Human Cell Atlas

Mozhgan Orujlu



The Human Cell Alas

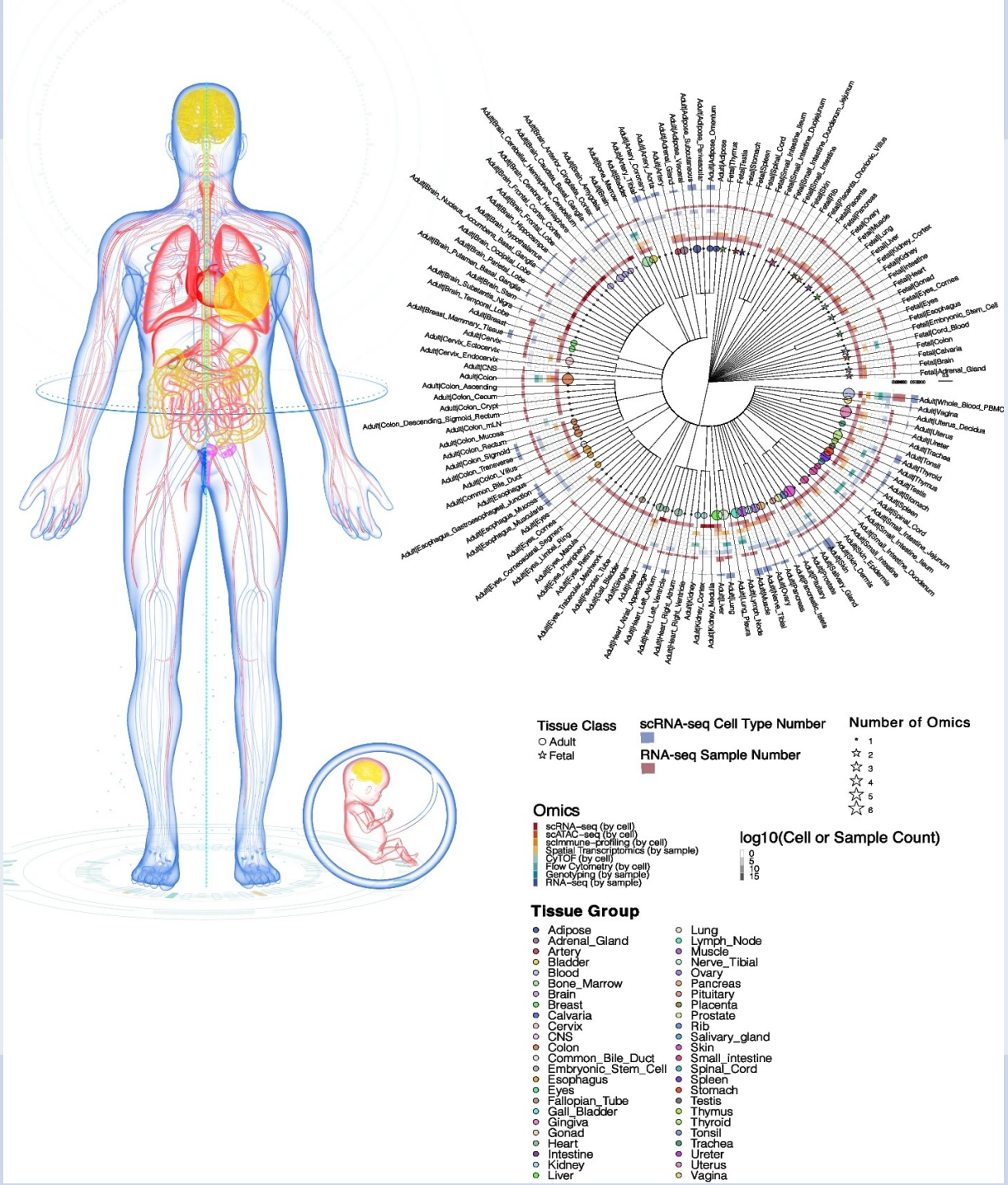


Image from: **Single Cell Atlas: a single-cell multi-omics human cell encyclopedia**

The White Paper

Sarah Amalia Teichmann

Aviv Regev

THE HUMAN CELL ATLAS

White Paper

The HCA Consortium

October 18, 2017

The Human Cell Alas

- The Human Cell Atlas (HCA) will be made up of comprehensive reference maps of all human cells — the fundamental units of life — as a basis for understanding fundamental human biological processes and diagnosing, monitoring, and treating disease.
- It will help scientists understand how **genetic variations** impact disease risk, define drug toxicities, discover better therapies, and advance regenerative medicine.
- As of 2024, the project has mapped approximately 62 million human cells into 18 biological networks, which includes cells from vital systems such as the nervous system, lungs, heart, intestine and immune system.

The Human Cell Alas

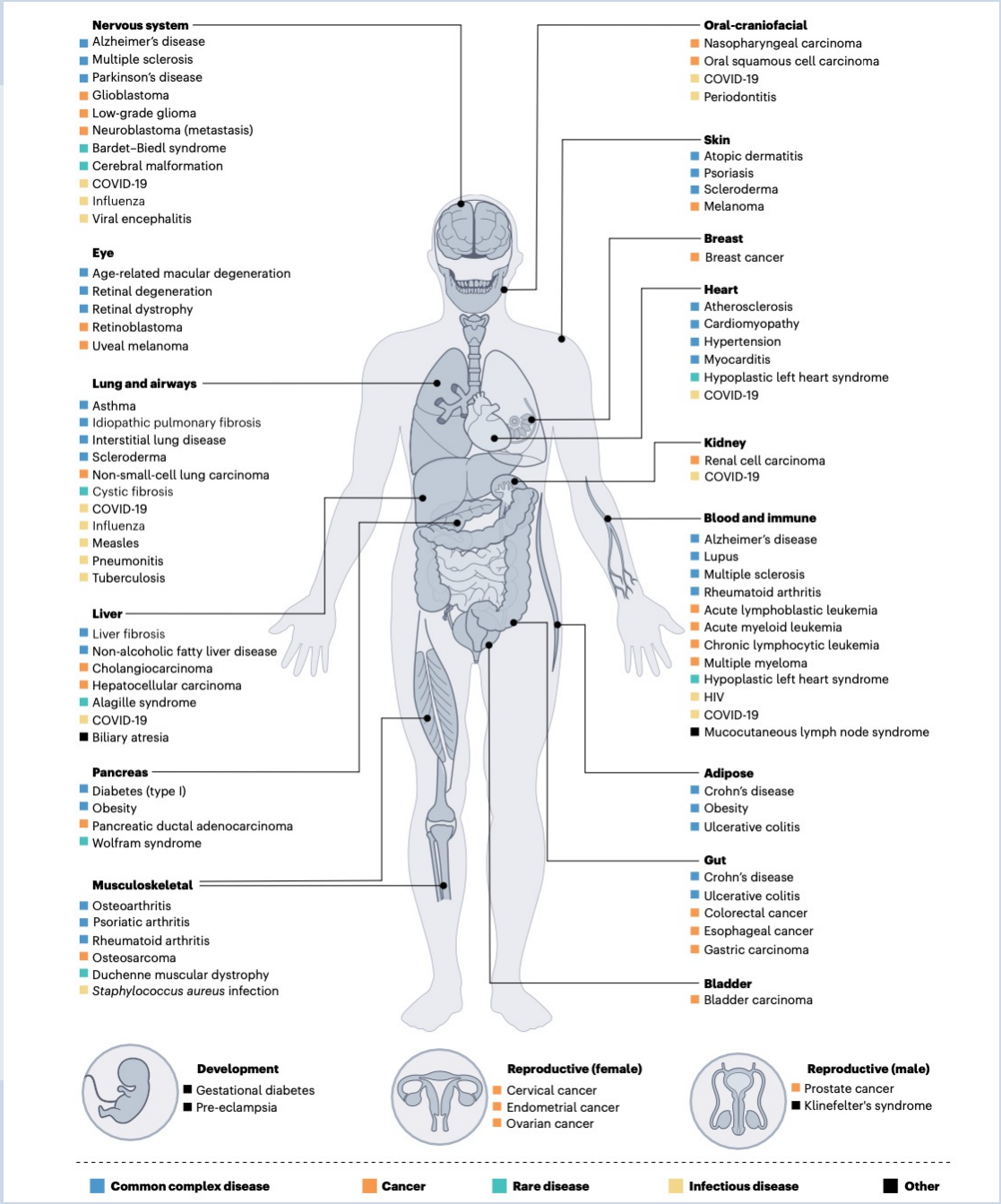


Image from: Impact of the Human Cell Atlas on medicine
Sarah A. Teichmann & Aviv Regev

Key Experimental Methods

- Key experimental methods for construction of cell atlases at different levels of biological organization

- Clinical data
- Tissue imaging and histology

- Spatial
- Multimodal
- **Transcriptomics**
- Genome and epigenomics

Key Experimental Methods

- Key experimental methods for construction of cell atlases at different levels of biological organization

- Clinical data
  - health-related information collected from patients, including medical history, laboratory test results, lifestyle factors, and clinical diagnoses.
- Tissue imaging and histology
  - visualize tissue architecture and cellular organization and provides spatial context by revealing how cells are arranged within tissues by staining tissue sections and examining them under a microscope
  - It helps map cell types and their interactions within tissues
  - Example: Photoacoustic imaging,
    - Atomic force microscopy (AFC)

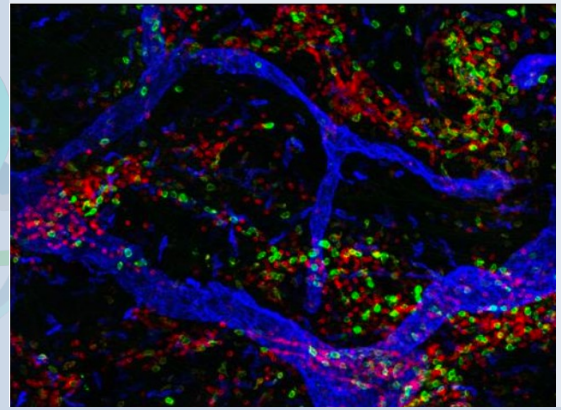


Image from: the white paper

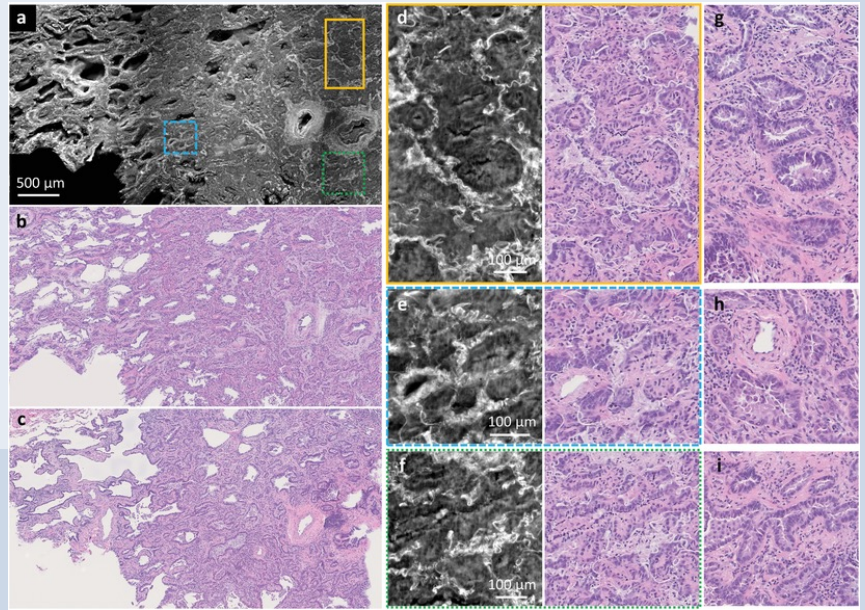


Image from: High-Throughput, Label-Free and Slide-Free Histological Imaging by Computational Microscopy and Unsupervised Learning

Key Experimental Methods

- Spatial
  - Allows researchers to see where specific genes are active within the tissue, providing insights into the cellular environment.
  - Example: FISH , INSTA-seq
- Genome and epigenomics
  - Techniques that study the genetic makeup (genome)

  and chemical modifications to DNA and histones (epigenomics)

  that regulate gene expression
  - Example: ATAC-seq
    - **23 pairs of chromosomes** (46 in total)

      chr1 to chr22 (autosomes)

      chrX, chrY (sex chromosomes).

```
chr1   3456000  3456500  peak1  120
chr2   7890000  7890500  peak2  230
chrX   1234500  1235000  peak3  190
```

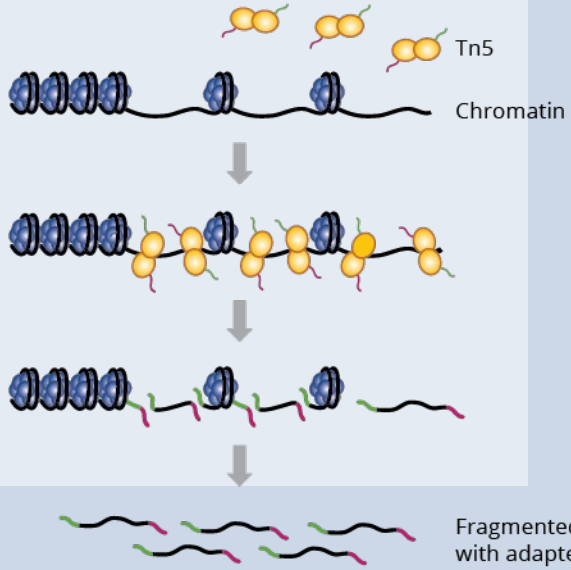
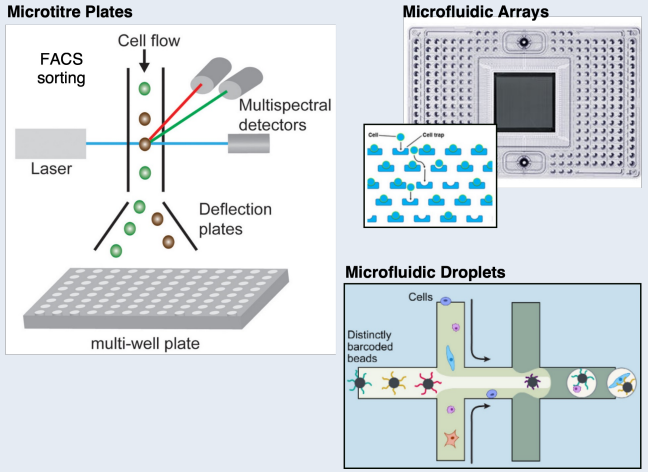


Image from: www.genewiz.com

Key Experimental Methods

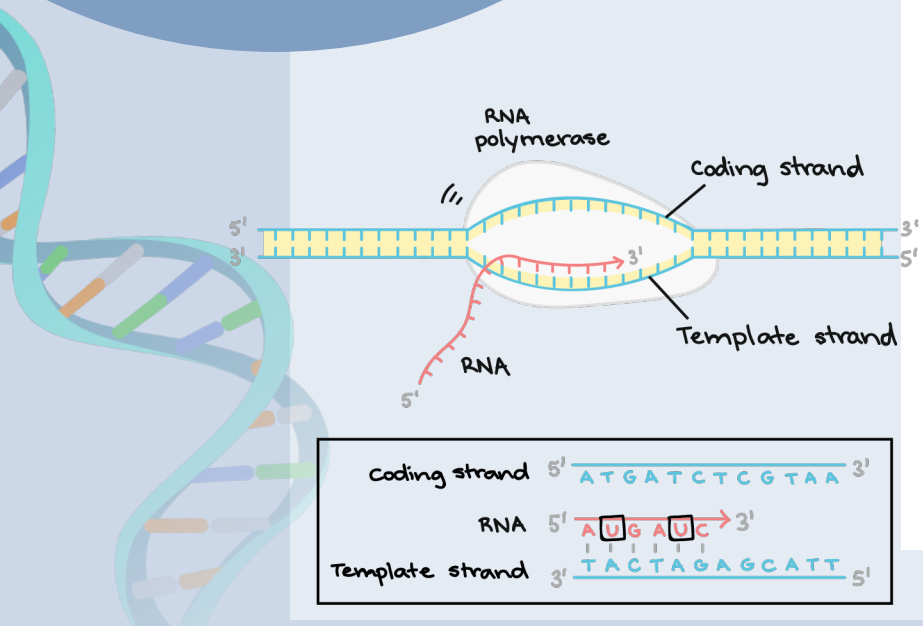
- **Transcriptomics**
  - Purpose: To measure which genes are active (expressed) and how much they're expressed in cells.
  - Examples
    - scRNA-seq
    - MARS-seq
    - Smart-seq
    - Drop-seq
    - Seq-Well
    - Chromium

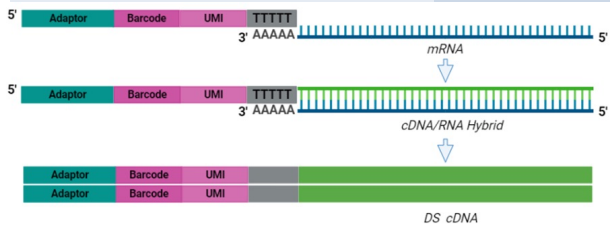


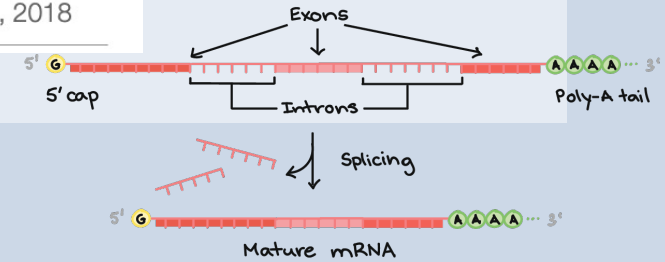
- Multimodal
  - Combines multiple types of data, such as gene expression, protein levels, and chromatin accessibility, within the same cells
  - Example:
    - ASAP-seq
    - SHARE-seq

Key Experimental Methods

Methods	Transcript coverage	UMI possibility	Strand specific	References
Tang method	Nearly full-length	No	No	Tang et al., 2009
Quartz-Seq	Full-length	No	No	Sasagawa et al., 2013
SUPeR-seq	Full-length	No	No	Fan X. et al., 2015
Smart-seq	Full-length	No	No	Ramskold et al., 2012
Smart-seq2	Full-length	No	No	Picelli et al., 2013
MATQ-seq	Full-length	Yes	Yes	Sheng et al., 2017
STRT-seq and STRT/C1	5′-only	Yes	Yes	Is Go to page 12 , 2012
CEL-seq	3′-only	Yes	Yes	Hashimshony et al., 2012
CEL-seq2	3′-only	Yes	Yes	Hashimshony et al., 2016
MARS-seq	3′-only	Yes	Yes	Jaitin et al., 2014
CytoSeq	3′-only	Yes	Yes	Fan H.C. et al., 2015
Drop-seq	3′-only	Yes	Yes	Macosko et al., 2015
InDrop	3′-only	Yes	Yes	Klein et al., 2015
Chromium	3′-only	Yes	Yes	Zheng et al., 2017
SPLiT-seq	3′-only	Yes	Yes	Rosenberg et al., 2018
sci-RNA-seq	3′-only	Yes	Yes	Cao et al., 2017
Seq-Well	3′-only	Yes	Yes	Gierahn et al., 2017
DroNC-seq	3′-only	Yes	Yes	Habib et al., 2017
Quartz-Seq2	3′-only	Yes	Yes	Sasagawa et al., 2018

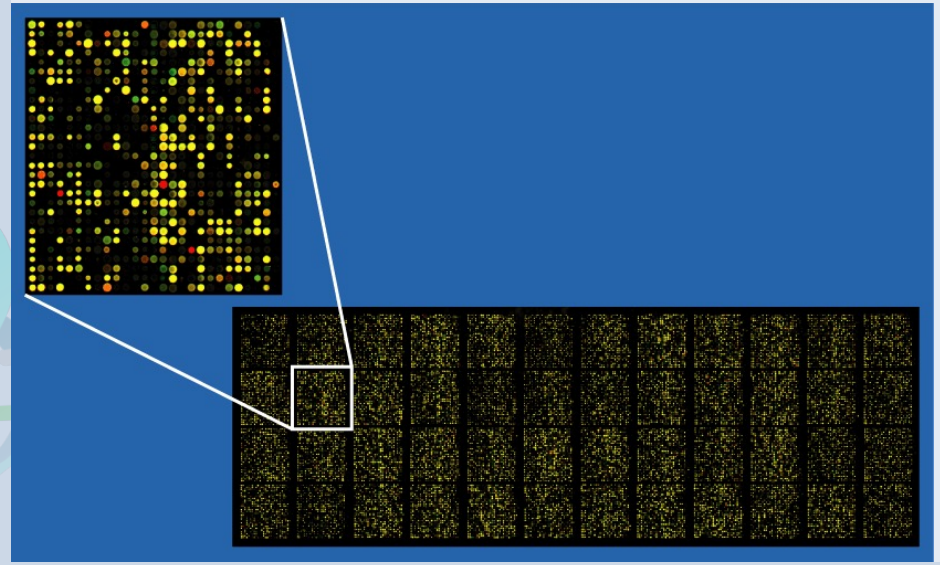


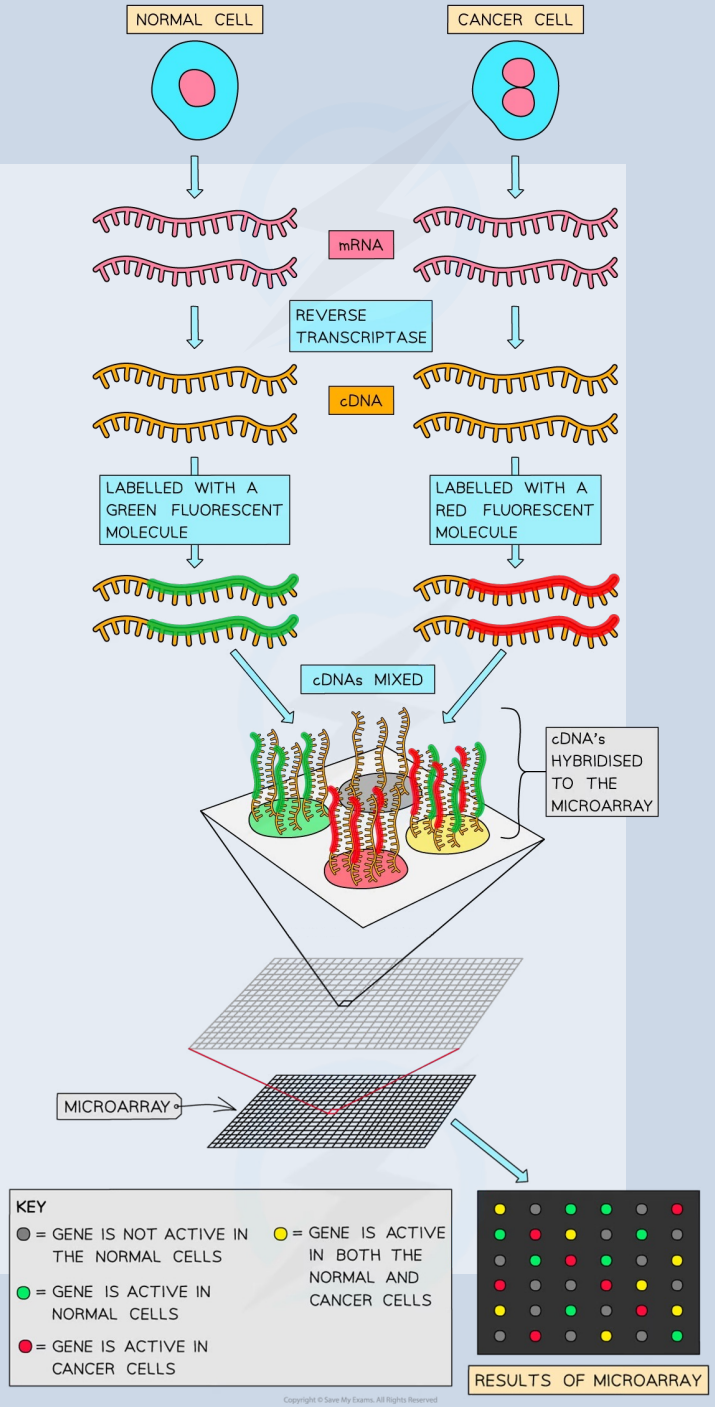




Microarrays

- Microarrays (DNA chip)
  - a high-throughput technology that allows researchers to study the **expression levels** of many genes or to detect **genetic variations** in a single experiment. It consists of a solid surface (e.g., a glass slide or silicon chip) with thousands of microscopic spots, each containing a specific DNA, RNA, or protein sequence.





Clustering

Briefings in Bioinformatics, 00(00), 2019, 1–13

doi: 10.1093/bib/bbz062
Advance Access Publication Date:
Review article

Clustering and classification methods for single-cell RNA-sequencing data

Ren Qi, Anjun Ma, Qin Ma and Quan Zou

Corresponding authors: Quan Zou, Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China, and Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, China. Tel: 170-9226-1008; E-mail: zouquan@nclab.net; Qin Ma, Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH, USA. Tel: 614-688-9857(O); E-mail: qin.ma@osumc.edu

Abstract

Appropriate ways to measure the similarity between single-cell RNA-sequencing (scRNA-seq) data are ubiquitous in bioinformatics, but using single clustering or classification methods to process scRNA-seq data is generally difficult. This has led to the emergence of integrated methods and tools that aim to automatically process specific problems associated with scRNA-seq data. These approaches have attracted a lot of interest in bioinformatics and related fields. In this paper, we systematically review the integrated methods and tools, highlighting the pros and cons of each approach. We not only pay particular attention to clustering and classification methods but also discuss methods that have emerged recently as powerful alternatives, including nonlinear and linear methods and descending dimension methods. Finally, we focus on clustering and classification methods for scRNA-seq data, in particular, integrated methods, and provide a comprehensive description of scRNA-seq data and download URLs.

Key words: single-cell RNA-seq; clustering; classification; similarity metric; sequences analysis; machine learning

Data Availability

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

There are no restrictions on data availability. The HLCA is fully public.
Data Availability Statement
The HLCA (raw and normalized counts, integrated embedding, cell type annotations and clinical and technical metadata) is publicly available and can be downloaded via cellxgene:
https://cellxgene.cziscience.com/collections/6f6d381a-7701-4781-935c-db10d30de293
The HLCA core reference model and embedding for mapping of new data to the HLCA can moreover be found on Zenodo, doi: 10.5281/zenodo.7599104.

The original, published datasets that were included in the HLCA can also be accessed under GEO accession numbers GSE135893, GSE143868, GSE128033, GSE121611, GSE134174, GSE150674, GSE151928, GSE136831, GSE128169, GSE171668, GSE132771, GSE126030, GSE161382, GSE155249, GSE135851, GSE145926, GSE178360, EGA study IDs EGAS00001004082, EGAS00001004344, EGAD00001005064, EGAD00001005065, and under urls https://www.synapse.org/#!Synapse:syn21041850, https://data.humancellatlas.org/explore/projects/c4077b3c-5c98-4d26-a614-246d12c2e5d7, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001750.v1.p1, https://www.nupulmonary.org/covid-19-ms2/?ds=full&meta=SampleName, https://figshare.com/articles/dataset/Single-cell_RNA-Seq_of_human_primary_lung_and_bronchial_epithelium_cells/11981034/1, https://covid19.lambrechtslab.org/downloads/Allcells.counts.rds, https://s3.amazonaws.com/dp-lab-data-public/lung-development-cancer-progression/PATIENT_LUNG_ADENOCARCINOMA_ANNOTATED.h5, https://github.com/theislab/2020_Mayr, https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-018-0449-8/MediaObjects/41586_2018_449_MOESM4_ESM.zip, http://blueprint.lambrechtslab.org/#/099de49a-cd68-4db1-82c1-cc7acd3c6d14/*/welcome, https://www.covid19cellatlas.org/index.patient.html (see also Supplementary Data Table 1).
GWAS summary statistics of COPD(GWAS catalog ID: GCST007692, dbGaP accession number: phs000179.v6.p2), IPF, and of lung adenocarcinoma(GWAS catalog ID: GCST004748, dbGaP accession number: phs001273.v3.p2) were made available on dbGap upon request. Summary statistics of lung function (GWAS catalog ID: GCST007429), of asthma (GWAS catalog ID: GCST010043), and of depression (used as negative control, GWAS catalog ID: GCST005902) were publicly available.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-15647-5 OPEN

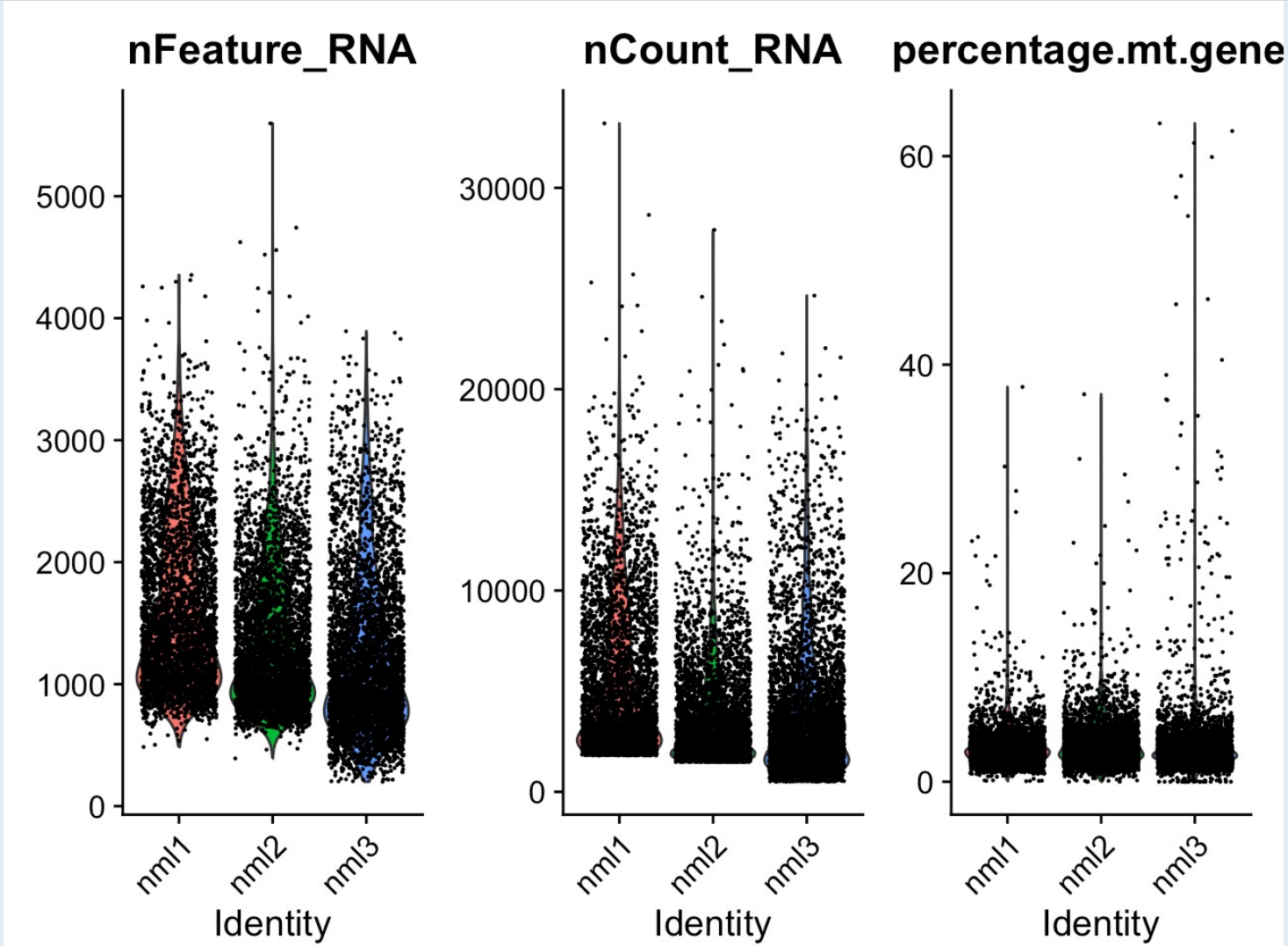
Collagen-producing lung cell atlas identifies multiple subsets with distinct localization and relevance to fibrosis

Tatsuya Tsukui[1], Kai-Hui Sun[1], Joseph B. Wetter[2], John R. Wilson-Kanamori[3], Lisa A. Hazelwood[2], Neil C. Henderson[3], Taylor S. Adams[4], Jonas C. Schupp[4], Sergio D. Poli[5], Ivan O. Rosas[5], Naftali Kaminski[4], Michael A. Matthay[6], Paul J. Wolters[7] & Dean Sheppard[1✉]

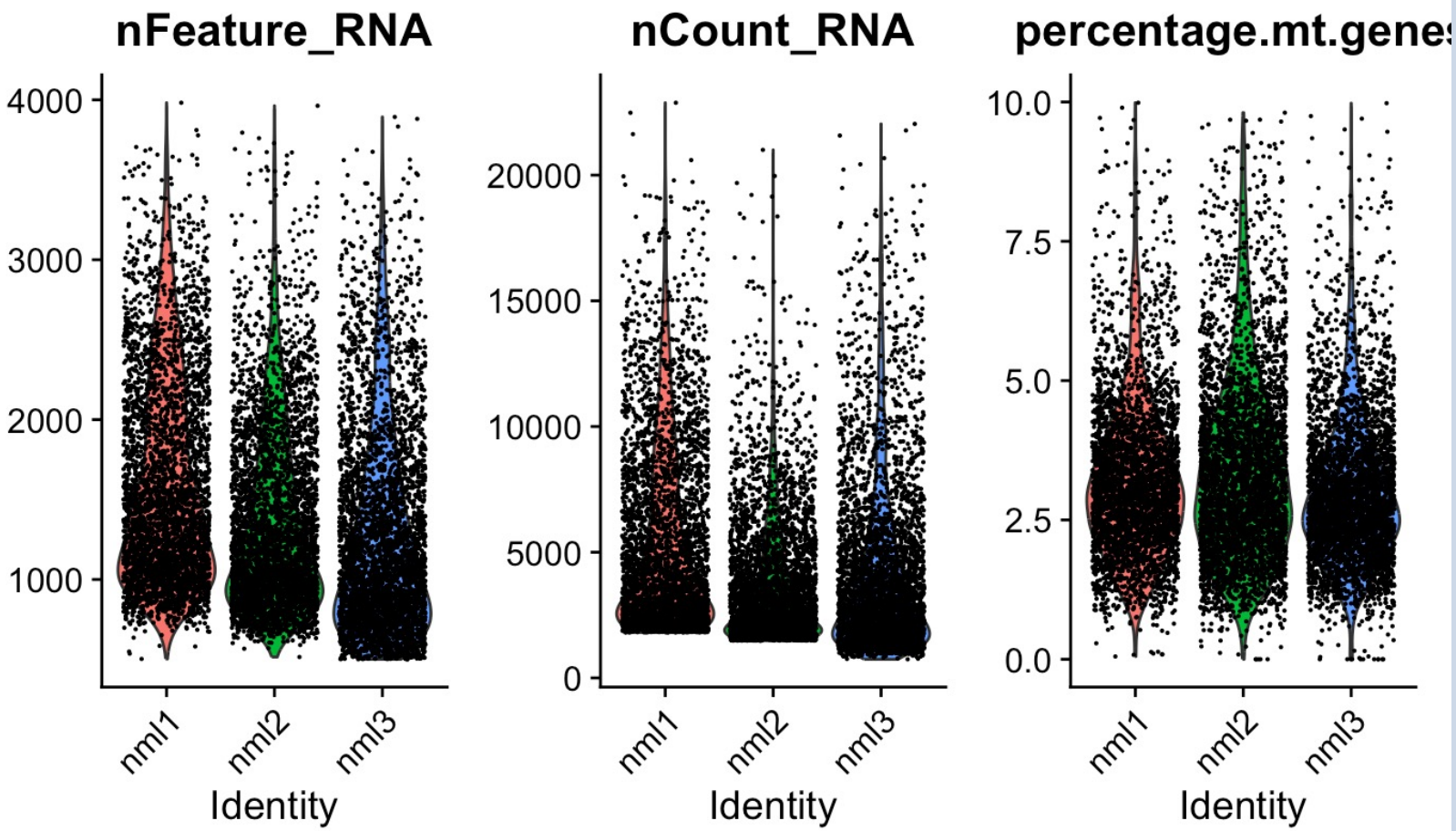
Quality Control

- Three main key factors : QC covariates
- Count depth
  - Low count depth ➡️ damaged or broken cells
  - Very high count depth ➡️ doublets
  - Threshold ➡️ 500, 20,000
- The number of genes per barcode
  - Low ➡️ dead cells or empty droplets
  - Very high ➡️ doublets
- The fraction of counts from mitochondrial genes per barcode
  - Very high ➡️ dying or stressed cells, broken cells
  - A normal mitochondrial fraction for a healthy cell is typically less than 5–10%

Before
quality control



after
Quality control



Normalization

$$log - normalized\ Value_{ij} = \ ln\left(\frac{X_{ij}}{\sum_i X_{ij}} \times scale\ factor + 1\right)$$

X_{ij} : Raw count of $gene_i$ in $cell_j$

$\sum_i X_{ij}$: Total counts in $cell_j$

scale factor: Arbitrary scaling factor (default is 10,000).

Find Variable Features

- The goal of **FindVariableFeatures()** is to identify genes whose expression varies significantly across cells.
- hese genes are often biologically meaningful, as they may represent **cell-type-specific markers** or genes involved in dynamic processes like **differentiation.**
- For each gene, the mean expression level across all cells is calculated.
- The variance of each gene's expression across all cells is also calculated.
- In single-cell RNA-seq data, there is typically a strong relationship between a gene’s mean expression and its variance.
    - Genes with higher mean expression tend to have higher variance.

Find Variable Features

- The **variance-stabilizing transformation (VST)** method models this relationship using a **local polynomial regression (loess)**.

$$\sigma_i^2 = f(\mu_i) + \epsilon_i$$

- f(μ_i) is the expected variance for a given mean µi, and ϵ_i is the residual variance.
- The residual variance (ϵ_i) represents the deviation of a gene's variance from the expected variance given its mean expression. Genes with **high residual variance** are considered **highly variable.**

Find Variable Features

- Seurat uses a **local polynomial regression (loess)** to fit a smooth curve to the observed mean-variance relationship.

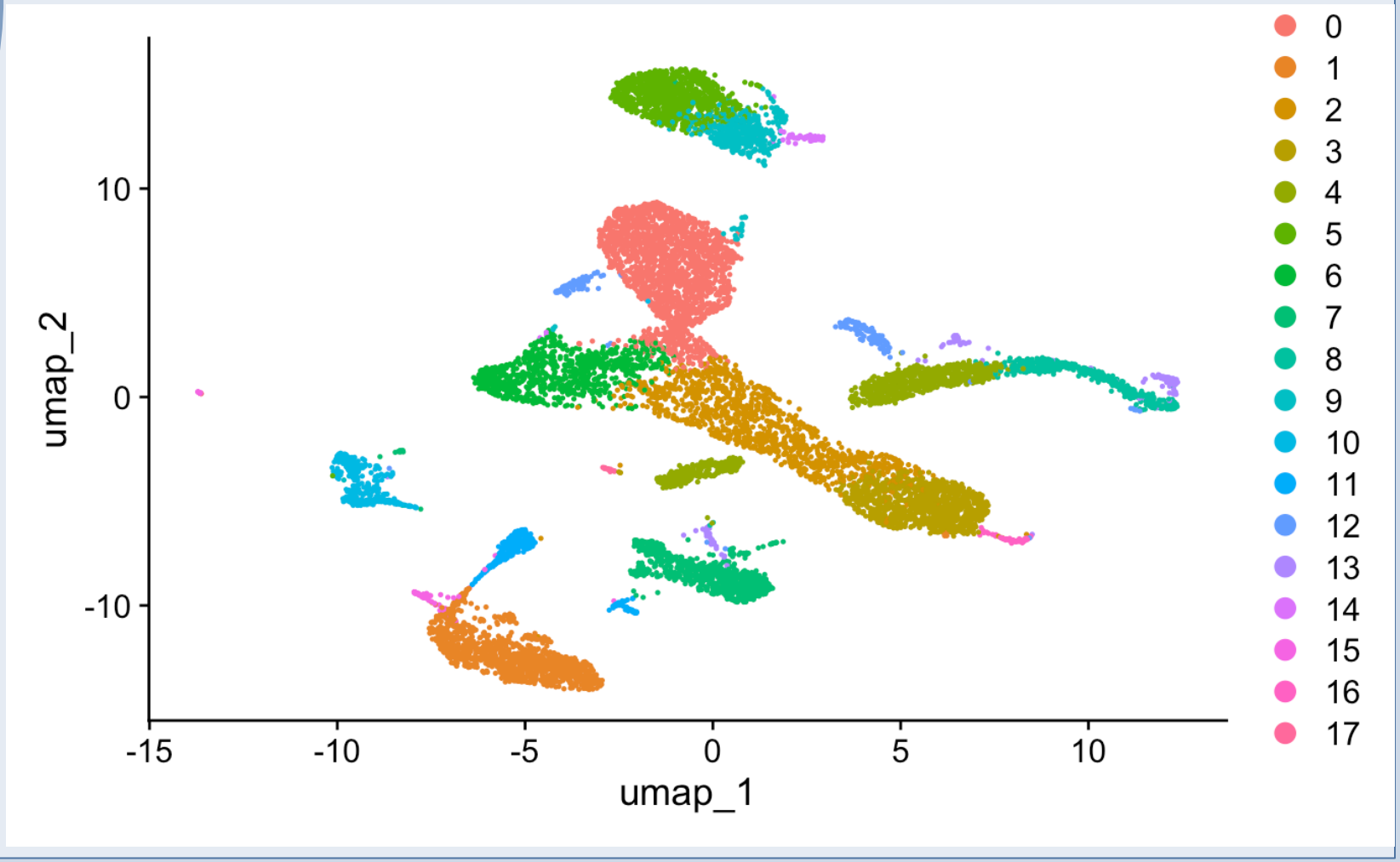
- The loess model is a non-parametric method that fits a smooth curve to the data by performing **weighted linear regression** in local neighborhoods. It is robust to outliers and adapts well to the shape of the data.

Find Variable Features

PCA

```
PC_ 1
Positive:  ADIRF, DSTN, CAV1, TIMP3, EMP2, TSC22D1, NGFRAP1, SPARCL1, CNN3, PTRF
           TFPI, CYB5A, AK1, ID1, TM4SF1, LIMCH1, C8orf4, CALD1, MT1E, SOD3
           SPTBN1, CYR61, SLPI, CAV2, SFTA2, TPM1, SFTPB, SELENBP1, SEPP1, IFITM3
Negative:  TYROBP, FCER1G, CYBA, LAPTM5, LYZ, AIF1, CTSS, MS4A7, C1orf162, VSIG4
           SH3BGRL3, FTL, FTH1, HLA-DRA, CD68, CD163, C1QA, SPI1, ALOX5AP, CD74
           OLR1, COTL1, C1QC, CTSB, C1QB, HLA-DPB1, MS4A6A, CAPG, EVI2B, MS4A4A
PC_ 2
Positive:  SPARCL1, TIMP3, GPX3, CALD1, A2M, SPARC, MT2A, COX7A1, MGP, VAMP5
           RAMP2, TPM1, MT1M, GNG11, B2M, CLDN5, TMEM100, TCF4, IFITM3, LMCD1
           MT1X, EGFL7, RAMP3, DCN, EPAS1, NPDC1, HYAL2, PECAM1, CLEC14A, PPP1R14A
Negative:  SFTA2, SFTPB, NAPSA, PEBP4, SFTPD, MUC1, SLC34A2, SFTA3, S100A14, SFTPA1
           SFTPA2, SLPI, PGC, ELF3, KRT19, SDR16C5, FOLR1, CXCL17, AGR3, HOPX
           SFTPC, LAMP3, KRT8, C16orf89, CLDN4, FXYD3, KRT18, NKX2-1, MGST1, GKN2
```

Clusters



Resolution:0.5

Next session