Single cell RNA-seq data analysis

Lecturer: Yao-Ming Chang

Assistant Research Scientist, IBMS

Nov. 14, 2022

Computational Medicine Core Lab in IBMS

Contents

- Basic concept of single cell RNA sequencing (scRNA-seq)
  - Bulk vs. single cell RNA-seq
  - Why scRNA-seq?
  - Droplet based scRNA-seq data
  - What dose scRNA-seq data look like?
  - What do we expect to get from the scRNA-seq data?

- Standard analysis workflow
  - Import → QC → dimensionality reduction → data correction → clustering → marker genes → cell type annotation → functional enrichment → gene expression dynamics (trajectory prediction)

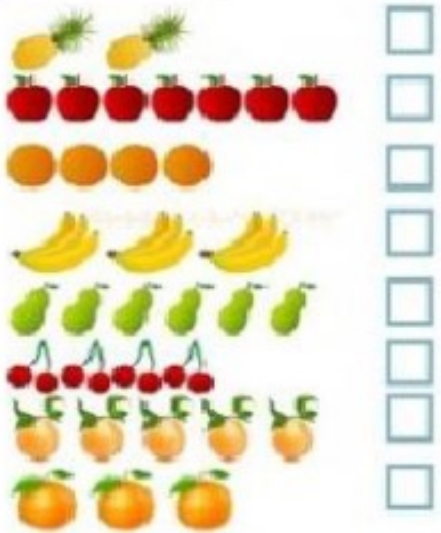
Basic concept of single cell RNA-seq

BULK VS SINGLE CELL RNA-SEQ

RNA-Seq

scRNA-Seq

Average expression level
- Comparative transcriptomics
- Disease biomarker
- Homogenous systems



Separate populations
- Define heterogeneity
- Identify rare cell populations
- Cell population dynamics

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

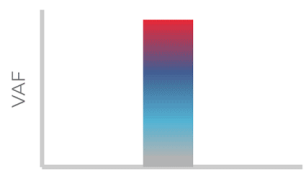
HETEROGENEOUS CELL POPULATION

BULK SEQUENCING

RNA IS MIXED TOGETHER

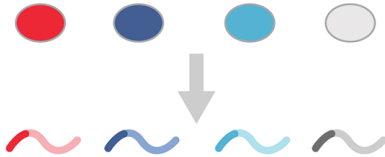
X

SUBPOPULATIONS NOT DEFINED

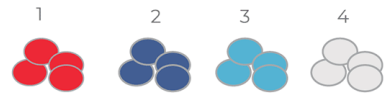


"AVERAGED" READOUT

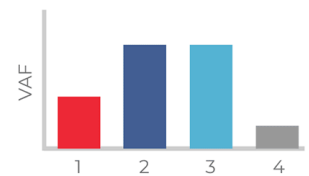
SINGLE-CELL RNA SEQUENCING



RNA FROM EACH CELL IS BARCODED



SUBPOPULATIONS DEFINED

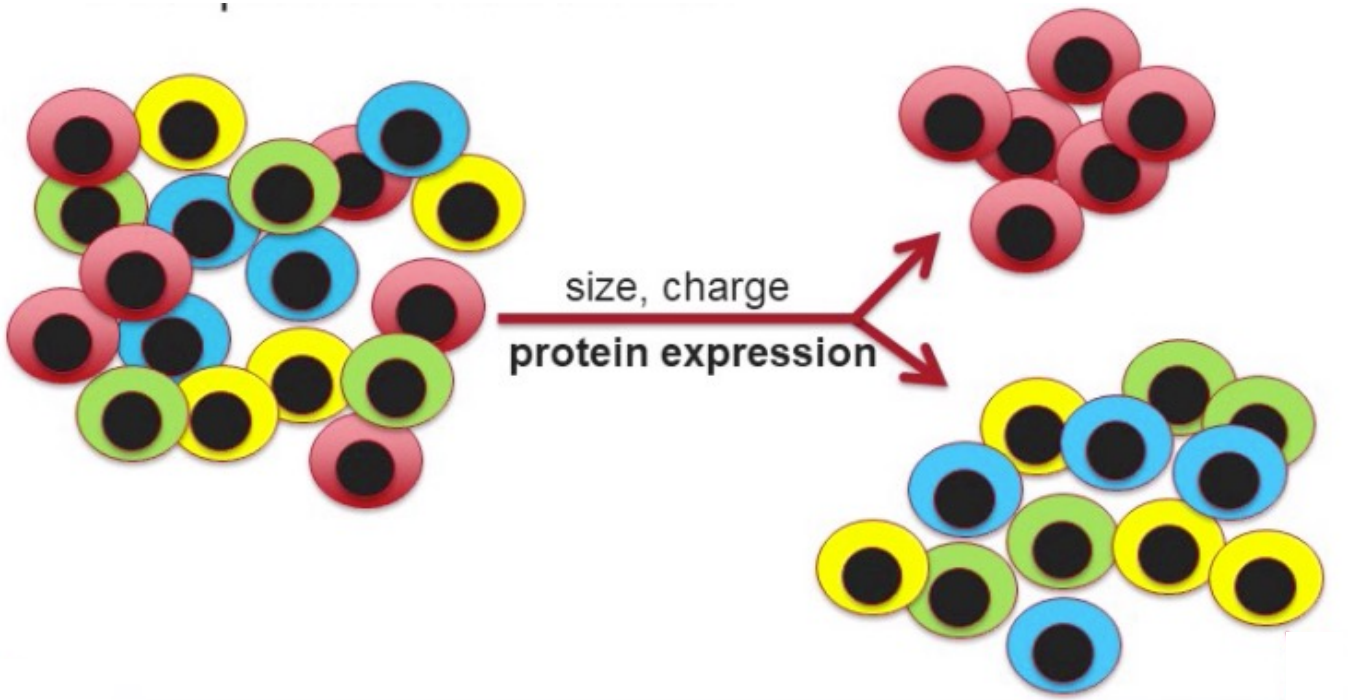


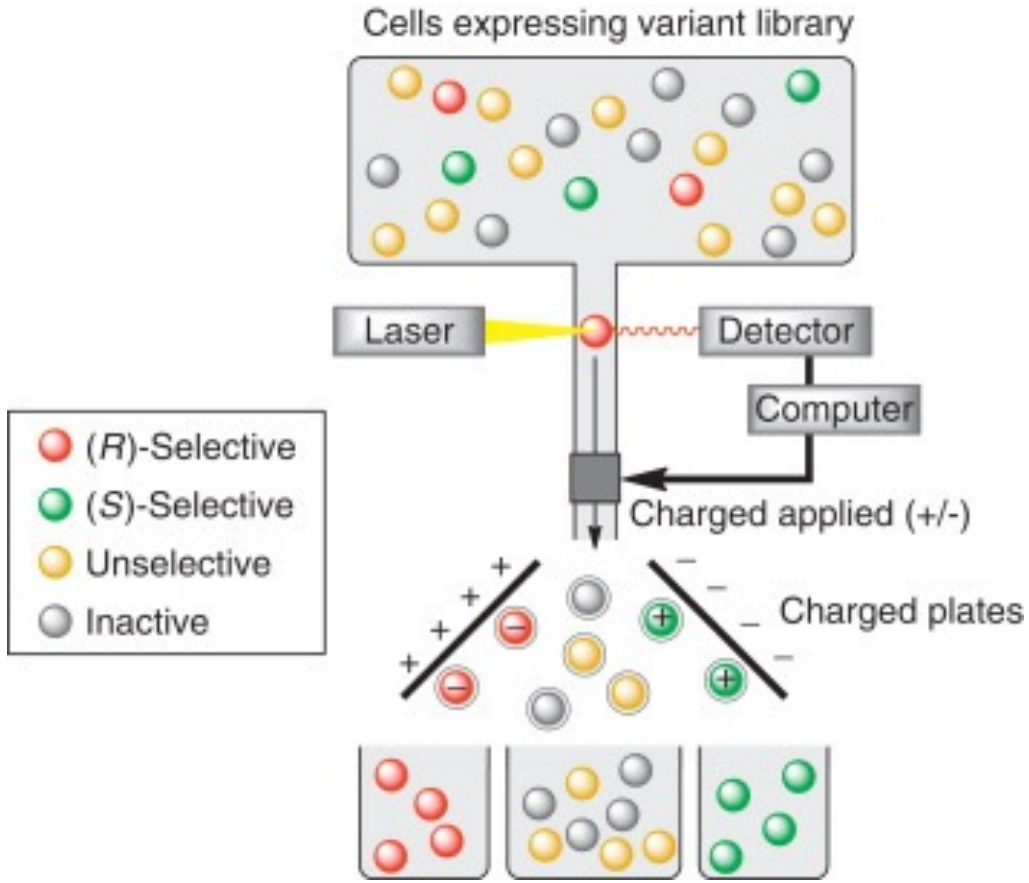
CELL TYPE-SPECIFIC READOUT

Why consider performing scRNA-seq?

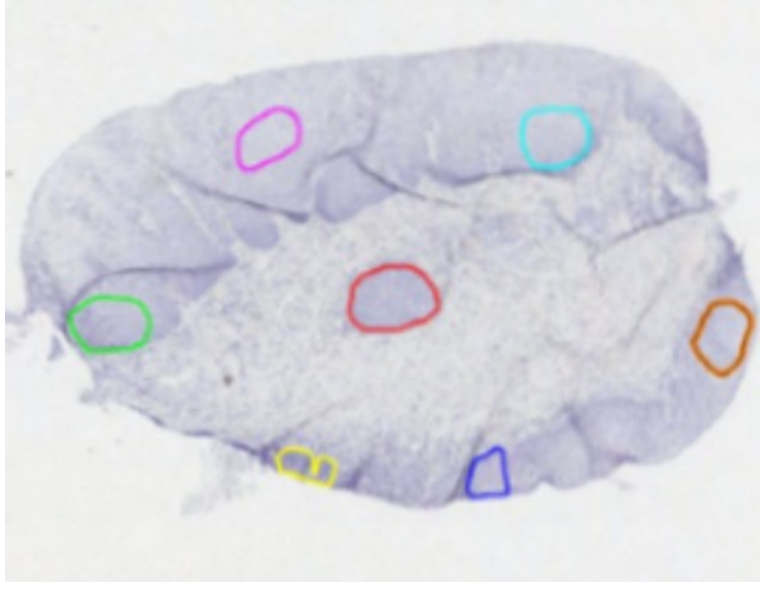
- scRNA-seq permits comparison of the transcriptomes of individual cells. Therefore, a major use of scRNA-seq has been to assess transcriptional similarities and differences within a population of cells, with early reports revealing previously unappreciated levels of heterogeneity, for example in embryonic and immune cells

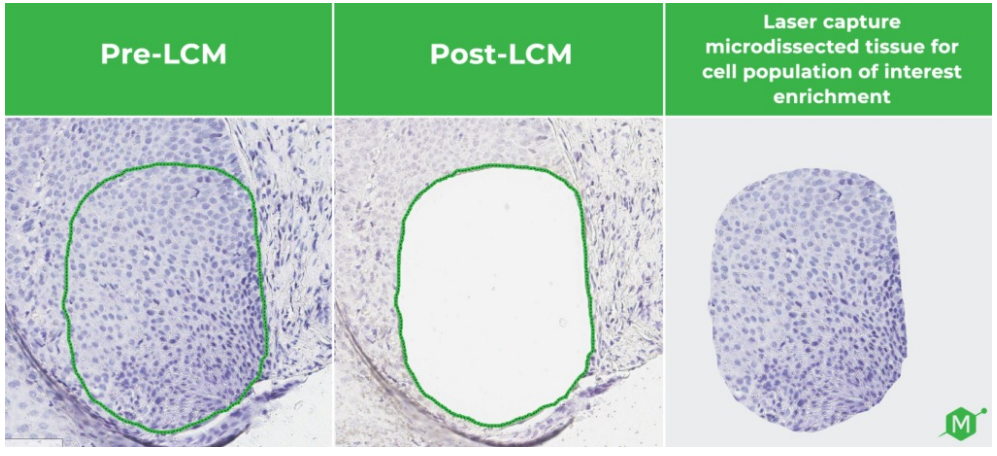
FACS (fluorescence activated cell sorting)

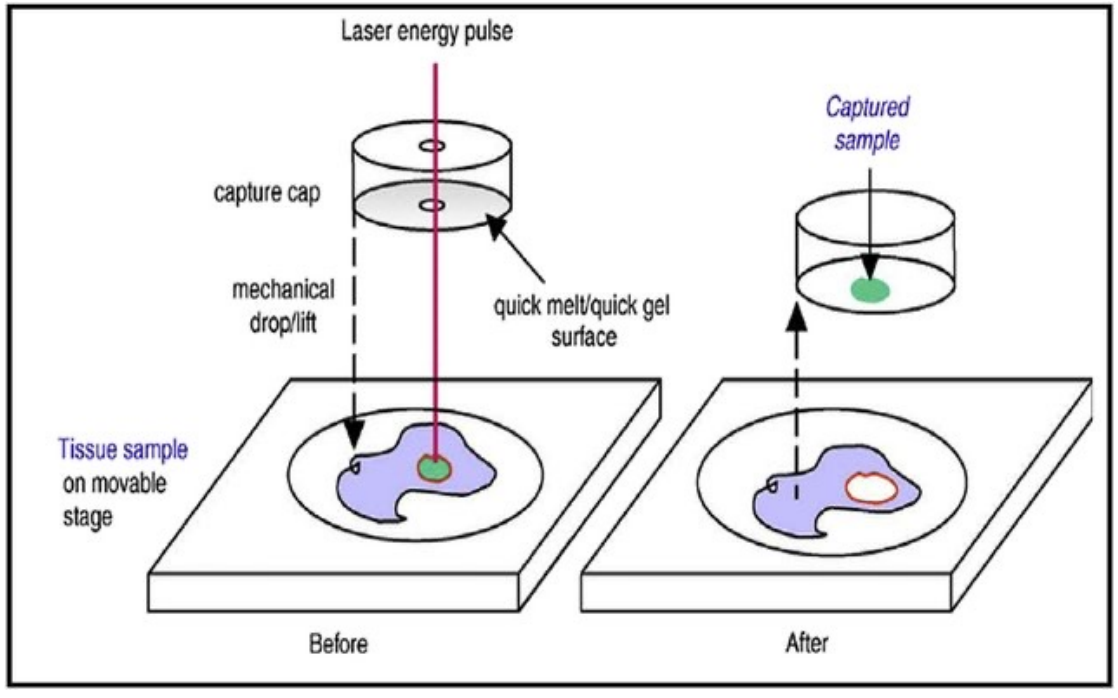




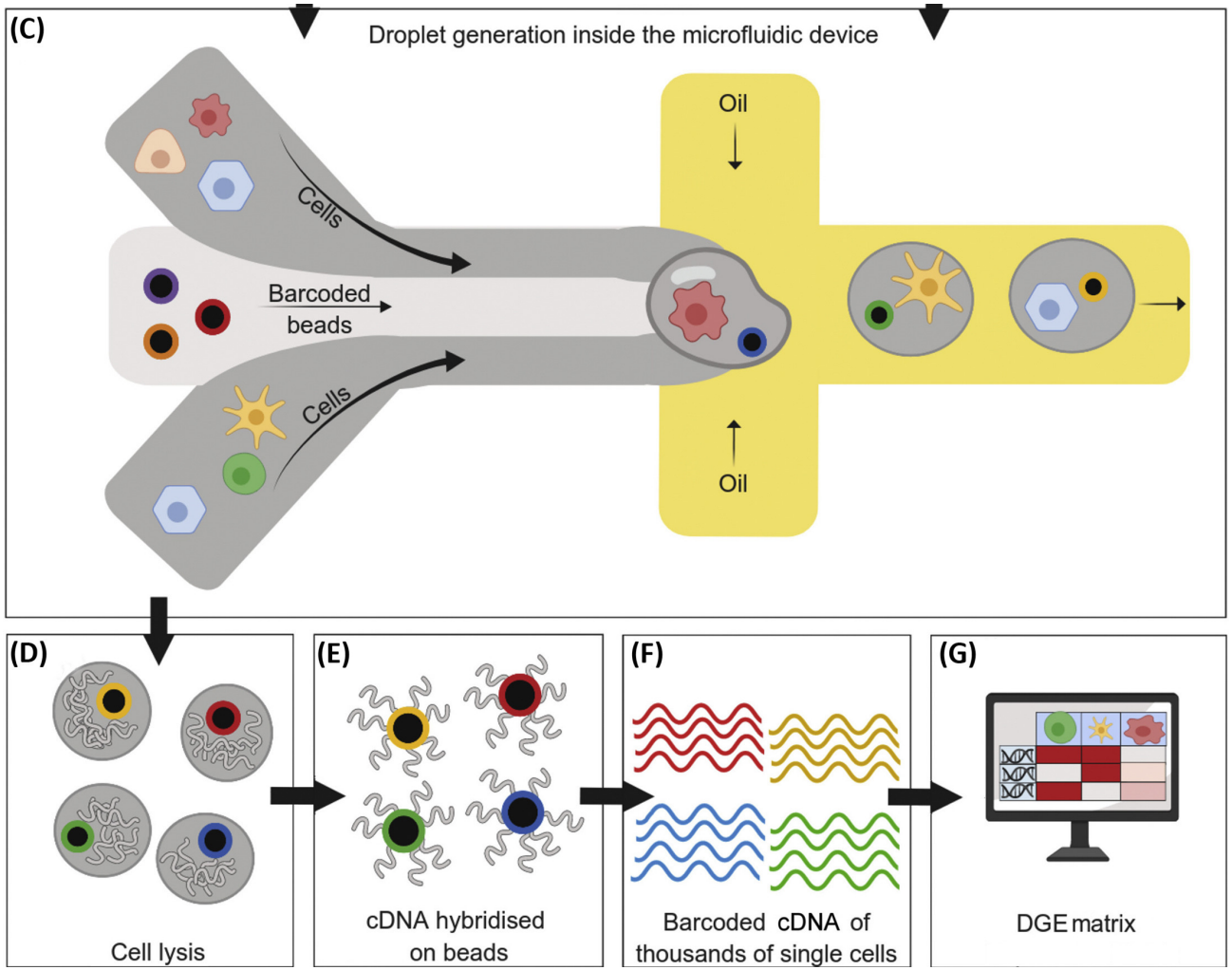
Laser capture microdissection (LCM)

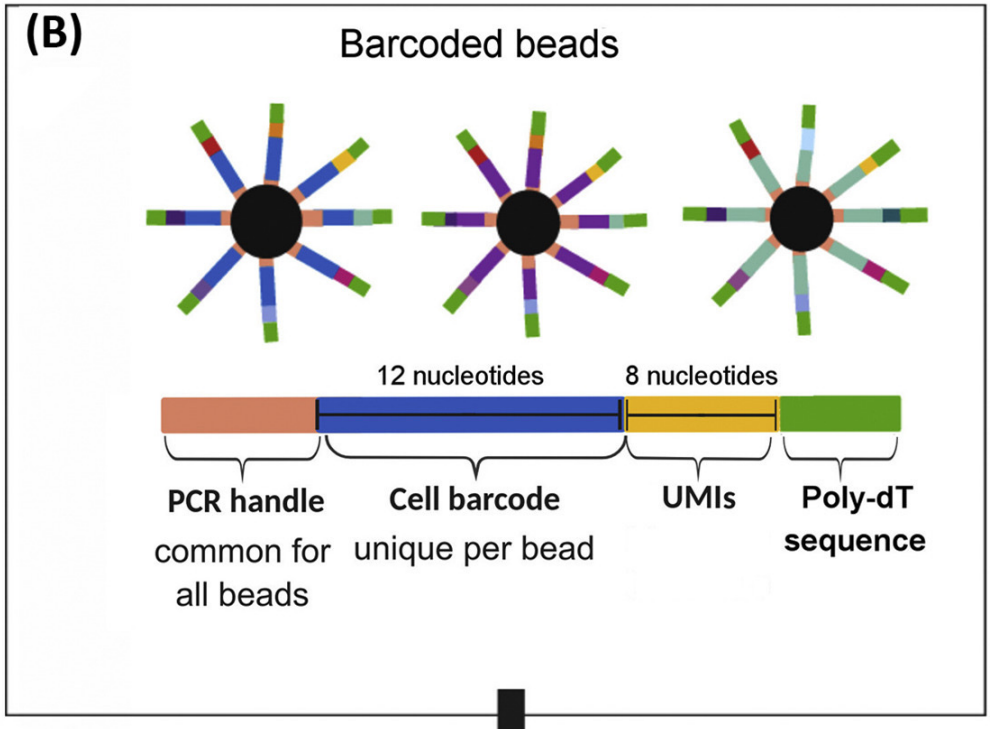




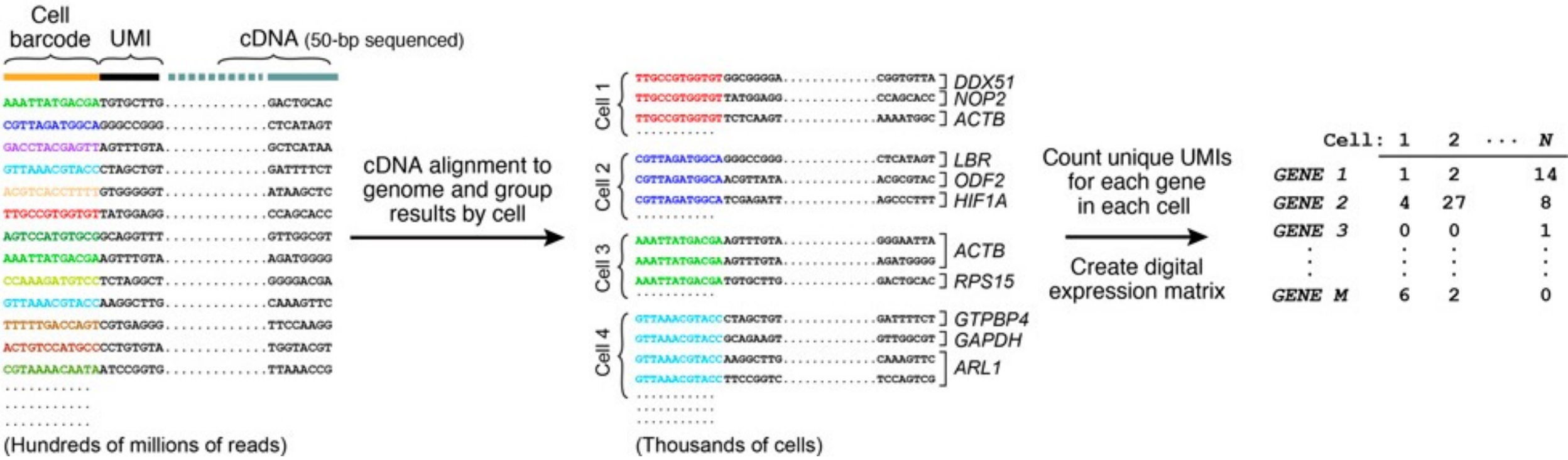


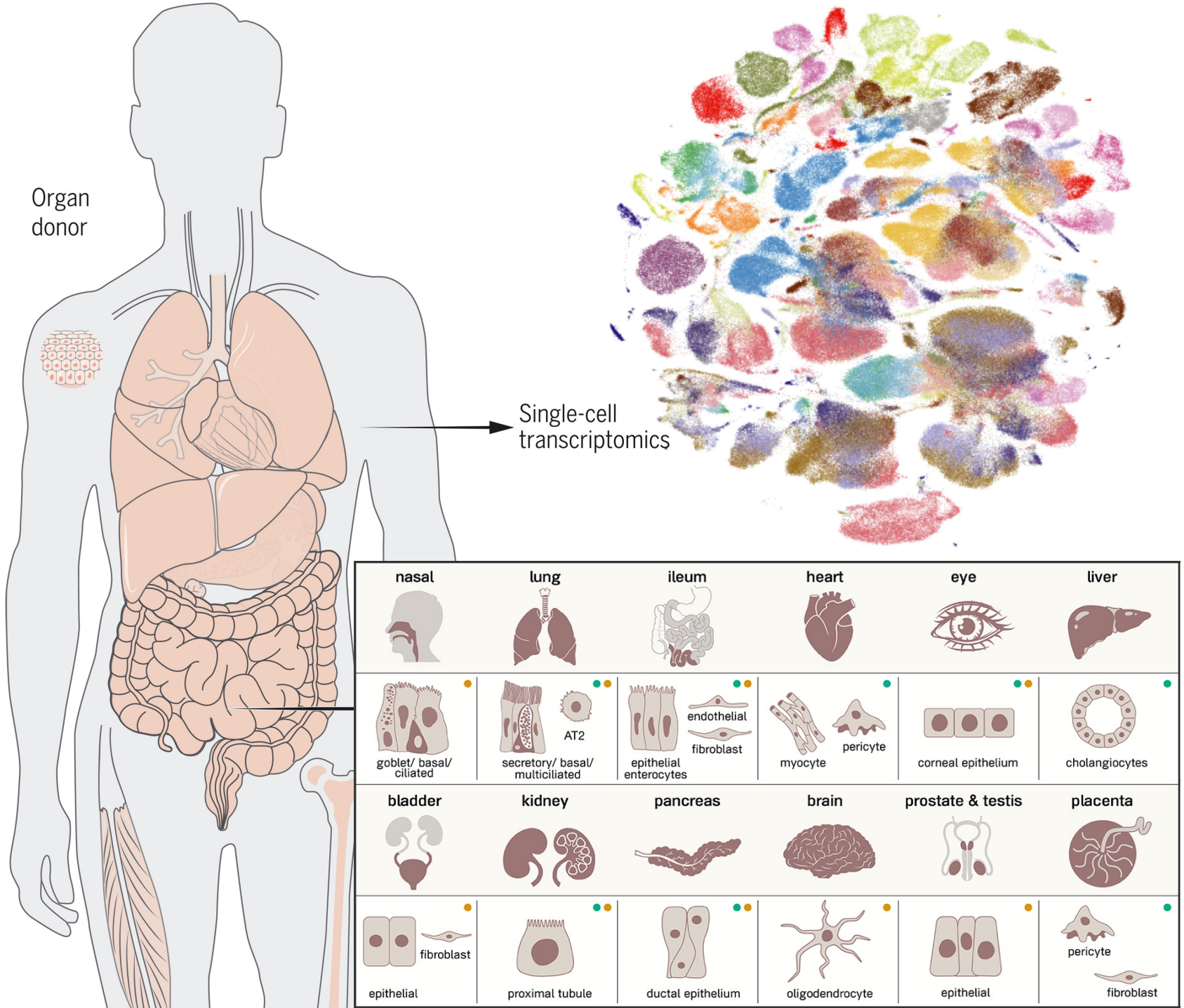
Droplet based single cell RNA sequencing





Droplet based single cell RNA sequencing





doi:10.1126/science.abl4896

However, there are limitations in scRNA-seq:

- low capture efficiency (~8% mRNAs in a cell were captured)
- higher level of technical noise than bulk RNA-seq data
- multi-cells in a droplet (doublet)
- dead cells (high proportion of mitochondrial RNAs)

Standard scRNA-seq data analysis

Read alignment (similar to bulk RNA-seq)

10X GENOMICS

Products Research Areas Resources Support Company

Support › Single Cell Gene Expression › Software

SEARCH Q&A CONTACT SUPPORT

SOFTWARE › PIPELINES

CELL RANGER

Introduction

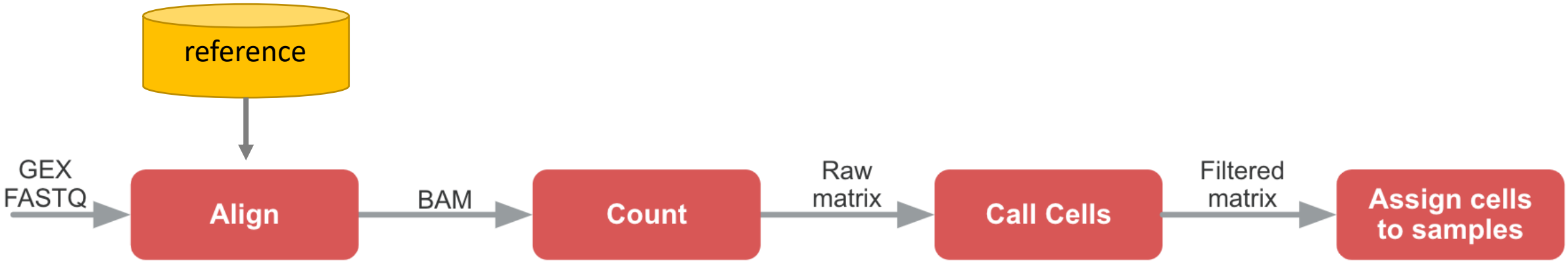
Downloads

- Download Links
- System Requirements
- Installing Cell Ranger
- Release Notes

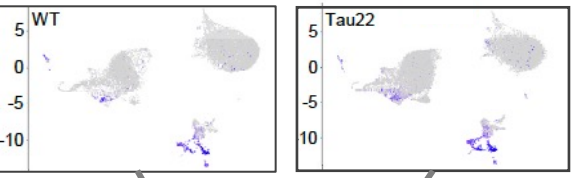
Creating a Reference Package with cellranger mkref

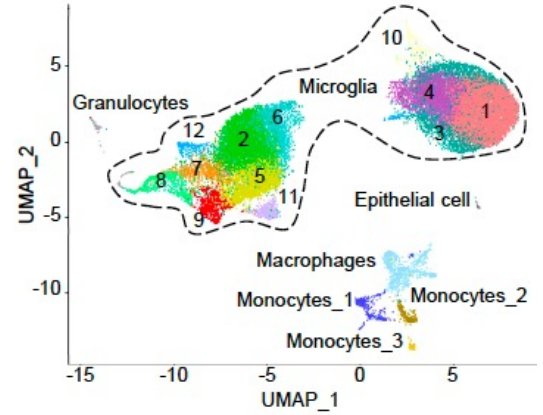
Cell Ranger provides pre-built human (hg19, GRCh38), mouse (mm10), and ercc92 reference packages for read alignment and gene expression quantification in `cellranger count`.
To create and use a custom reference package, Cell Ranger requires a reference genome sequence (FASTA file) and gene annotations (GTF file).

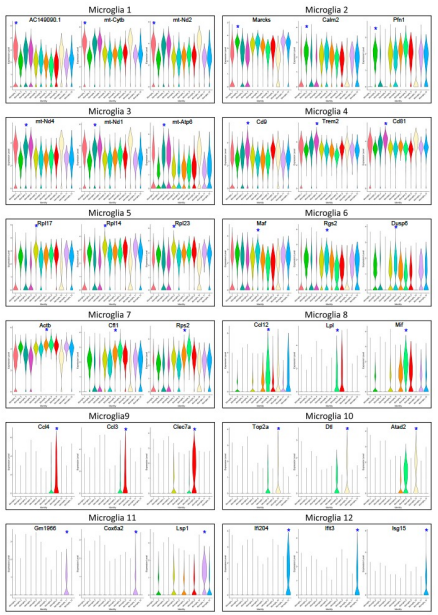
A tutorial 'Build a Custom Reference With cellranger mkref' is available to walk you through the steps.



Asc-Seurat – Analytical single-cell Seurat-based web application







SEURAT

Sample
Sample
Sample
Sample
Sample
Integration pipeline
Quality Control
PCA
Clustering
Exclude or select clusters
Yes
Select clusters of interest?
No
Lists of markers
Markers identification
Differential expression analysis
Lists of DE genes
Expression visualization
High resolution plots

v2.1 May 26, 2021

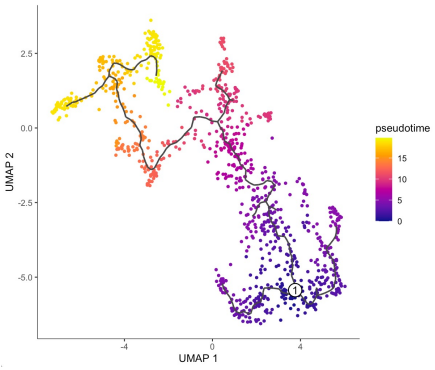
Dynverse

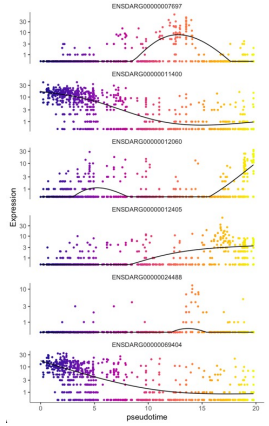
Clustered data
Trajectory inference
Identification of the most predictive genes
Expression visualization within the trajectory
High resolution plots
Lists of DE genes

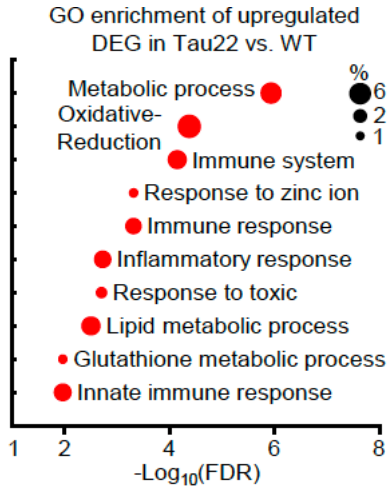
BioMart

BioMart annotation
GO terms enrichment analysis
List of enriched GO terms
Annotated genes

Pereira et al., bioRxiv, 2021







RStudio

Go to file/function | Addins | Project: (None)

ingleR.R | Basic_analysis.R | Biomart.R | NOIseq_no_replicate_example_2.r | NOIseq_with_replicate_example.r | DEseq2_PCA_e

Source on Save | Run | Source

```
1  library(dplyr)
2  library(Seurat)
3  library(patchwork)
4
5  ##import data from 10x output folder, QC and selecting cells for further analysis
6  input.data <- Read10X(data.dir = "/Volumes/CONTAX/DR_RB_Yang_Scube_genes_expression_profiles_in_endothelial_scRNA-seq_da
7  W0 <- CreateSeuratObject(counts = input.data, project = "W0", min.cells = 3, min.features = 200)
8  W0
9  W0$source <- "W0"
10 W0[["percent.mt"]] <- PercentageFeatureSet(W0, pattern = "^mt-")
11 VlnPlot(W0, features = c("nFeature_RNA", "nCount_RNA", "percent.mt"), ncol = 3)
12 #W0 <- subset(W0, subset = nFeature_RNA > 200 & nFeature_RNA < 6000 & percent.mt < 15)
13 W0 <- NormalizeData(W0, verbose = FALSE)
14 W0 <- FindVariableFeatures(W0, selection.method = "vst")
15
16 ##import data from 10x output folder, QC and selecting cells for further analysis
17 input.data <- Read10X(data.dir = "/Volumes/CONTAX/DR_RB_Yang_Scube_genes_expression_profiles_in_endothelial_scRNA-seq_da
18 W1 <- CreateSeuratObject(counts = input.data, project = "W1", min.cells = 3, min.features = 200)
19 W1
20 W1$source <- "W1"
21 W1[["percent.mt"]] <- PercentageFeatureSet(W1, pattern = "^mt-")
22 VlnPlot(W1, features = c("nFeature_RNA", "nCount_RNA", "percent.mt"), ncol = 3)
23 W1 <- subset(W1, subset = nFeature_RNA > 200 & nFeature_RNA < 6000 & percent.mt < 15)
24 W1 <- NormalizeData(W1, verbose = FALSE)
25 W1 <- FindVariableFeatures(W1, selection.method = "vst")
26
```

16:50 (Top Level) R Script

Console | Terminal | Jobs

/Volumes/Extented-Data/My Drive/temp/LSL_course_sample_data/

```
R version 4.1.0 (2021-05-18) -- "Camp Pontanezen"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
```

Environment | History | Connections | Tutorial

Import Dataset | List

Global Environment

Environment is empty

Files | Plots | Packages | Help | Viewer

R: Save a ggplot (or other grid object) with sensible defaults | Find in Topic

ggsave {ggplot2} R Documentation

Save a ggplot (or other grid object) with sensible defaults

Description

`ggsave()` is a convenient function for saving a plot. It defaults to saving the last plot that you displayed, using the size of the current graphics device. It also guesses the type of graphics device from the extension.

Usage

```
ggsave(
  filename,
  plot = last_plot(),
  device = NULL,
  path = NULL,
  scale = 1,
  width = NA,
  height = NA,
  units = c("in", "cm", "mm", "px"),
  dpi = 300,
  limitsize = TRUE,
  bg = NULL,
  ...
)
```

Arguments

Web-based scRNA-seq data analysis tool

Tool name	Published on	Journal name	Data	Need registration
SingleCAnalyzer	23 May 2022	Frontiers in Bioinformatics	FASTQ	yes
ICARUS	10 May 2022	Nucleic Acids Research	gene-cell count matrix	no
SC1	5 Aug 2021	Journal of Computational Biology	gene-cell count matrix	no

Published online 10 May 2022 *Nucleic Acids Research, 2022, Vol. 50, Web Server issue* ***W427–W433***
https://doi.org/10.1093/nar/gkac322

ICARUS, an interactive web server for single cell RNA-seq analysis

Andrew Jiang[1,*], Klaus Lehnert[1], Linya You[2] and Russell G. Snell[1]

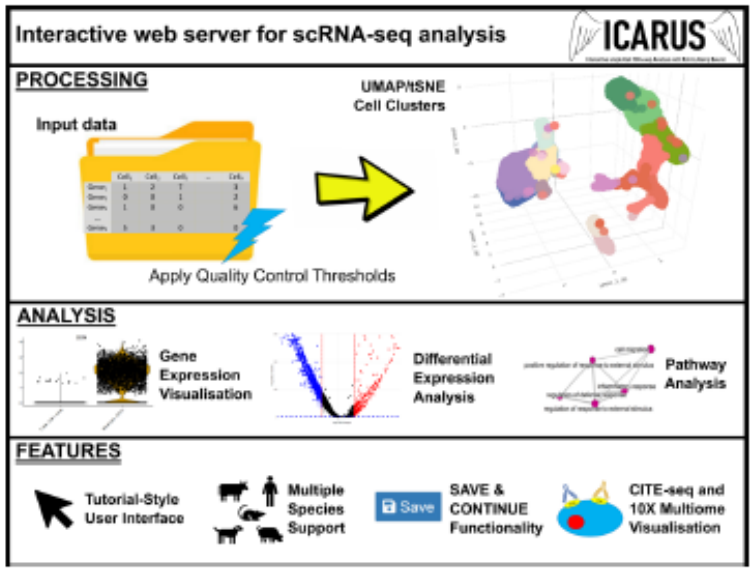
[1]Applied Translational Genetics Group, School of Biological Sciences, The University of Auckland, Auckland, New Zealand and [2]Department of Human Anatomy & Histoembryology, School of Basic Medical Sciences, Fudan University, Shanghai, China

Received March 24, 2022; Revised April 14, 2022; Editorial Decision April 19, 2022; Accepted April 21, 2022

ABSTRACT

Here we present ICARUS, a web server to enable users without experience in R to undertake single cell RNA-seq analysis. The focal point of ICARUS is its intuitive tutorial-style user interface, designed to guide logical navigation through the multitude of pre-processing, analysis and visualization steps. ICARUS is easily accessible through a dedicated web server (https://launch.icarus-scrnaseq.cloud.edu.au/) and avoids installation of software on the user's computer. Notable features include the facility to apply quality control thresholds and adjust dimensionality reduction and cell clustering parameters. Data is visualized through 2D/3D UMAP and t-SNE plots and may be curated to remove potential confounders such as cell cycle heterogeneity.

GRAPHICAL ABSTRACT



Welcome to ICARUS (**I**nteractive single **C**ell RNA-seq **A**nalysis with **R** shiny **U**sing **S**eurat)

This application was designed to guide the user through single cell RNA-seq analysis using the Seurat scRNA-seq analysis toolkit via a tutorial style interface. It offers user control over each of the steps to personalise analysis based on the dataset of interest. Graphical outputs at each analysis step ensures easy and logical interpretation.

The purpose of this application is to allow the user to interactively visualize single cell RNA-seq data without the requirement of previous R programming knowledge.

Features include:

1. Tutorial inspired user interface!
2. Support for 11 common species!
3. Adjust your own quality control thresholds!
4. Adjust your own dimensionality reduction and clustering parameters!
5. 3D UMAP and t-SNE plots!
6. Data correction for cell cycle effects!
7. Removal of cell doublets (multiplets) with DoubletFinder!
8. Labelling of cell clusters with sctype and SingleR!
9. Gene expression and gene pathway visualisation!
10. Trajectory analysis with Monocle3!
11. Differential expression analysis and gene set enrichment analysis with ClusterProfiler and ReactomePA!
12. Custom differential expression analysis with user selected cell groups to compare!
13. Integration with second dataset and adjustment for batch effects!
14. Support for multimodal analysis (i.e. CITE-seq, 10X multiome kit)!
15. Save and continue functionality!
16. Downloadable tables and plots!

Please refer to the "Help" tab on the sidebar menu for troubleshooting.

NEW

Choose from the following species

HOMO SAPIENS

START

PRESS START TO BEGIN

CONTINUE

Load previously saved enviroment

BROWSE... No file selecte

Applied Translational Genetics

THE UNIVERSITY OF AUCKLAND NEW ZEALAND | CENTRE FOR BRAIN RESEARCH

Created by: Andrew Jiang
Email: ajia169@aucklanduni.ac.nz
Applied Translational Genetics Group
Centre for Brain Research
The University of Auckland

Step 1: Load your data

ICARUS

Interactive single Cell RNA-seq Analysis with Rshiny Using Seurat

DATA INPUT

A

Single sample

	$Cell_1$	$Cell_2$	$Cell_3$	...	$Cell_X$
$Gene_1$	1	2	7		
$Gene_2$	0	0	1		
$Gene_3$	1	0	0		6
...					
$Gene_X$	5	3	0		0

Column names must not contain any underscores

B

Multiple samples

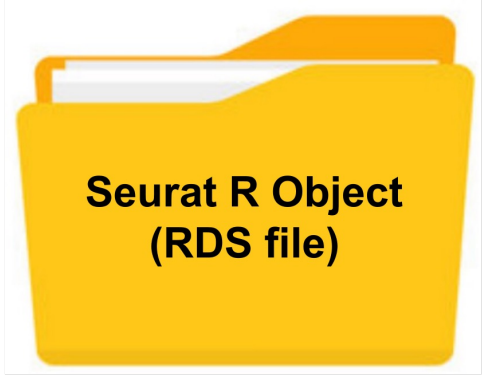
	$S1\_Cell_1$	$S1\_Cell_2$	$S2\_Cell_3$	...	$S2\_Cell_X$
$Gene_1$	1	2	7		
$Gene_2$	0	0	1		
$Gene_3$	1	0	0		6
...					
$Gene_X$	5	3	0		0

Each sample can be denoted by an identifier separated by an underscore

C

10x GENOMICS®

D

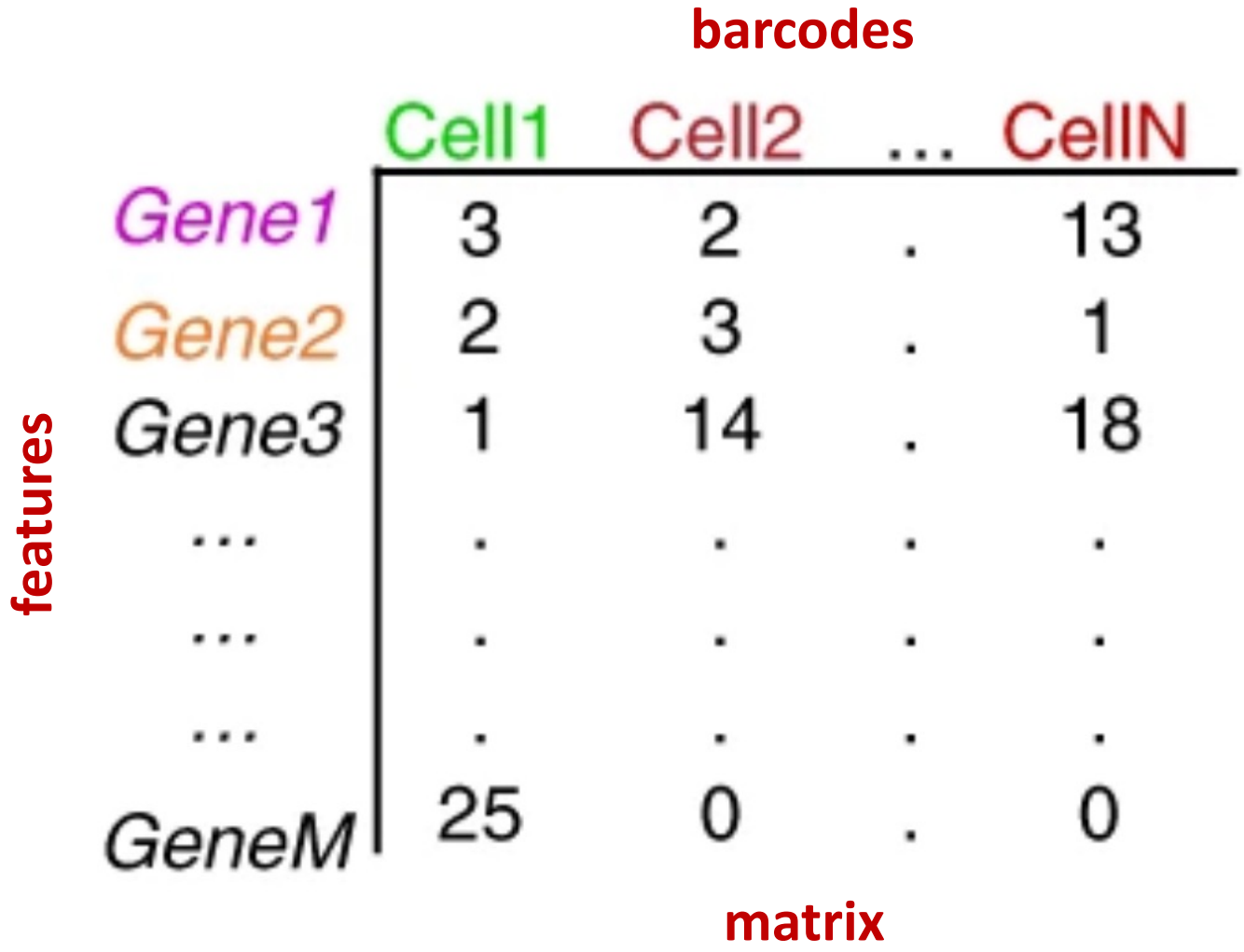


R

Processed data from 10X cellranger

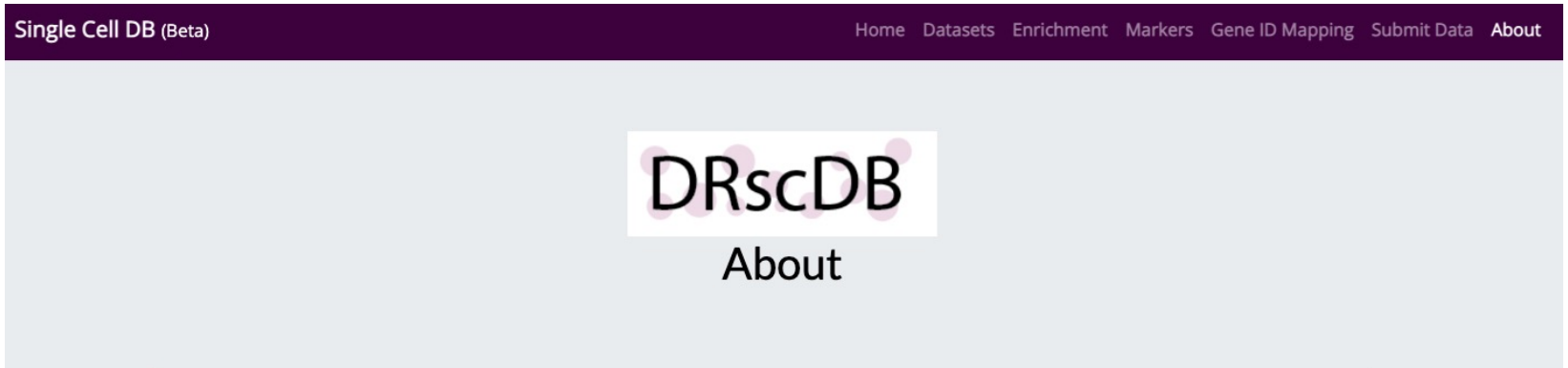
```
$ cd /home/jdoe/runs/sample345/outs
$ tree filtered_feature_bc_matrix
filtered_feature_bc_matrix
├── barcodes.tsv.gz
├── features.tsv.gz
└── matrix.mtx.gz

0 directories, 3 files
```



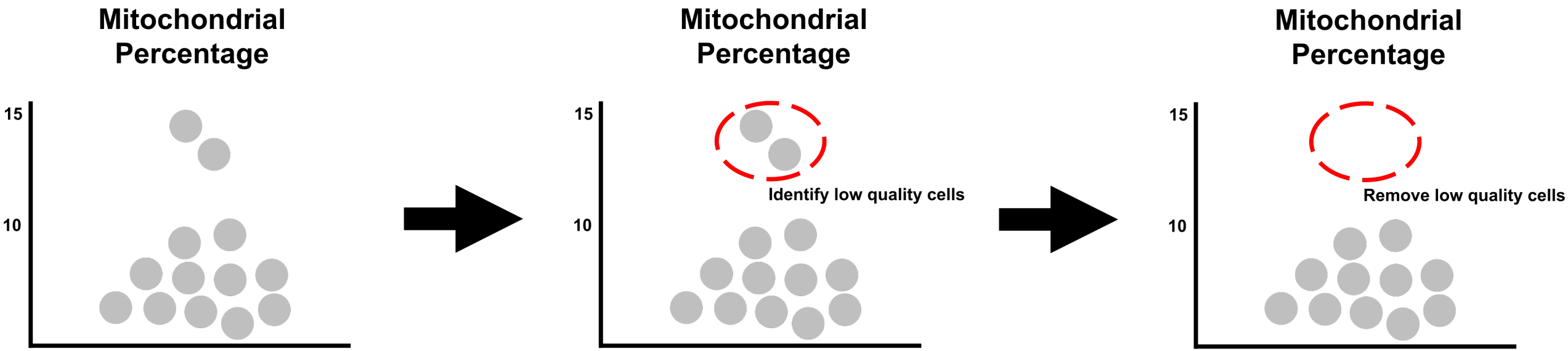
Download gene-cell matrix from database

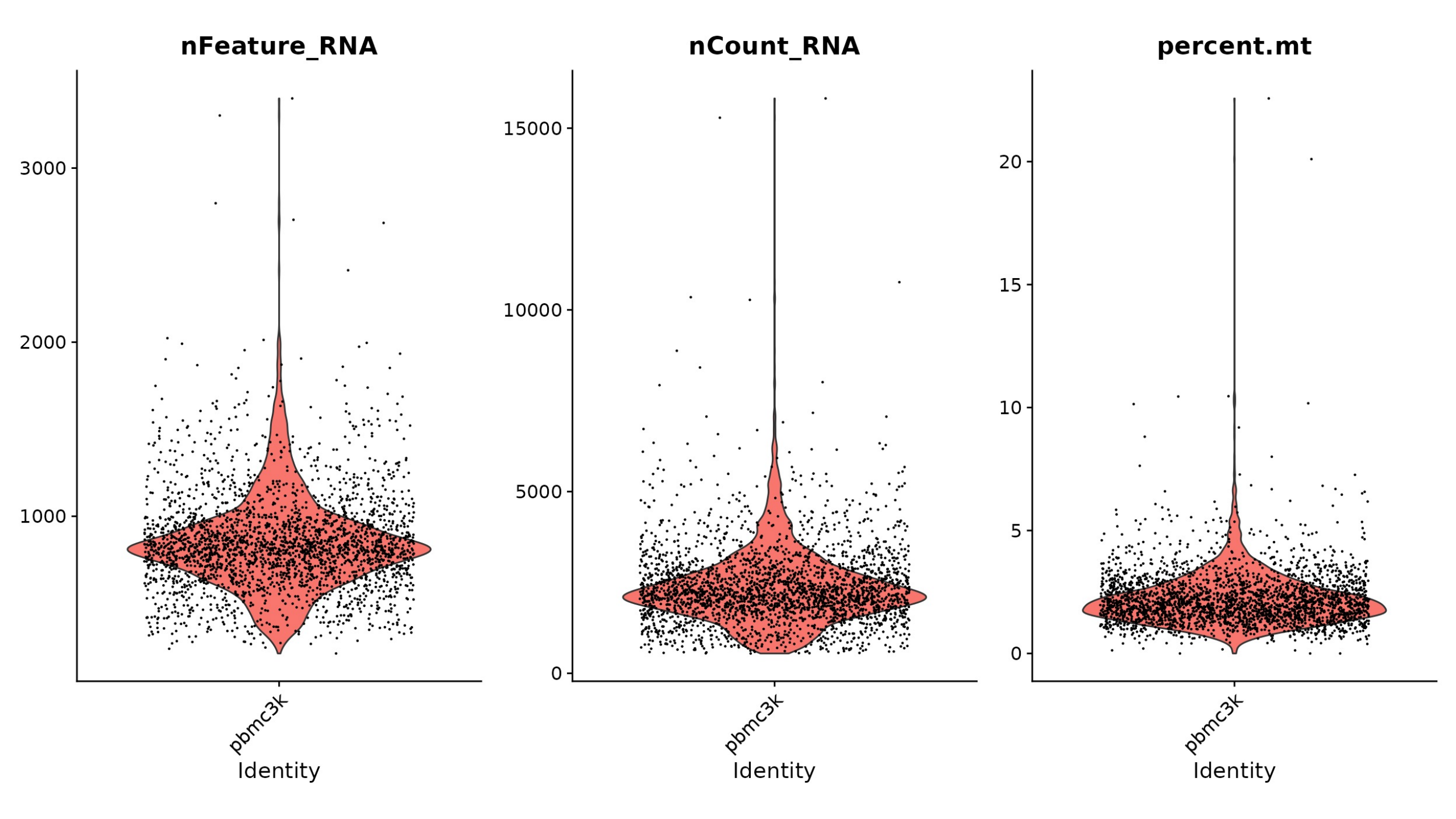
Single Cell DB

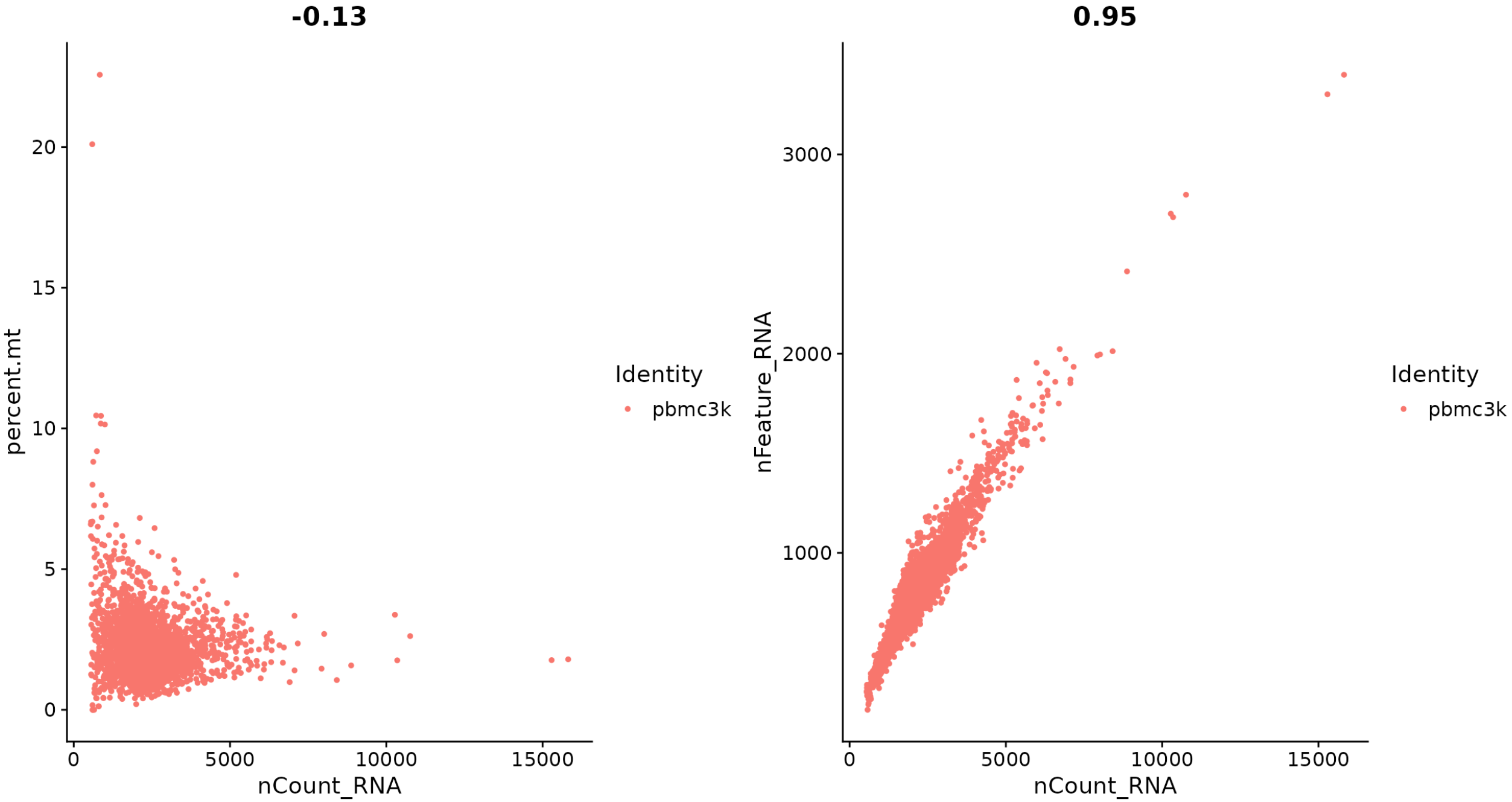


Step 2: Quality Control

QUALITY CONTROL WORKFLOW

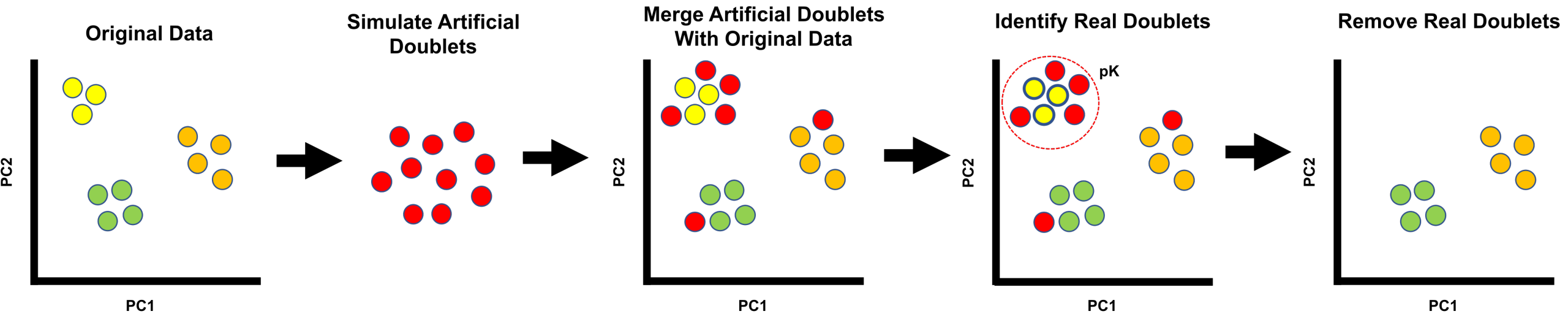



nFeature_RNA
3000
2000
1000
pbmc3k
Identity
nCount_RNA
15000
10000
5000
0
pbmc3k
Identity
percent.mt
20
15
10
5
0
pbmc3k
Identity

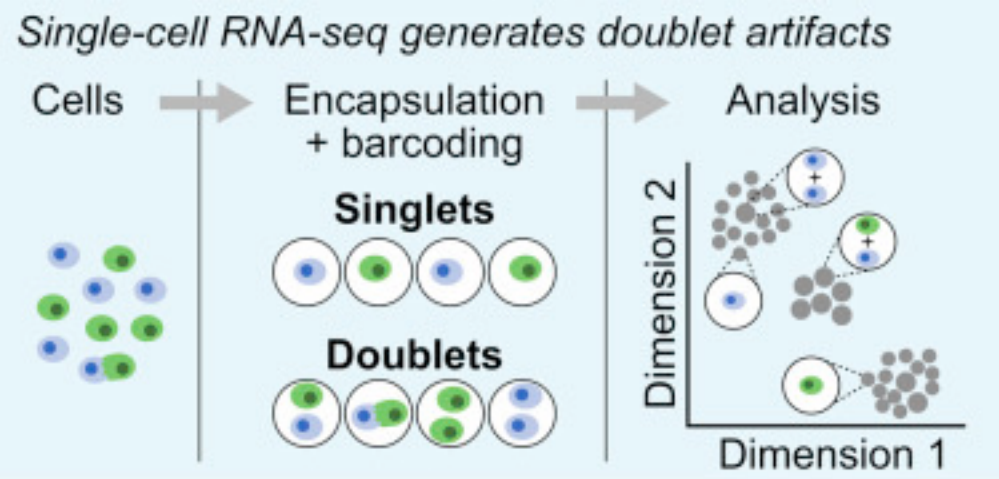

-0.13
percent.mt
20
15
10
5
0
0
5000
10000
15000
nCount_RNA
Identity
pbmc3k
0.95
nFeature_RNA
3000
2000
1000
0
5000
10000
15000
nCount_RNA
Identity
pbmc3k

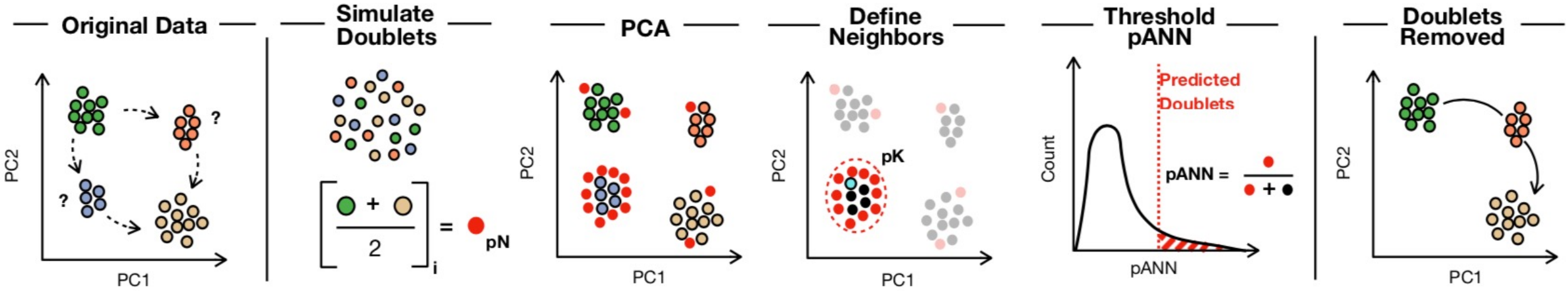
Step 3: Doublet Removal

DOUBLETFINDER WORKFLOW

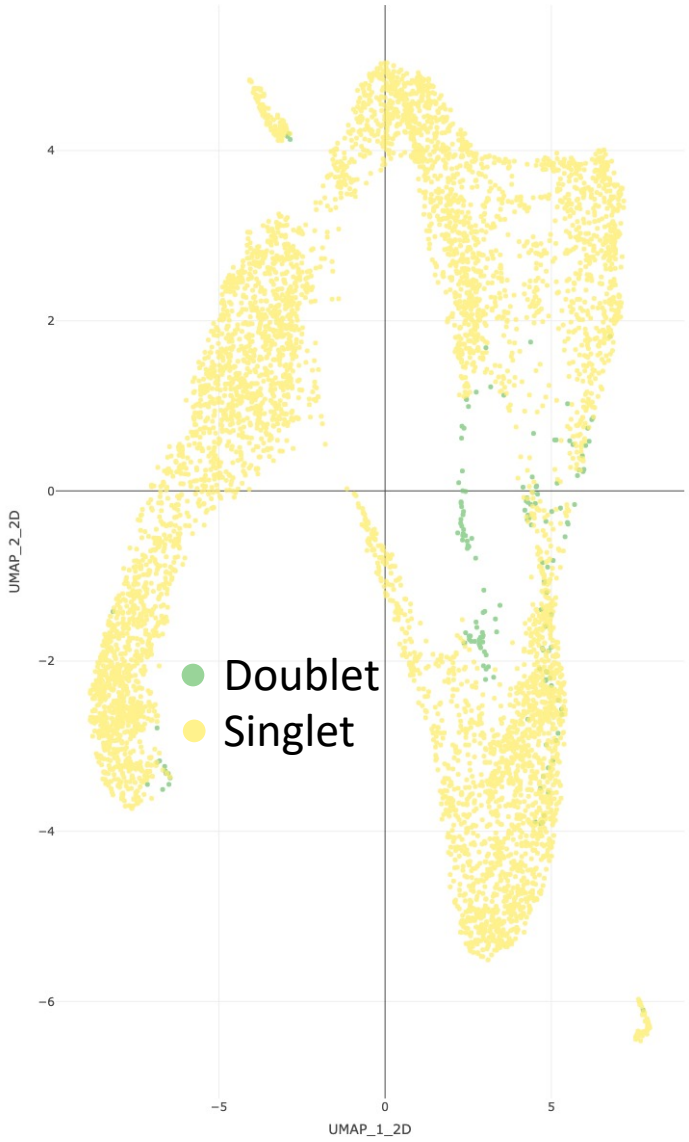


Droplet based single cell RNA sequencing

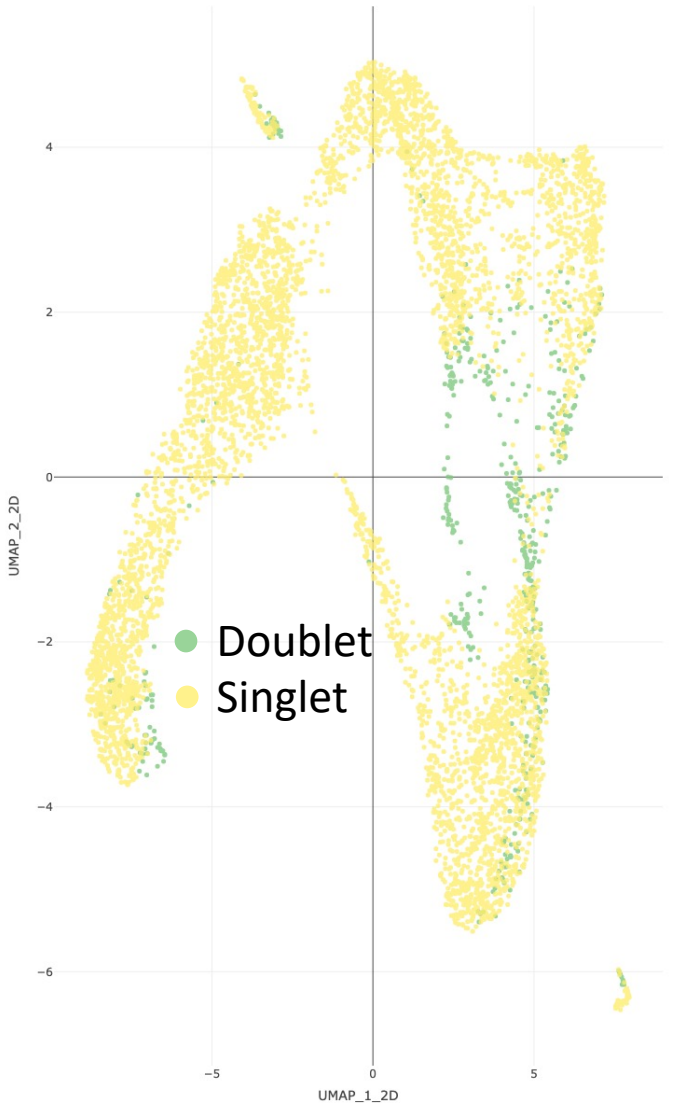




3% of doublets



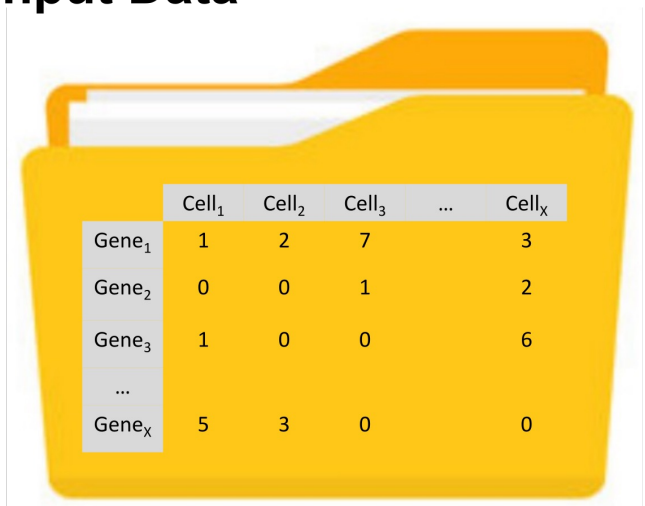
10% of doublets



Step 4: Dimensionality Reduction

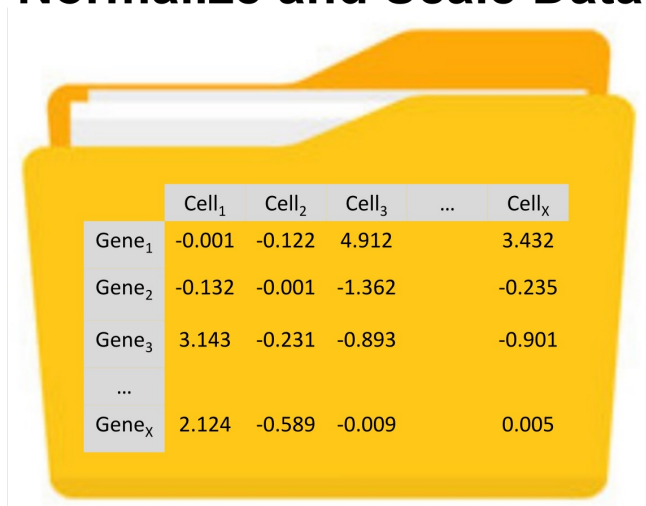
DIMENSIONALITY REDUCTION WORKFLOW

Input Data



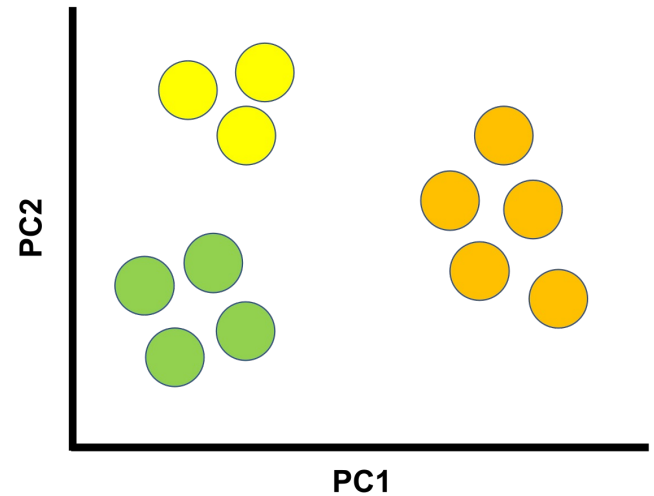
	$Cell_1$	$Cell_2$	$Cell_3$	...	$Cell_X$
$Gene_1$	1	2	7		3
$Gene_2$	0	0	1		2
$Gene_3$	1	0	0		6
...					
$Gene_X$	5	3	0		0

Normalize and Scale Data

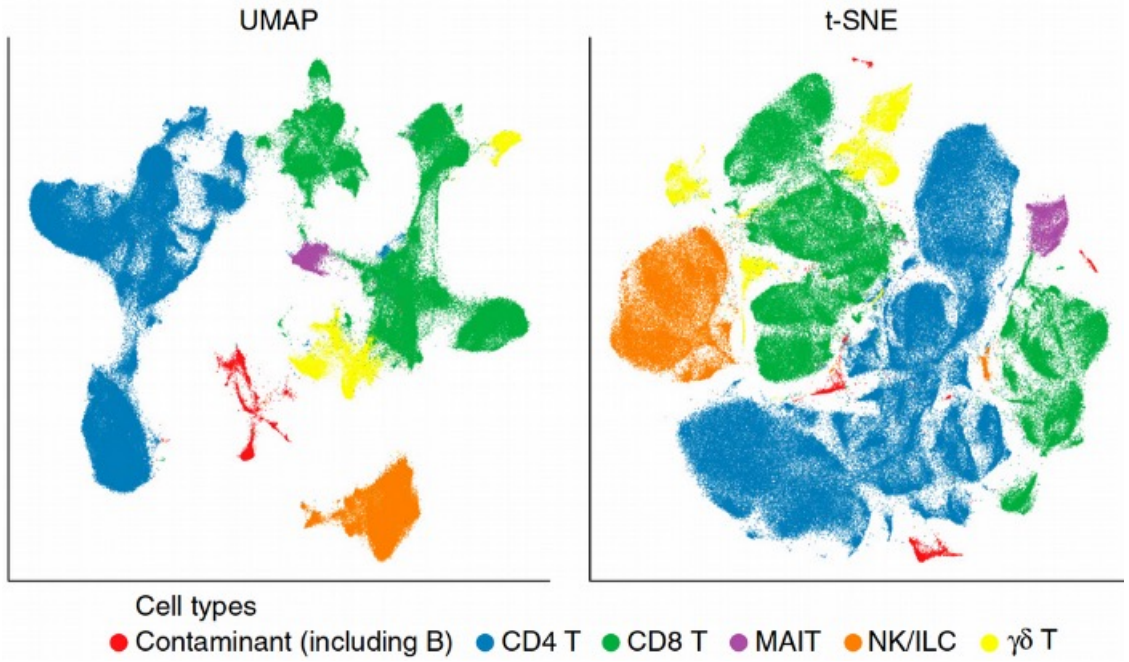


	$Cell_1$	$Cell_2$	$Cell_3$	...	$Cell_X$
$Gene_1$	-0.001	-0.122	4.912		3.432
$Gene_2$	-0.132	-0.001	-1.362		-0.235
$Gene_3$	3.143	-0.231	-0.893		-0.901
...					
$Gene_X$	2.124	-0.589	-0.009		0.005

Dimensionality Reduction



Dimensionality reduction



- PCA (principal component analysis)

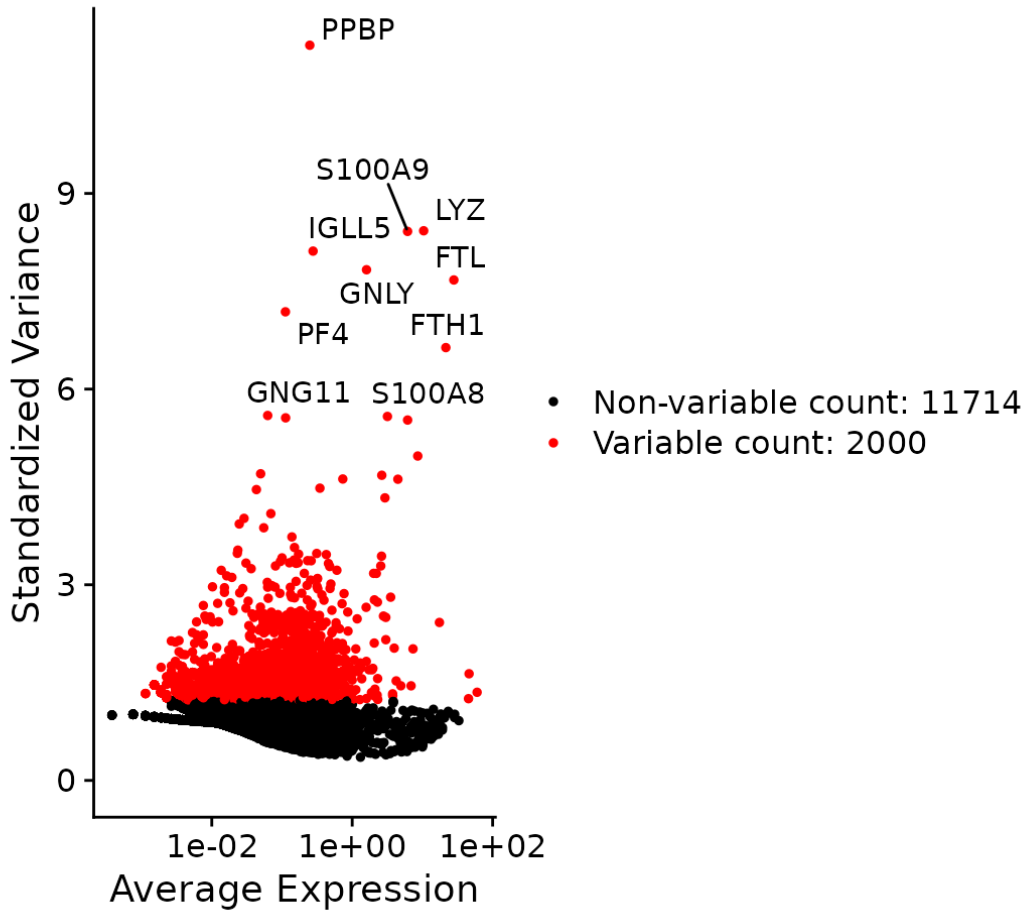
Visualization:

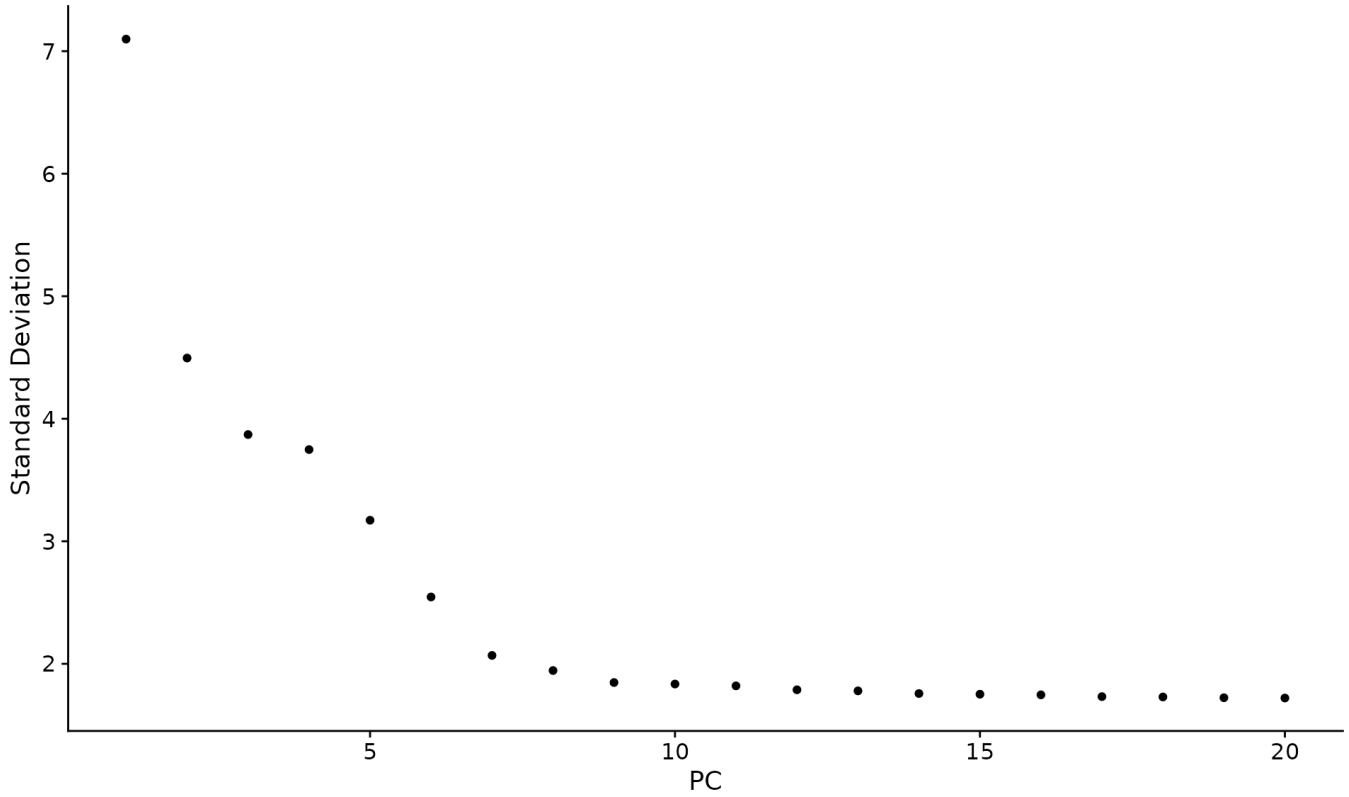
- tSNE (t-distributed Stochastic Neighbor Embedding)

  #2008 #old #slow

- UMAP (Uniform Manifold Approximation and Projection)

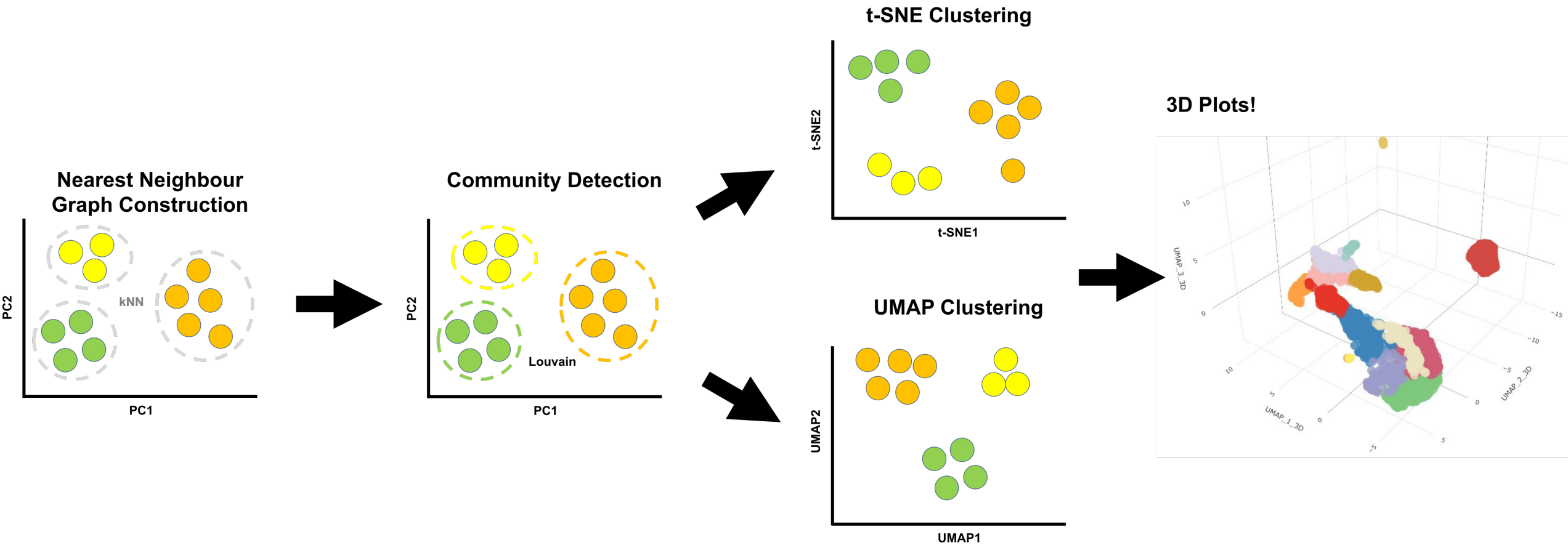
  #2018 #new #fast

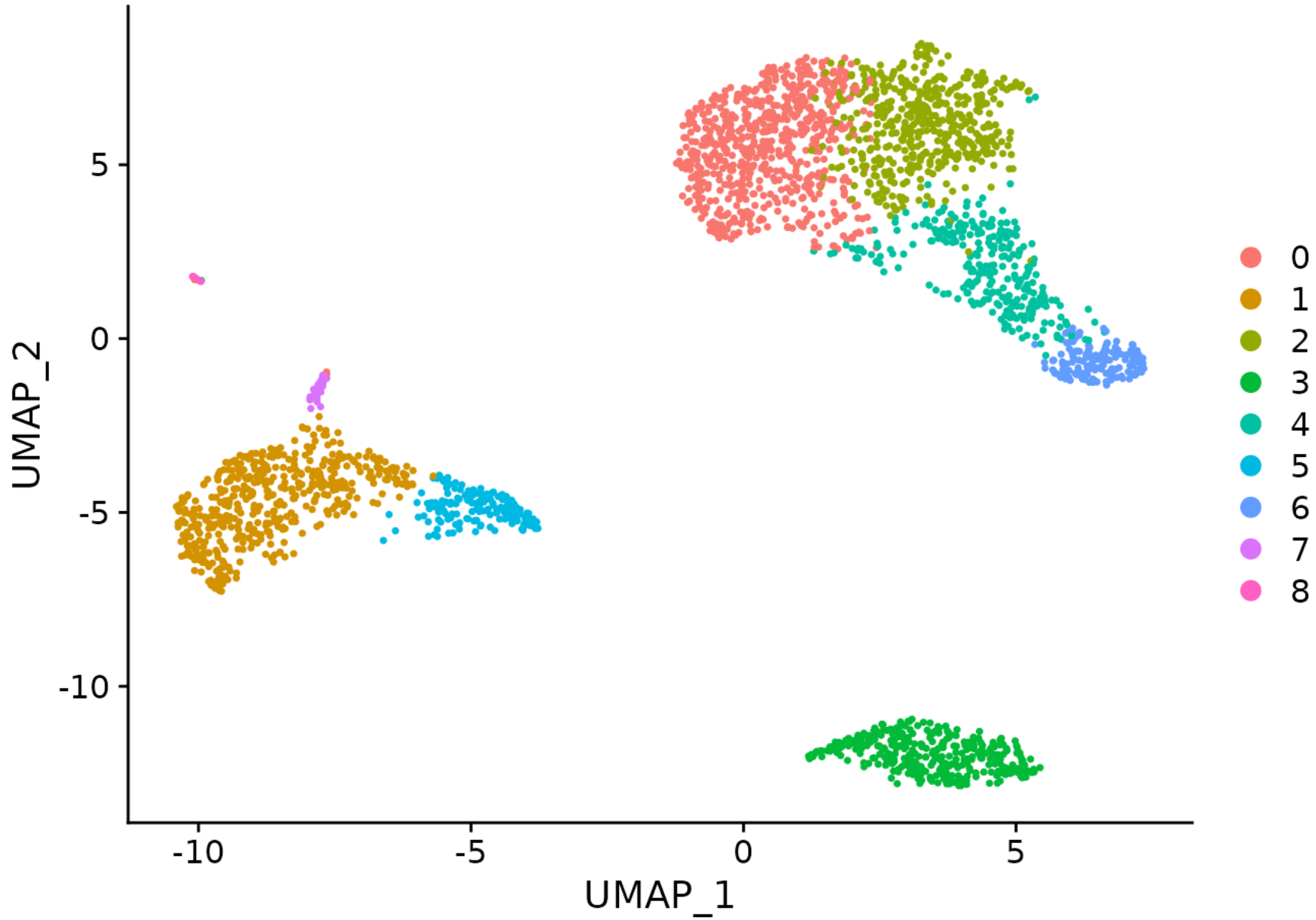

PPBP
S100A9
LYZ
IGLL5
FTL
GNLY
FTH1
PF4
GNG11
S100A8
Standardized Variance
0
3
6
9
1e-02
1e+00
1e+02
Average Expression
Non-variable count: 11714
Variable count: 2000


Standard Deviation
2
3
4
5
6
7
5
10
15
20
PC

Step 5: Clustering

CLUSTERING WORKFLOW



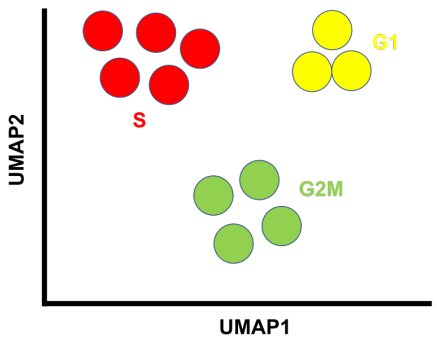

5
0
-5
-10
UMAP_2
-10
-5
0
5
UMAP_1
0
1
2
3
4
5
6
7
8

Step 6: Data Correction

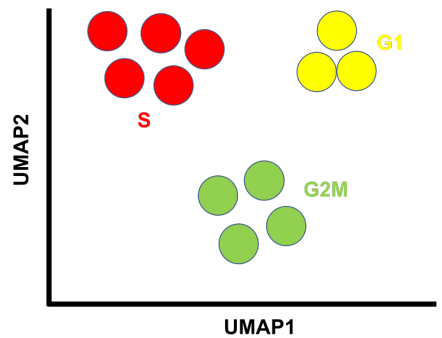
DATA CORRECTION WORKFLOW

Cell Cycle Effect

Cell Cycle Scoring

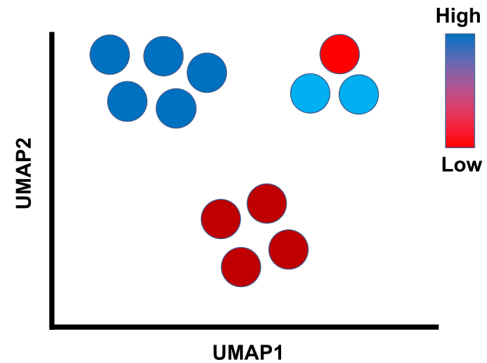


Regress Cell Cycle Effect

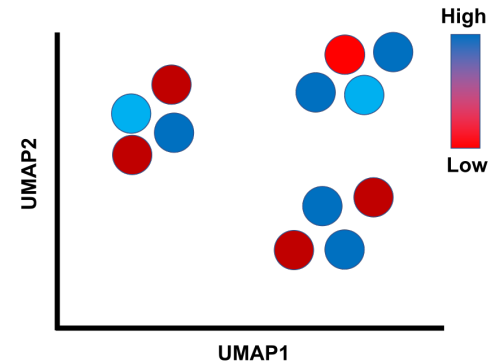


Gene Effect

Gene Expression



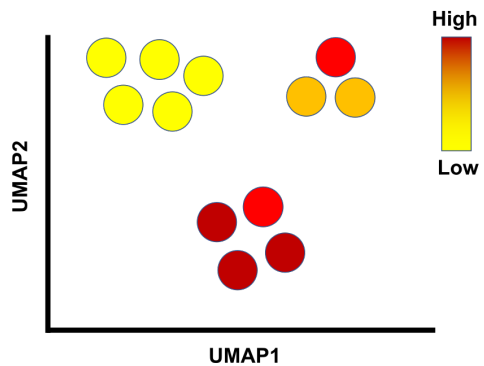
Regress Gene Effect



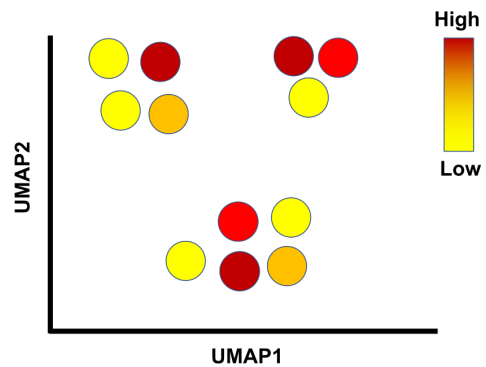
Gene Pathway Effect

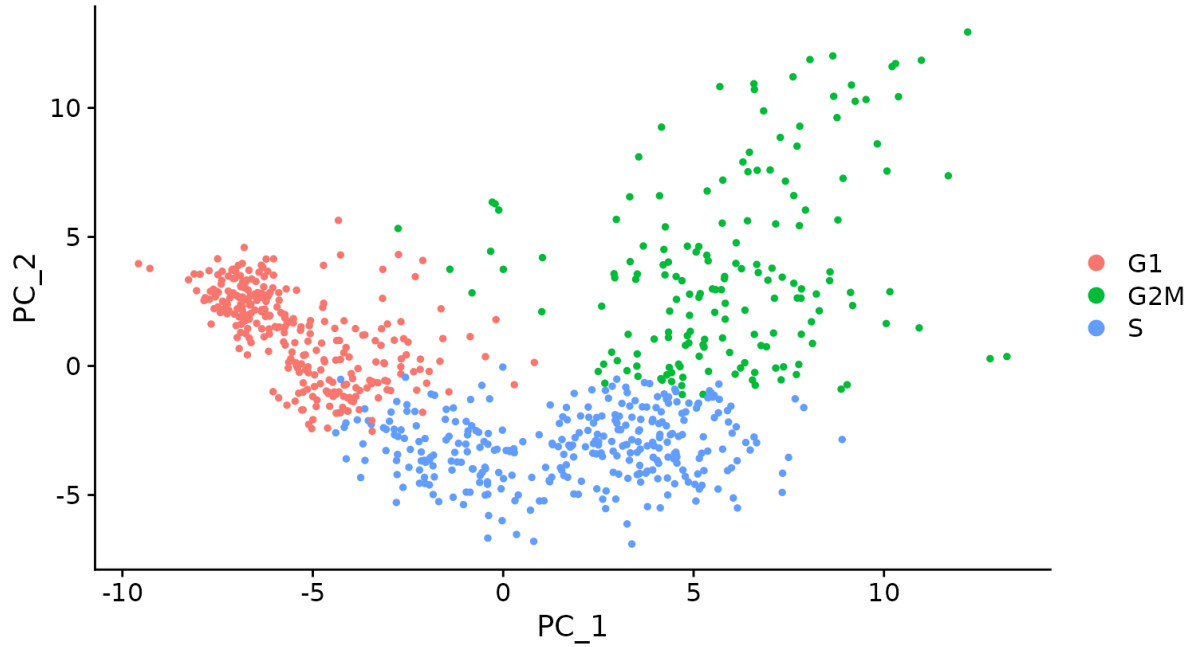
MSigDB Molecular Signatures Database

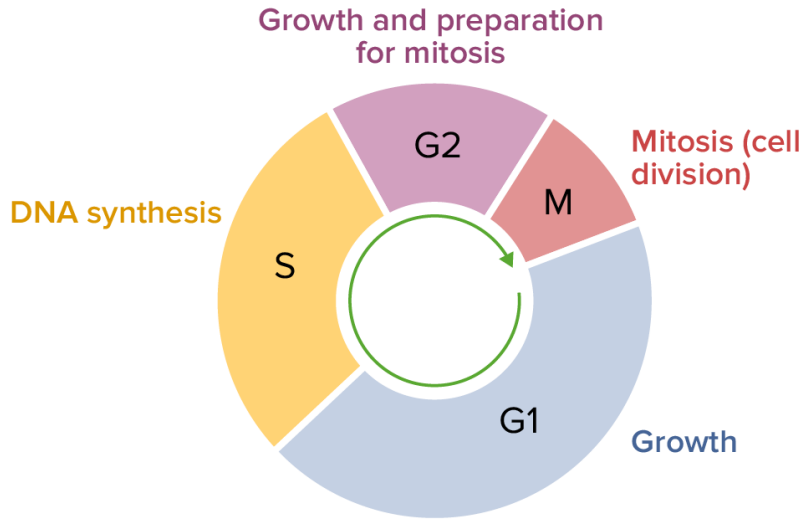
Gene Pathway Expression

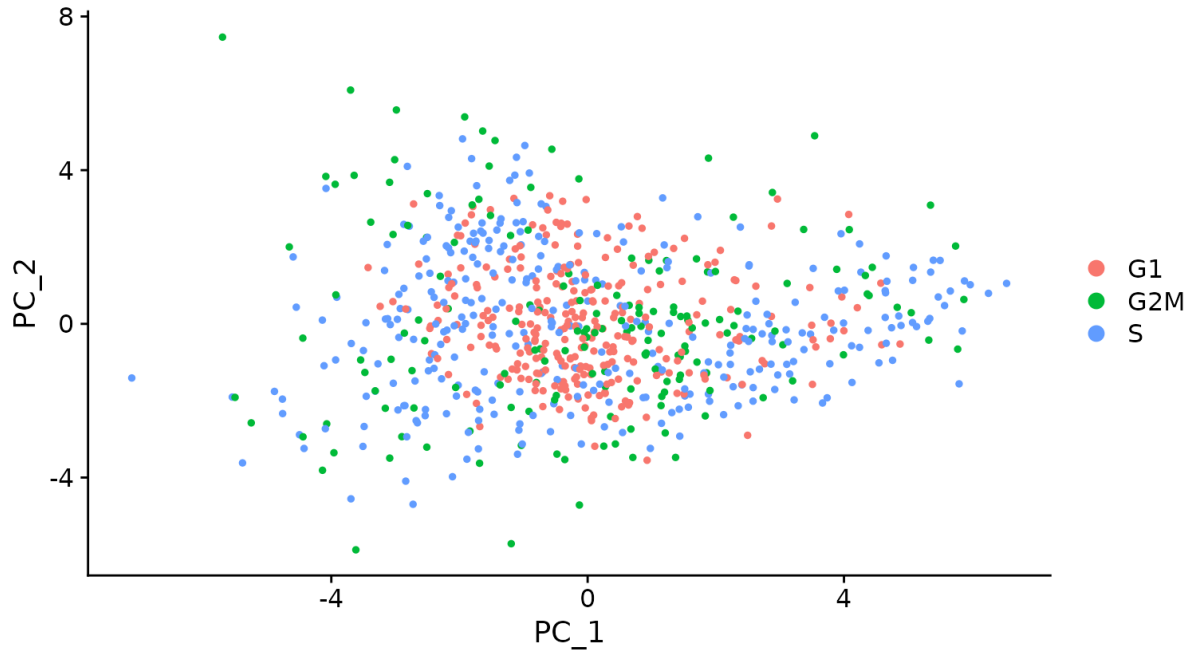


Regress Gene Pathway Effect



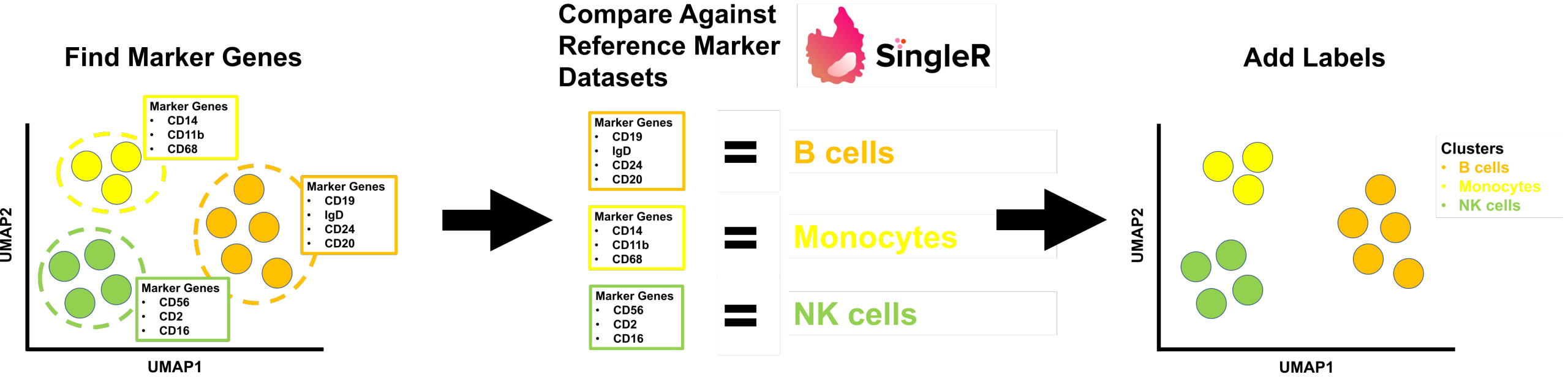

PC_2
PC_1
10
5
0
-5
-10
-5
0
5
10
G1
G2M
S

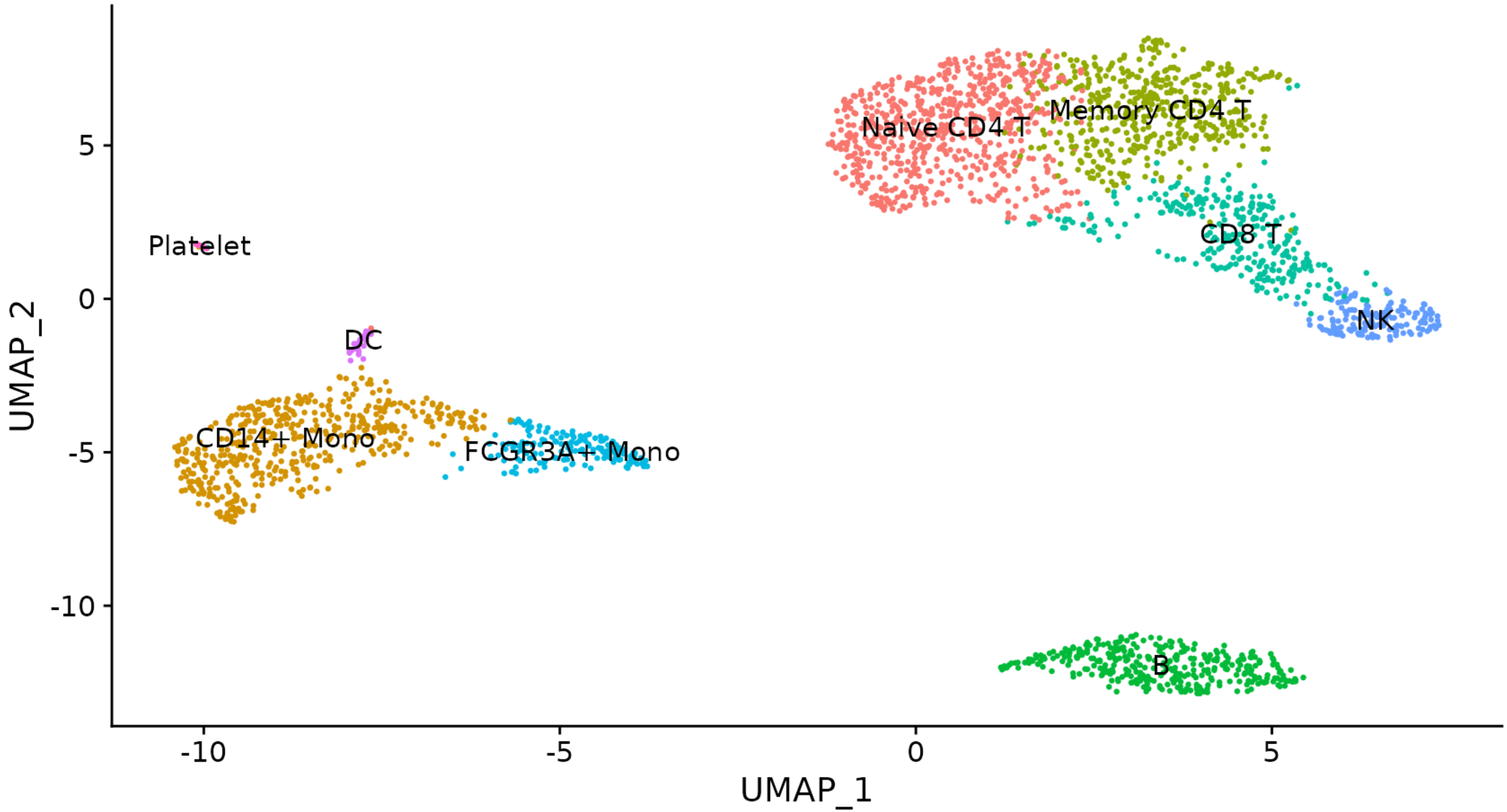

Growth and preparation for mitosis
G2
Mitosis (cell division)
M
DNA synthesis
S
G1
Growth


PC_2
PC_1
8
4
0
-4
-4
0
4
G1
G2M
S

Step 7: Labelling Clusters

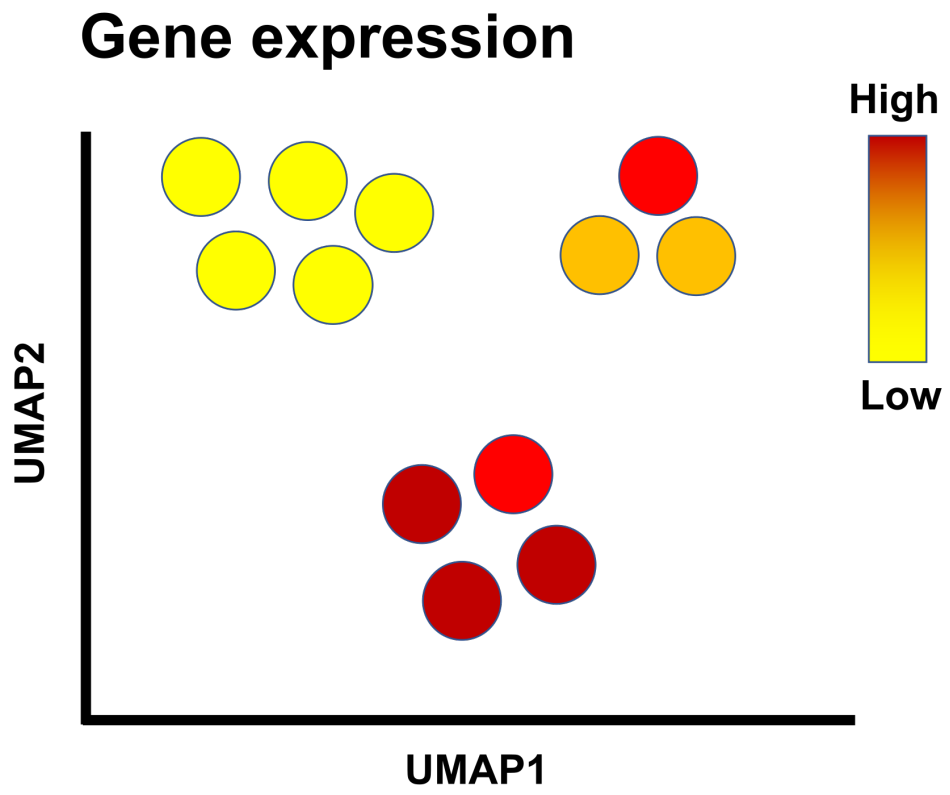
LABELLING WORKFLOW




Naive CD4 T
Memory CD4 T
CD8 T
NK
Platelet
DC
CD14+ Mono
FCGR3A+ Mono
B
UMAP_2
UMAP_1
5
0
-5
-10
-10
-5
0
5

Step 8: Gene Expression

GENE EXPRESSION VISUALISATION



Visualise Gene Pathways From:

MSigDB Molecular Signatures Database

WikiPathways Pathways for the People

GeneOntology Unifying Biology

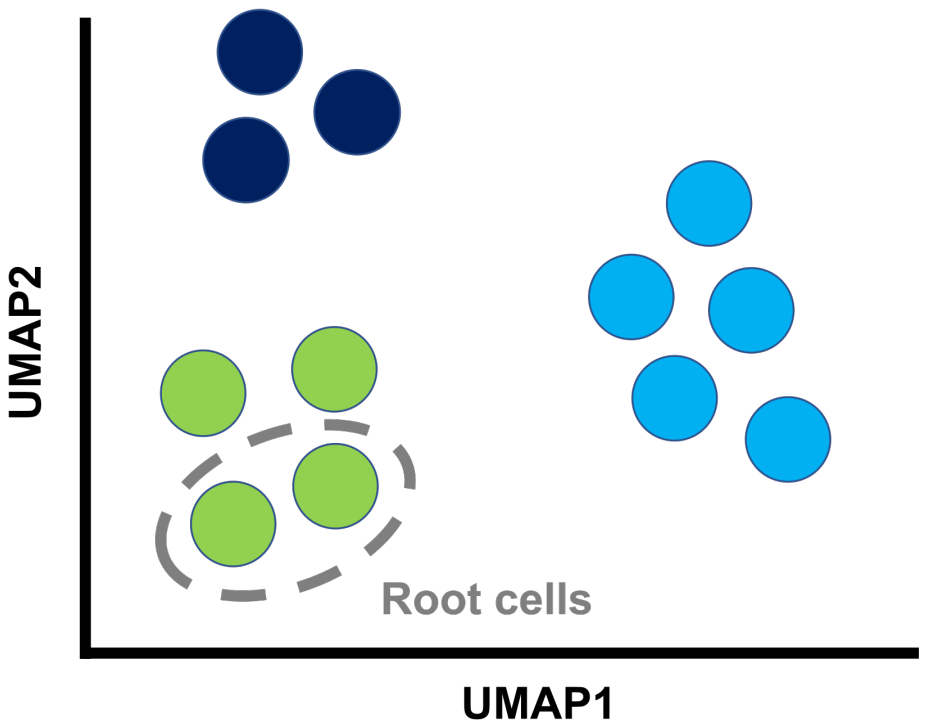
KEGG Kyoto Encyclopedia of Genes and Genomes

reactome

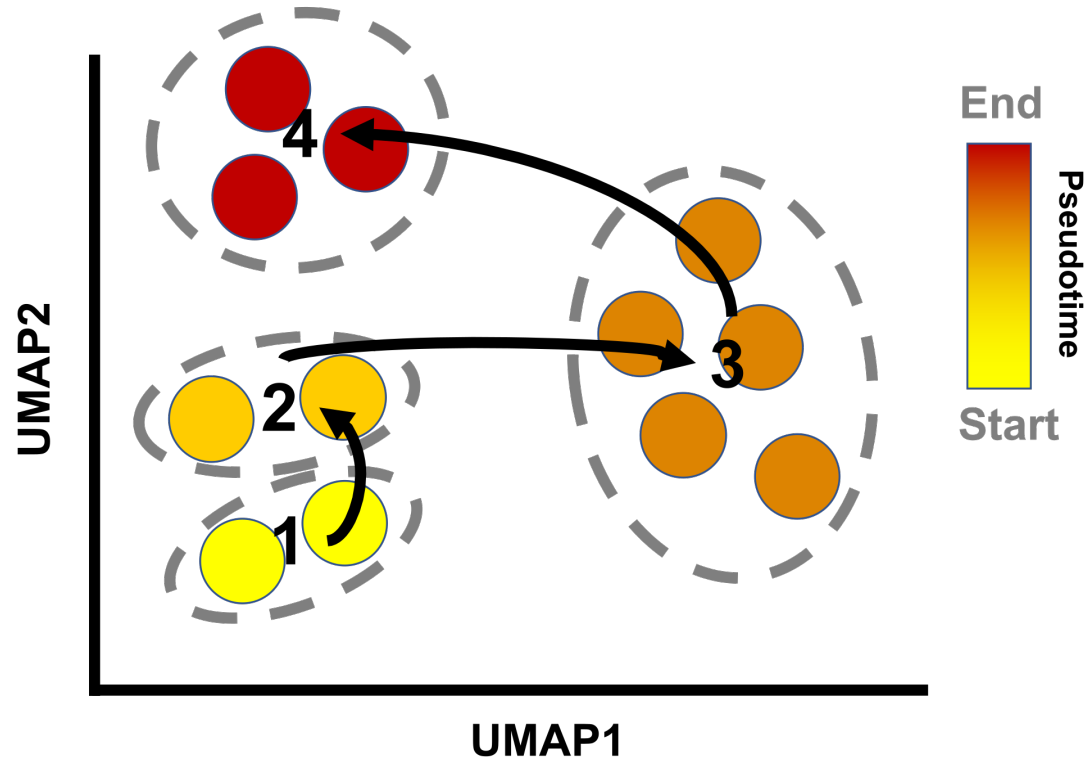
Step 9: Trajectory Analysis

TRAJECTORY ANALYSIS WORKFLOW

Assign Root Cells



Trajectory Analysis



Step 10: Multimodal Analysis

DATA INPUT (MULTIMODAL ANALYSIS)

A

Multimodal data

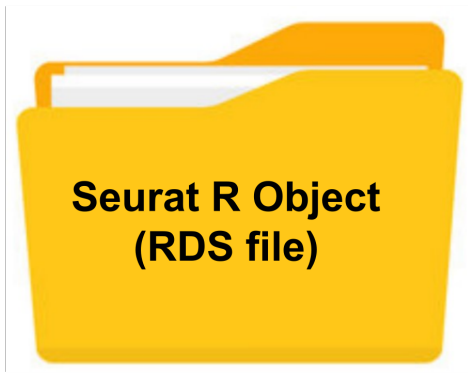
	$Cell_1$	$Cell_2$	$Cell_3$	...	$Cell_X$
$Gene_1$	1	2	7		
$Gene_2$	0	0	1		
$Gene_3$	1	0	0		6
...					
$Gene_X$	5	3	0		0

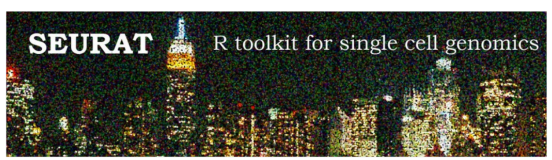
Column names must match those in the scRNA-seq data

B

10x GENOMICS®

C



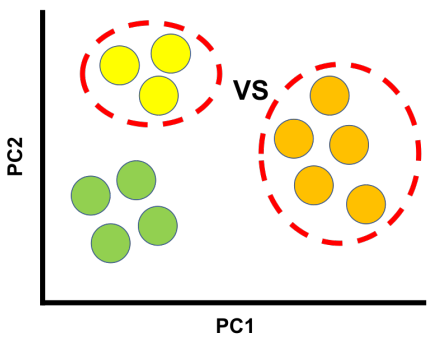


Make sure multimodal data is saved in "ADT" assay

Step 11: Differential Expression & Gene Set Enrichment Analysis

DIFFERENTIAL EXPRESSION AND GENE SET ENRICHMENT ANALYSIS WORKFLOW

Select Cells to Compare



List of Differentially Expressed Genes

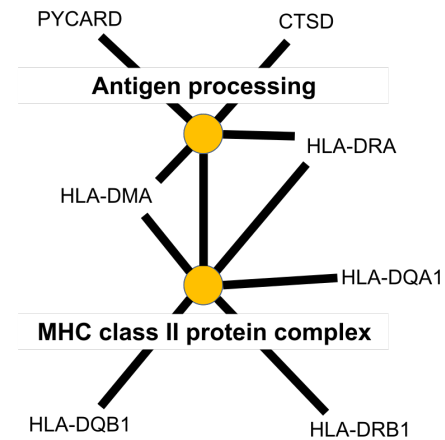
Differentially expressed genes	Fold change	P-value
IL7R	-4.150	0.002
CYBB	-4.037	0.004
TRAC	3.563	0.02
IL32	5	0.06

Gene Set Enrichment Analysis

clusterProfiler

ID	Description	P-value
GO:0002181	cytoplasmic translation	0.001
GO:0002399	MHC class II protein complex assembly	0.009
GO:0006364	rRNA processing	0.01
GO:0006518	peptide metabolic process	0.03

Pathway Analysis Visualisations




Tingki
Komapsumnida
Shukuria
atu
Paldies
Tashakkur
gozaimashita
bolzin
Maake
Ekhmet
Mehrbani
Arigato
You
Dankscheen
Yaqhanyelay
Efcharisto
suksama
Merci
Gracias
Shukria
Thank
Dhanyabaad
Merastawhy
Chaltu
Bïyan
Grazie
Juspaxar
謝謝