SCODA pipeline

for single-cell transcriptomics data analysis

MLBI Lab

SCODA?

❑ Web-based, Fully-automated, all-in-one computing service for **S**ingle-**C**ell (transcript)**O**mics **D**ata **A**nalysis (single-cell RNA-seq)

❑ It is useful for
  ❍ Tissue/tumor microenvironment study to elucidate underlying biological mechanism
  ❍ Discovery of therapeutic targets in many diseases & cancers

How SCODA works?

Single-cell RNA-seq raw data

Upload raw data (Count matrix)

Web-based, automated analysis pipeline: SCODA

Functional Analysis

- Cell-type annotation
- CNV estimation
- Identifying tumor cells
- Cell-cell interaction
- Diff. expressed gene
- Gene set analysis
- Marker discovery

mlbi-lab.net
mlbi-lab.com

Download result file with functional annotations

Client's own data mining using SCODA-viz package

Jupyter Notebook for client's own data mining
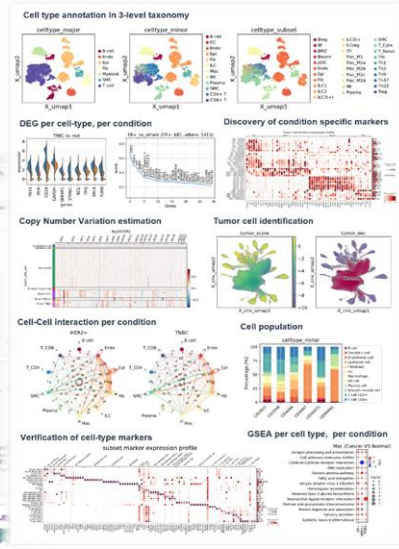
- Therapeutic target discovery
- Biological mechanism
- Design validation experiment

How SCODA works?

Client

Client's data

Upload

← SCODA Server →

SCODA pipeline

Results
Saved in anndata

Download

Client

Jupyter notebook for your own Data Mining

---

(1) Simply upload your data

1. Compressed 10x mtx files + optional meta_data.csv
2. Compressed CSV file + optional meta_data.csv
3. Compressed h5ad file (including meta data in obs field)

---

(2) Automated pipeline analyzes your data

1. Celltype annotation (HiCAT, MIT license)

2. DEG & GSA/GSEA (GSEApy, MIT license)

3. Cell-cell interaction (CellPhoneDB, MIT license)

4. Tumor cell identification (InferCNV, BSD-3c + InferPloidy, MIT)

❖ We use proven & standardized analysis tools
❖ The pipeline can be tailored with optional settings

---

(3) Once it is done, you can download the result

1. Saved in Anndata formatted h5ad file (tar.gz compressed)
2. It contains all the analysis results and the count matrix itself

---

(4) Explore the results by yourself on your web browser

1. We provide "jupyter notebook" with example codes
2. You can freely use SCODA-viz & SCANpy (free open SW)
3. With a little bit of programming, you can get deeper insight into the single-cell RNA-seq data you uploaded
4. We offer free workshop for data mining (upon request)

Visualization & data mining

Why SCODA?

Bio/Medical background

+

SCODA
with SCODA-viz tool

+

Little bit of programming skill
(Free Training workshop available)

||

Single-cell RNA-seq
Data analysis expert

Why SCODA?
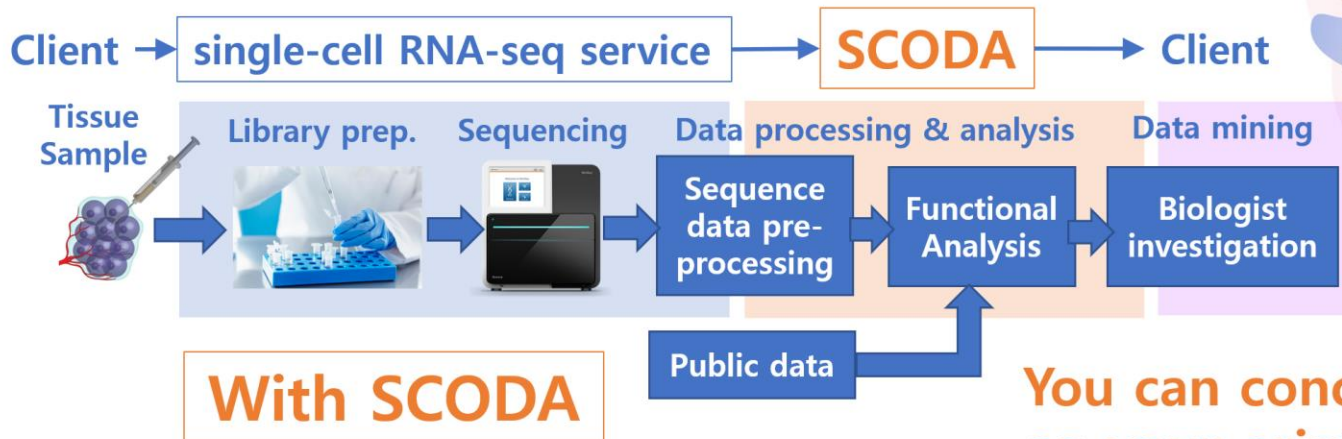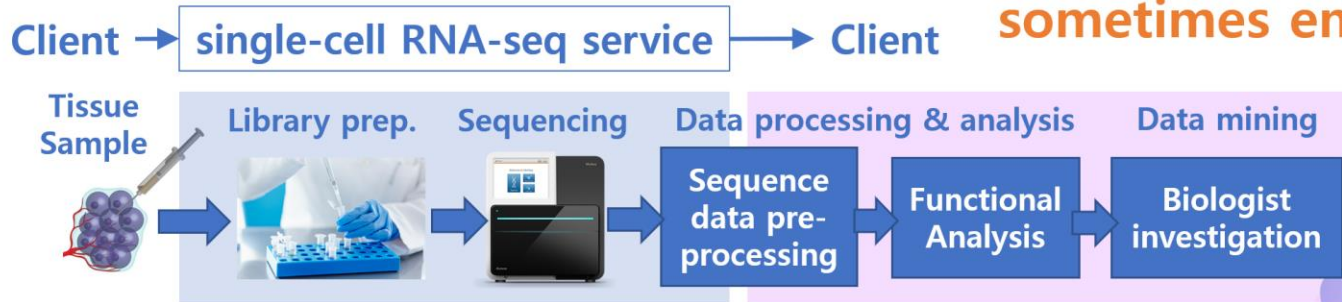
Related papers

❑ **HiCAT**

  ❍ Hierarchical cell-type identifier accurately distinguishes immune-cell subtypes enabling precise profiling of tissue microenvironment with single-cell RNA-sequencing, **Briefings in bioinformatics, March 2023**

❑ **Studies using SCODA**

  ❍ Colon-Targeted eNAMPT-Specific Peptide Systems for Treatment of DSS-Induced Acute and Chronic Colitis in Mouse, **Antioxidants, Nov. 2022**

  ❍ Integrative analysis of ulcerative colitis progression using single-cell RNA-seq and microbiome, **Communications Biology, June 2024**

  ❍ A Retrospective View on Triple Negative Breast Cancer Microenvironment: Novel Markers, Interactions, and Mechanisms of Tumor-Associated Components using public Single-cell RNA Seq Datasets, **Cancers, May 2024**

SCODA Optional configuration

Mandatory input

Cell-by-Gene count matrix

Upload

Optional input

Analysis_config.py

Optional input

markers_db.tsv

SCODA server

Cell-type annotation (HiCAT)

Tumor cell identification (InferCNVpy + addon)

cellphonedb.zip

Cell-cell interaction (CellPhoneDB)

pathways.gmt

DEG/GSA/GSEA (GSEApy)

- N_GENES_MIN = 200
- N_CELLS_MIN = 10
- PCT_COUNT_MT_MAX = 20
- REFERENCE_CELLTYPES_FOR_TUMOR_ID = ['T cell', 'B cell', 'Myeloid', 'Fibroblast']
- MIN_NUM_CELLS_FOR_CCI = 40
- CCI_MIN_OCC_FREQ_TO_COUNT = 0.5
- CCI_PVAL_CUTOFF = 0.1
- CCI_MEAN_CUTOFF = 0.1
- CCI_DEG_BASE = 'celltype_minor'
- MIN_NUM_CELLS_FOR_DEG = 100
- DEG_PVAL_CUTOFF_FOR_GSEA = 0.1
- DEG_REF_GROUP = None
- GSA_PVAL_CUTOFF = 0.1

Result h5ad file (compressed)

Download

Visualization & Data mining

SCODA?

❑ **SCODA is useful for**

  ○ In-vivo tissue/tumor micro-environment (TME) study.

  ○ Immune cell profiling in many diseases, e.g., autoimmune disease & cancer

  ○ Discovery of diagnostic/prognostic markers

  ○ Discovery of druggable targets and its biological mechanism around pathological tissue

  ○ Exploring drug response and mechanism of action

❑ **But not suitable yet for**

  ○ Studies with cell line

  ○ Differentiation study

Summary

❑ **SCODA utilizes proven open-source software**

- ❍ **HiCAT (MIT license)** for cell type annotation
- ❍ **CellPhoneDB (MIT license)** for inferring cell-cell interaction
- ❍ **InferCNVpy (BSC-3clause)** for CNV estimation
- ❍ **GSEApy (BSC-3clause)** for gene set enrichment analysis

❑ **SCODA-viz package and example jupyter notebook freely available for visualization and data mining**

- ❍ With a little bit of programming skill, you can create any kind of plots you want. (Free training workshop available upon request)

❑ **It accelerate your research with single-cell RNA-seq experiment, saving your time and the cost.**

- ❍ Use SCODA first to get insight into the tissue of your interest.
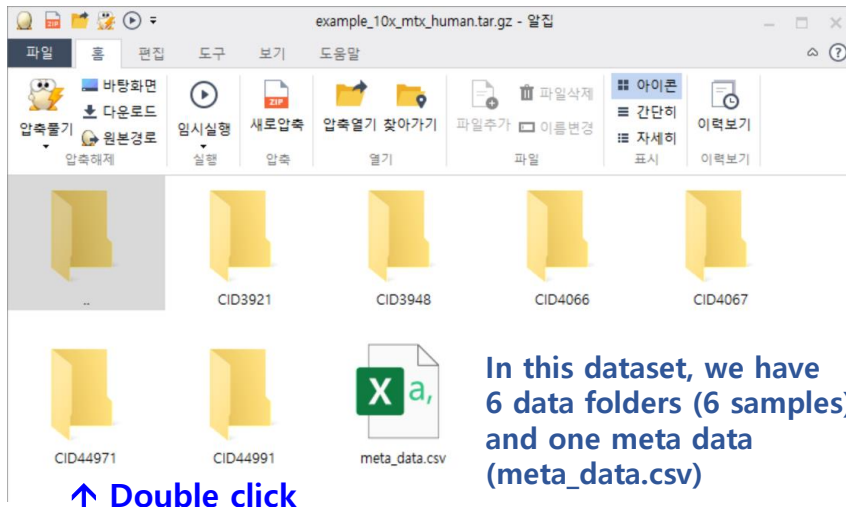- ❍ Then, plan biological experiment to verify your hypothesis.

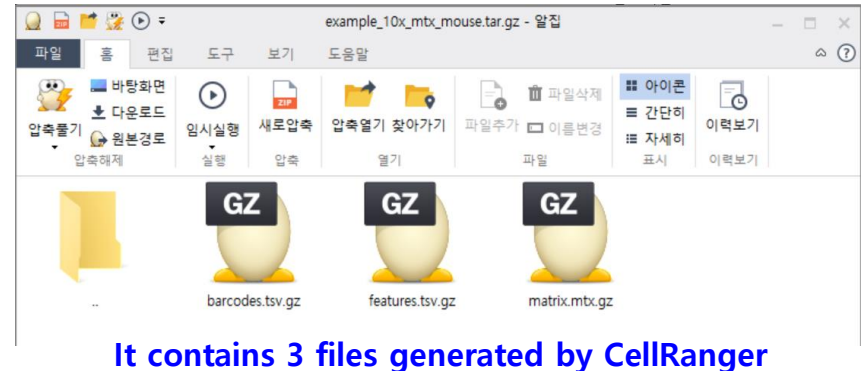# Thank you

Input data formatting (1) 10x_mtx

1. Contents of the compressed input file (.zip or .gz)

← **Main folder Double click**

2. Opening the main folder, we see the following

↑ **Double click**

In this dataset, we have 6 data folders (6 samples) and one meta data (meta_data.csv)

3. Further opening each data folder, we see

It contains 3 files generated by CellRanger

4. Opening the meta_data.csv

A: Index column
B: sample name
C: condition

- Each index corresponds to one of the folders (name).
- But we don't require sample name must correspond to a folder name.
- Condition is required to perform DEG, GSEA, cell-cell interaction to compare difference among conditions

Input data formatting (2) csv format

1. Contents of the compressed input file (.zip or .gz)

← **Main folder Double click**

GSE176078
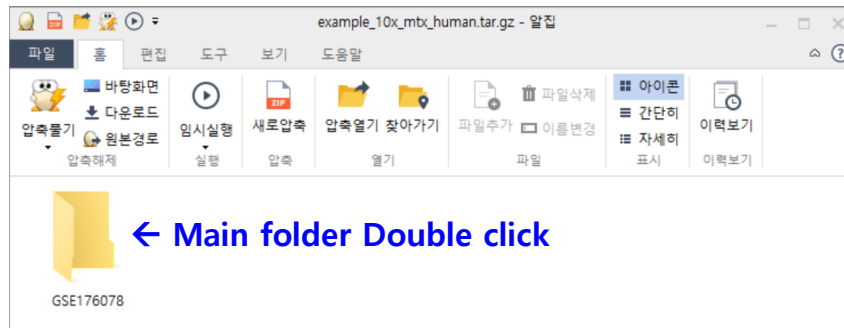
2. Opening the main folder, we see the following
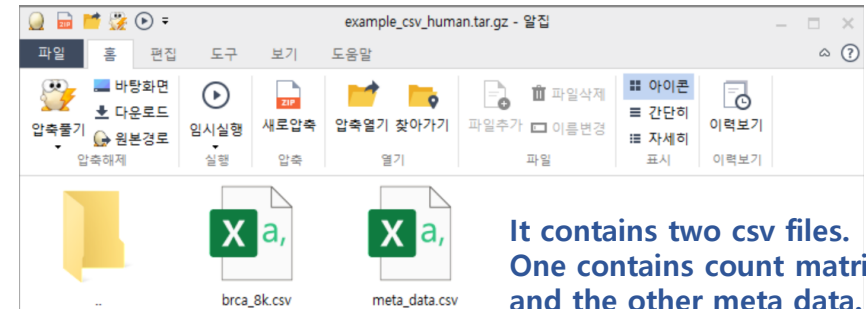
It contains two csv files. One contains count matrix and the other meta data. (The latter must be named "meta_data.csv"

3. Data csv file (containing count matrix)

Hugo symbol

Count matrix

**Cell barcode (cell ID)**

4. meta_data.csv 파일

Cell barcode (cell ID)

**A: Index column
B: sample
C: condition
D~: optional items**

- **The indices of the data matrix must have 1-to-1 correspondence to those of meta data matrix.**

- **Condition is required to perform DEG, GSEA, cell-cell interaction to compare difference among conditions**

Input data formatting (3) h5ad format

1. Contents of the compressed input file (.zip or .gz)

← h5ad file (anndata format)

https://anndata.readthedocs.io/en/latest/
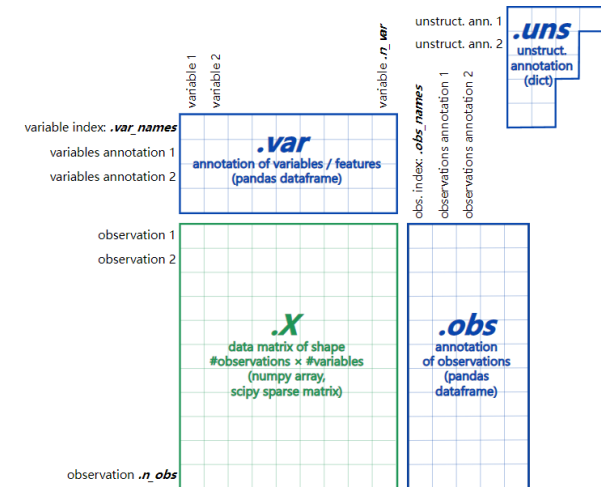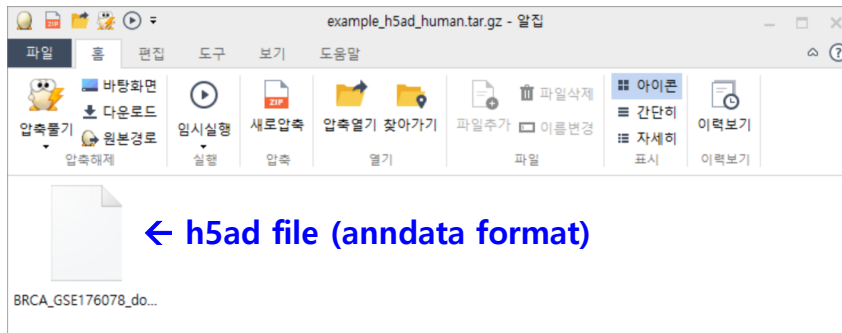
2. h5ad file contents

```
adata_t = sc.read_h5ad(file_h5ad)
adata_t
```

```
AnnData object with n_obs × n_vars = 12000 × 29733
    obs: 'Patient', 'Percent_mito', 'nCount_RNA', 'nFeature_RNA', 'Celltype_Major', 'Celltype_Minor', 'Celltype_Subset',
'subtype', 'gene_module', 'Calls', 'normal_cell_call', 'CNA_value', 'sample', 'condition'
    var: 'gene_ids'
```

- AnnData contains "sample" and "condition" columns to run DEG/GSEA. DEG/GSEA will not be performed if the obs field does not contain both "sample" and "condition" column.
- If the "sample" column exists in the obs field, cell-cell interaction will be performed per-sample the same as in the above.