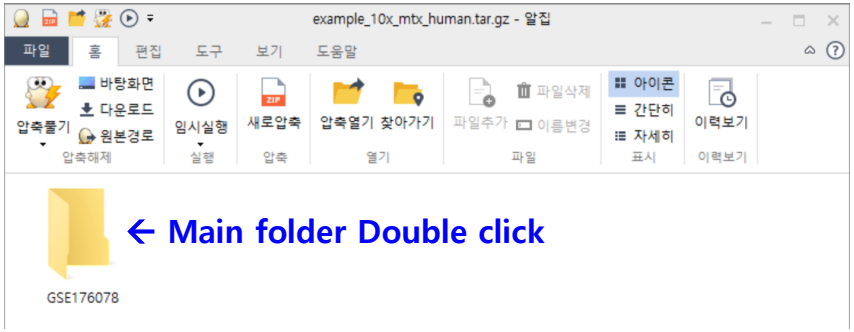
SCODA pipeline

Input data formatting

MLBI Lab

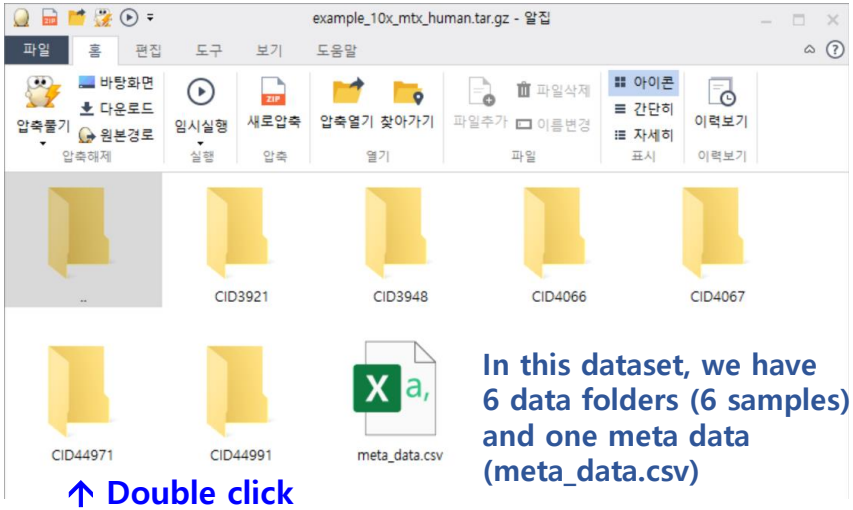
Input data formatting (1) 10x_mtx

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



← Main folder Double click

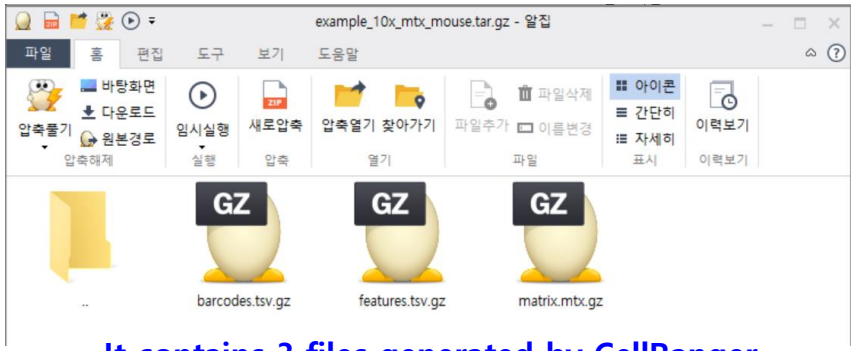
2. Opening the main folder, we see the following



↑ Double click

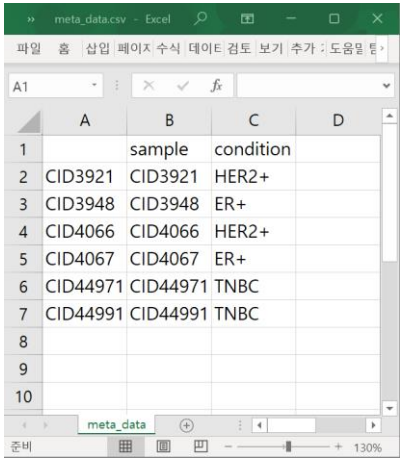
In this dataset, we have 6 data folders (6 samples) and one meta data (meta_data.csv)

3. Further opening each data folder, we see



It contains 3 files generated by CellRanger

4. Opening the meta_data.csv



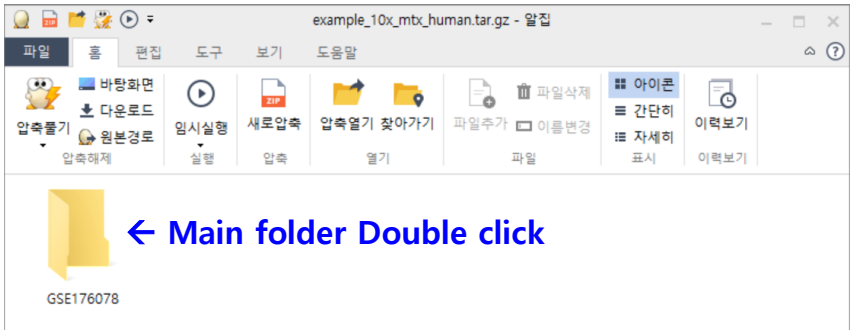
	A	B	C	D
1		sample	condition	
2	CID3921	CID3921	HER2+	
3	CID3948	CID3948	ER+	
4	CID4066	CID4066	HER2+	
5	CID4067	CID4067	ER+	
6	CID44971	CID44971	TNBC	
7	CID44991	CID44991	TNBC	
8				
9				
10				

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

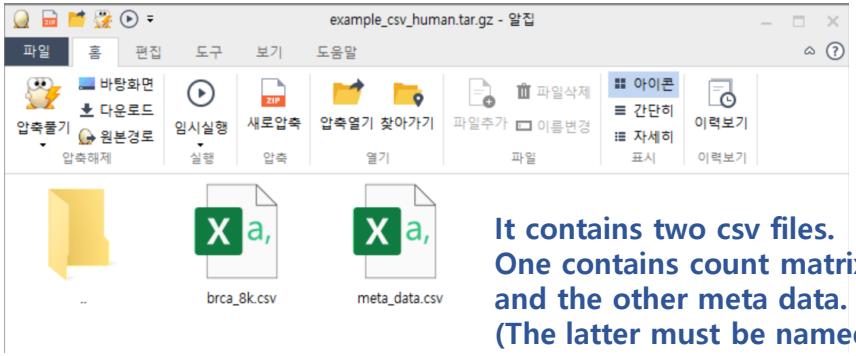
Input data formatting (2) csv format

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



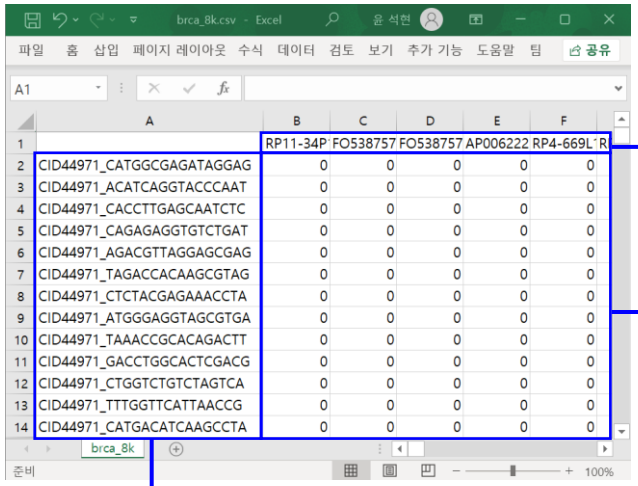
← Main folder Double click

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)

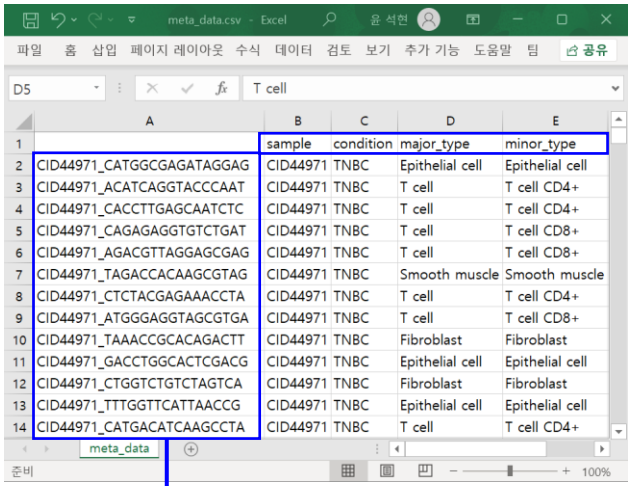


Hugo symbol

Count matrix

Cell barcode (cell ID)

4. meta_data.csv 파일



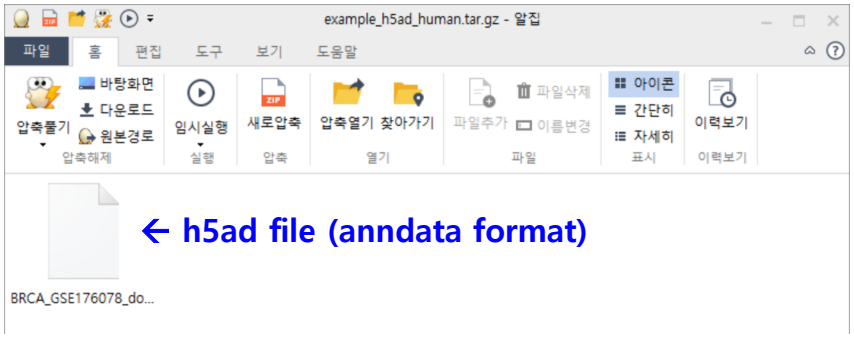
Cell barcode (cell ID)

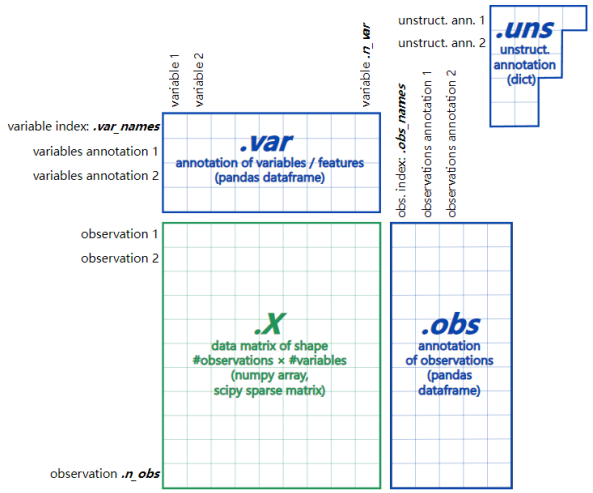
A: Index column
B: sample
C: condition
D~: optional items

- 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)





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.