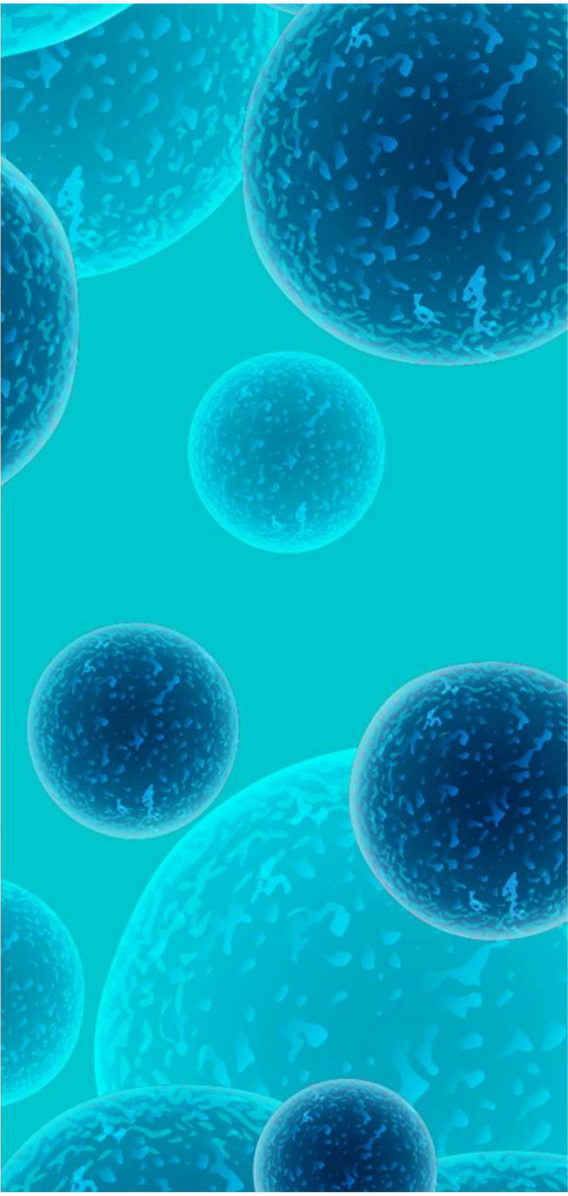
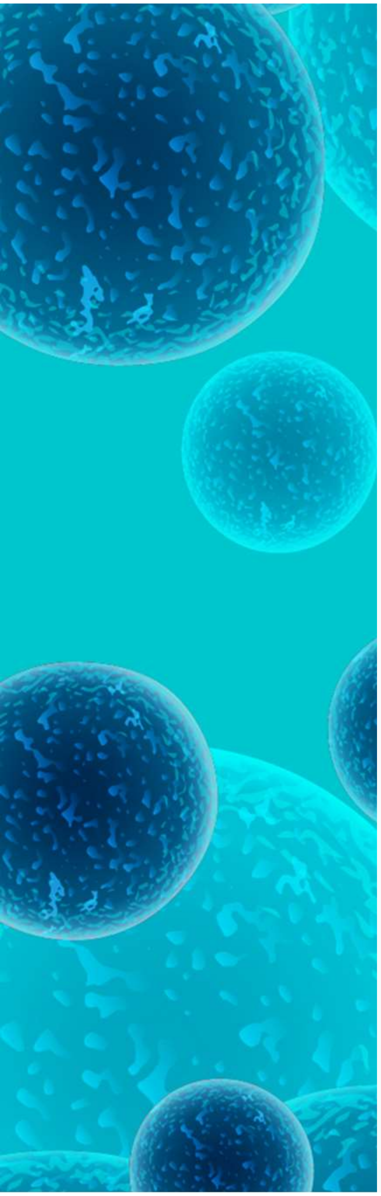
Clustering Human Organ Cells

CS-297



Name: Swathi M.V.S
ID: 016669757



Outline

01

Dataset

02

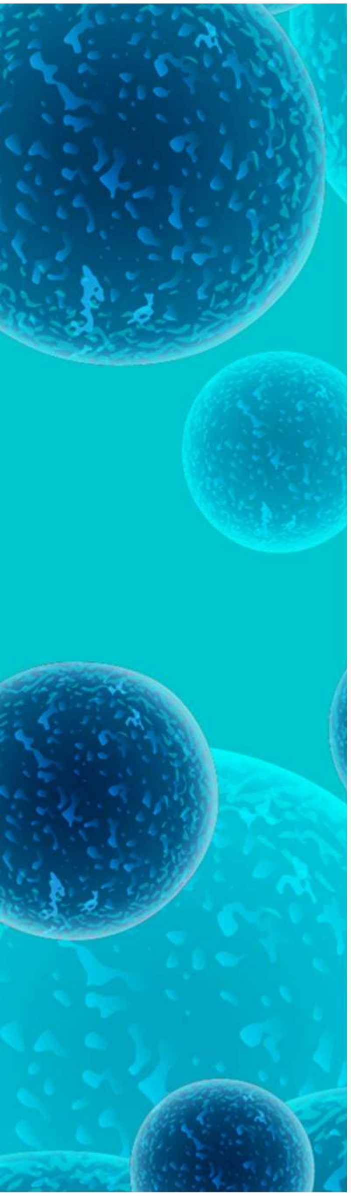
Libraries

03

Attributes

04

Dimensionality Reduction

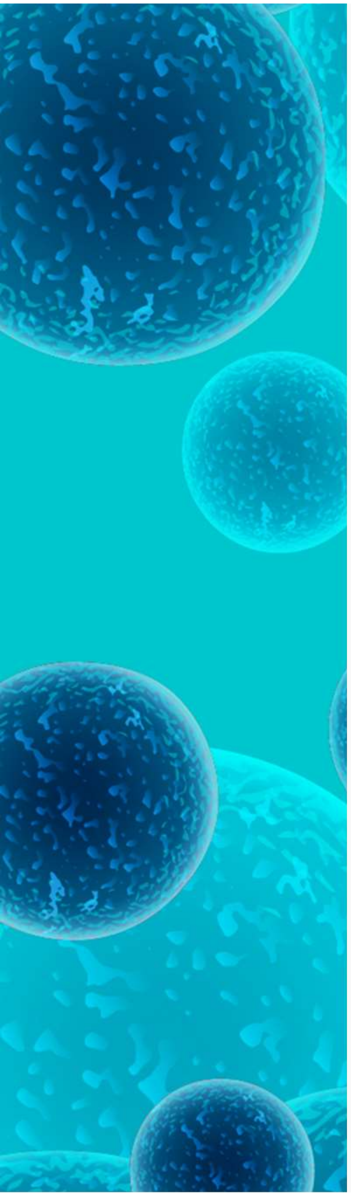


Dataset

H5ad Data Format

- Specific to R language for single-cell RNA sequencing
- Widely adapted and used persistent on-disk storage
- H5ad format - based on standard h5 format - Hierarchical Data Formats (HDF)
- HDFs - use multidimensional arrays to store large amounts of data
- H5 format - used to store scientific data that is well-organized for quick retrieval and analysis
- Developed by Theislab with extensive support in Python

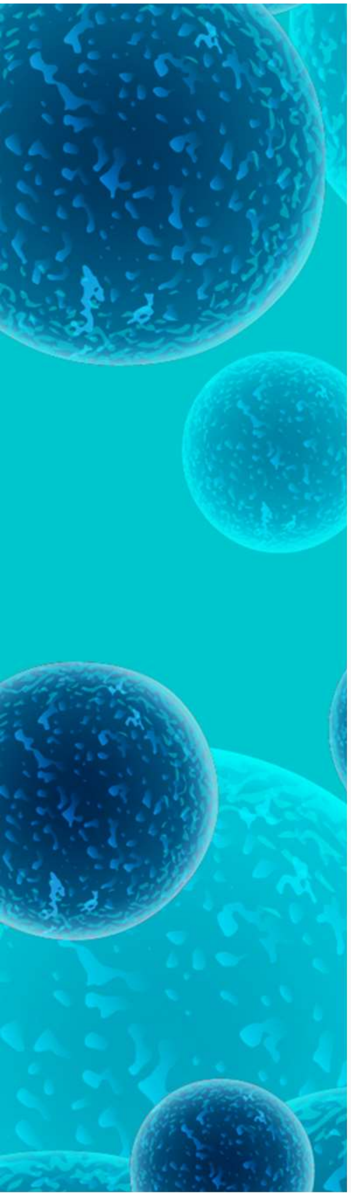
Source: Elucidata.io



Dataset

Tabula Sapiens - Heart

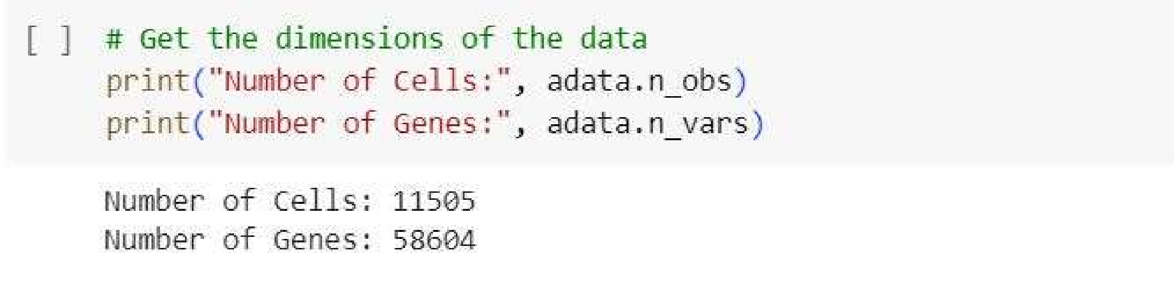
- Specific to R language.
- The H5ad format is based on the standard h5 format, a Hierarchical Data Formats (HDF) used to store large amounts of data in the form of multidimensional arrays. The H5 format is primarily used to store scientific data that is well-organized for quick retrieval and analysis.



Dataset

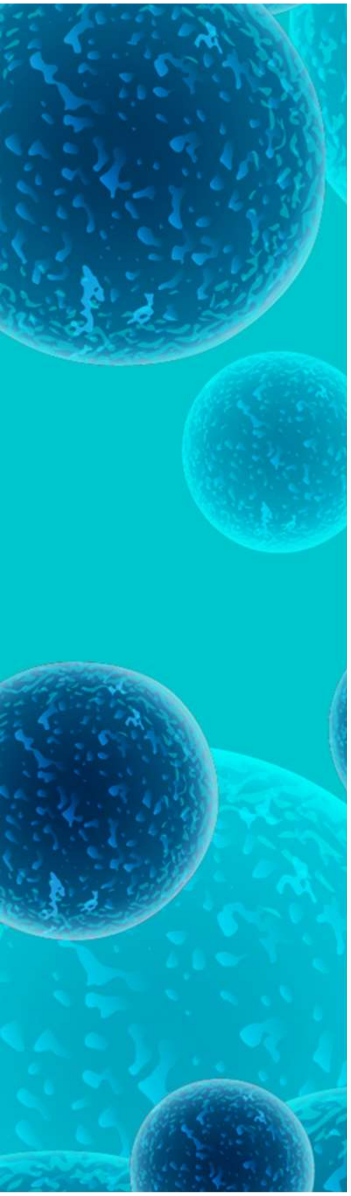
Tabula Sapiens - Heart

- The heart dataset contains 11,505 cells and 58,604 genes



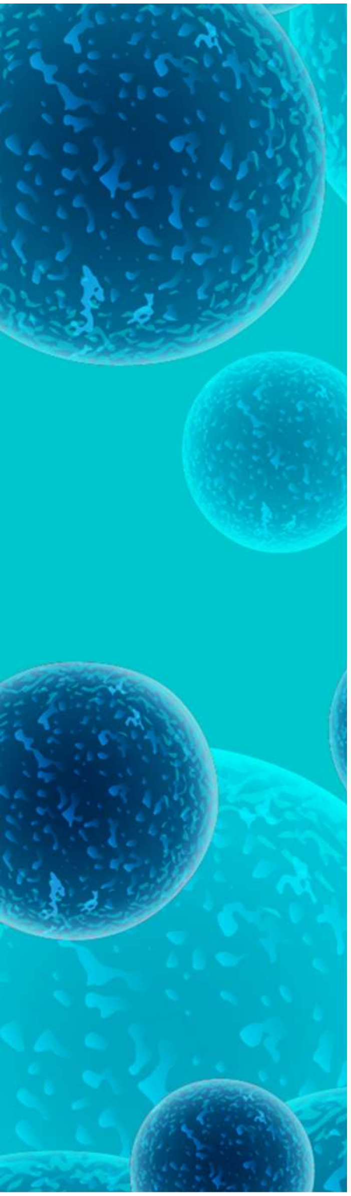
```
[ ]  # Get the dimensions of the data
     print("Number of Cells:", adata.n_obs)
     print("Number of Genes:", adata.n_vars)
```

```
     Number of Cells: 11505
     Number of Genes: 58604
```



Libraries

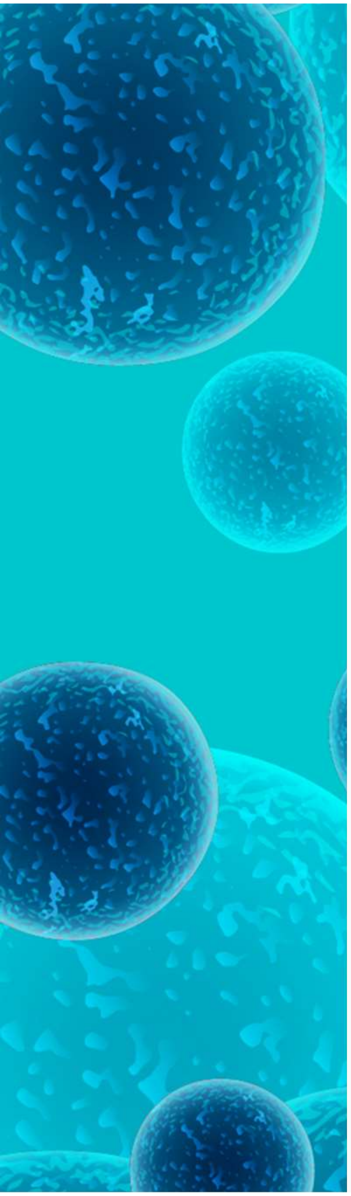
- Scanpy - It's commonly used in bioinformatics for analyzing single-cell genomics data.
- Anndata – It is used to to store and manipulate data associated with individual cells and handling annotated data
- Matplotlib - Library for creating static, animated, and interactive visualizations
- Seaborn - Built on top of Matplotlib and provides a high-level interface for creating informative and attractive statistical graphics
- Pandas - Data manipulation and analysis library. It provides data structures like DataFrame and Series
- Numpy - fundamental package for scientific computing in Python
- Os – module to interact with the operating system



Attributes

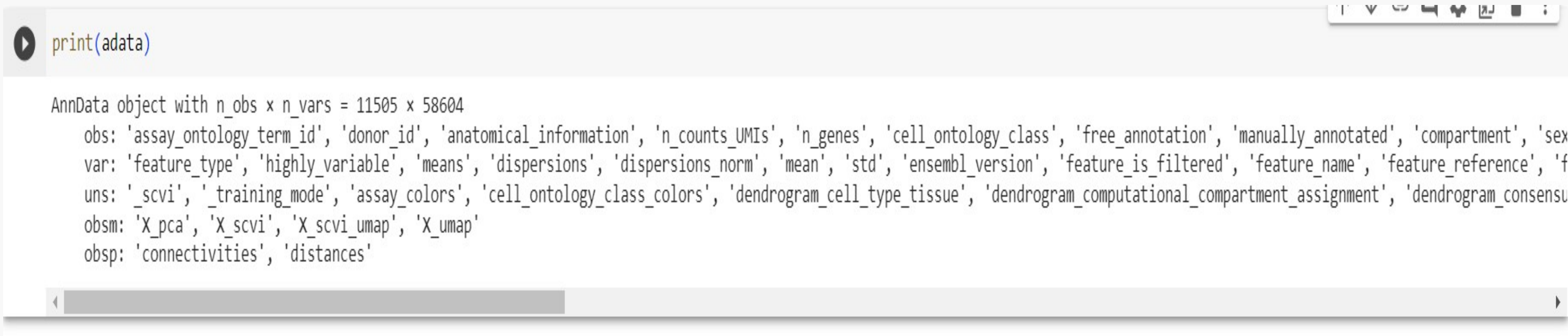
- The data contains the following attributes

```
AnnData object with n_obs × n_vars = 11505 × 58604 obs:
'assay_ontology_term_id', 'donor_id', 'anatomical_information',
'n_counts_UMIs', 'n_genes', 'cell_ontology_class',
'free_annotation', 'manually_annotated', 'compartment',
'sex_ontology_term_id', 'disease_ontology_term_id',
'is_primary_data', 'organism_ontology_term_id', 'suspension_type',
'cell_type_ontology_term_id', 'tissue_ontology_term_id',
'development_stage_ontology_term_id',
'self_reported_ethnicity_ontology_term_id', 'cell_type', 'assay',
'disease', 'organism', 'sex', 'tissue', 'self_reported_ethnicity',
'development_stage' var: 'feature_type', 'highly_variable',
'means', 'dispersions', 'dispersions_norm', 'mean', 'std',
'ensembl_version', 'feature_is_filtered', 'feature_name',
'feature_reference', 'feature_biotype' uns: '_scvi',
'_training_mode', 'assay_colors', 'cell_ontology_class_colors',
'dendrogram_cell_type_tissue',
'dendrogram_computational_compartment_assignment',
'dendrogram_consensus_prediction', 'dendrogram_tissue_cell_type',
'donor_id_colors', 'hvg', 'neighbors', 'schema_version',
'sex_colors', 'tissue_colors', 'title', 'umap' obsm: 'X_pca',
'X_scvi', 'X_scvi_umap', 'X_umap' obsp: 'connectivities',
'distances'
```



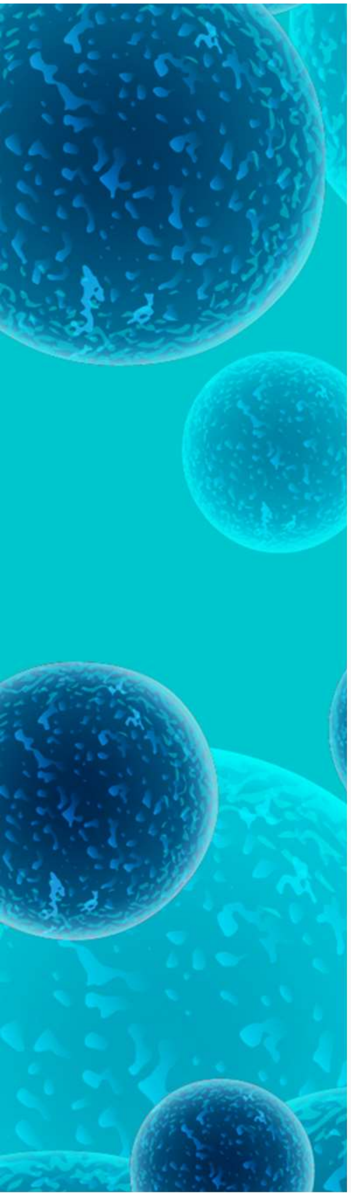
Attributes

- The data contains the following attributes



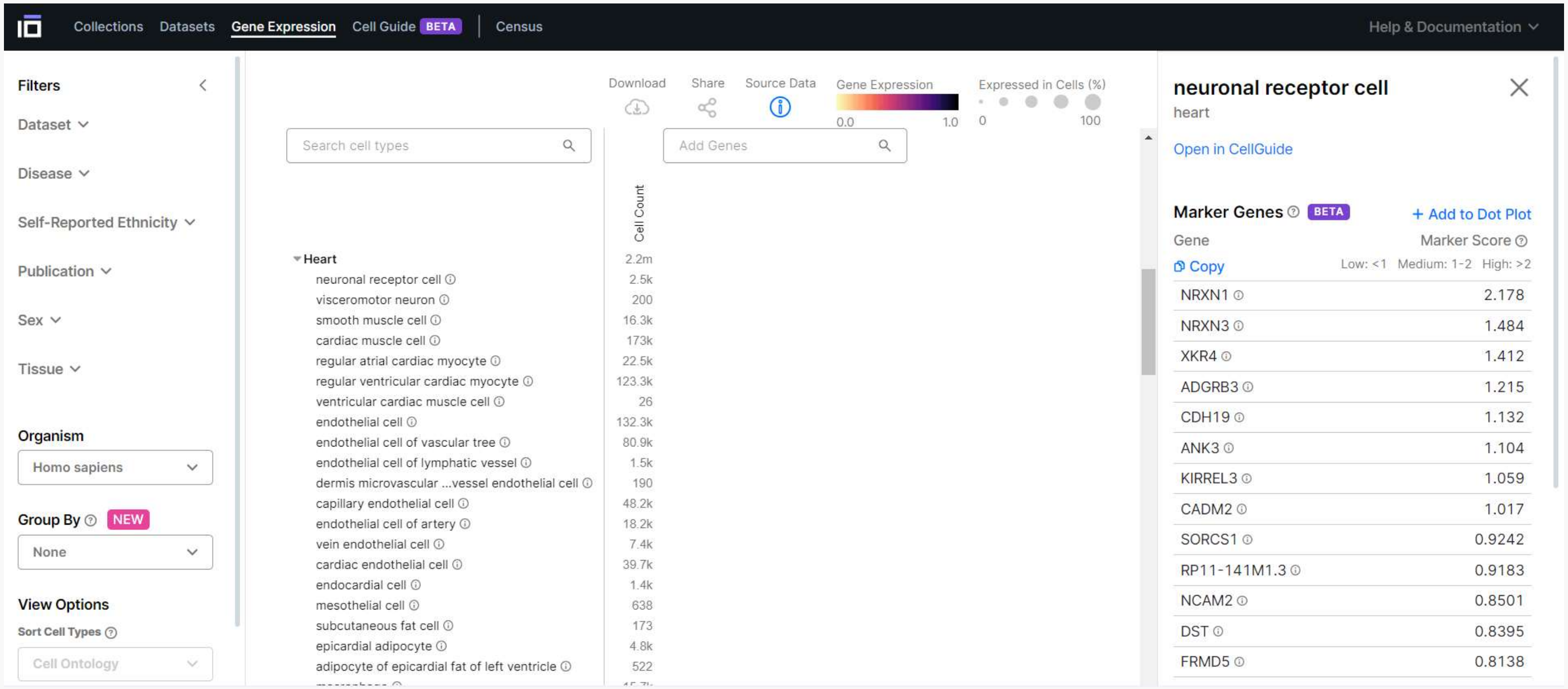
```
print(adata)
```

```
AnnData object with n_obs × n_vars = 11505 × 58604
    obs: 'assay_ontology_term_id', 'donor_id', 'anatomical_information', 'n_counts_UMIs', 'n_genes', 'cell_ontology_class', 'free_annotation', 'manually_annotated', 'compartment', 'sex
    var: 'feature_type', 'highly_variable', 'means', 'dispersions', 'dispersions_norm', 'mean', 'std', 'ensembl_version', 'feature_is_filtered', 'feature_name', 'feature_reference', 'f
    uns: '_scvi', '_training_mode', 'assay_colors', 'cell_ontology_class_colors', 'dendrogram_cell_type_tissue', 'dendrogram_computational_compartment_assignment', 'dendrogram_consensu
    obsm: 'X_pca', 'X_scvi', 'X_scvi_umap', 'X_umap'
    obsp: 'connectivities', 'distances'
```



Attributes

- Gene expression - the process by which the information encoded in a gene is turned into a function



Filters

Dataset

Disease

Self-Reported Ethnicity

Publication

Sex

Tissue

Organism: Homo sapiens

Group By NEW: None

View Options

Sort Cell Types: Cell Ontology

Heart	2.2m
neuronal receptor cell	2.5k
visceromotor neuron	200
smooth muscle cell	16.3k
cardiac muscle cell	173k
regular atrial cardiac myocyte	22.5k
regular ventricular cardiac myocyte	123.3k
ventricular cardiac muscle cell	26
endothelial cell	132.3k
endothelial cell of vascular tree	80.9k
endothelial cell of lymphatic vessel	1.5k
dermis microvascular ...vessel endothelial cell	190
capillary endothelial cell	48.2k
endothelial cell of artery	18.2k
vein endothelial cell	7.4k
cardiac endothelial cell	39.7k
endocardial cell	1.4k
mesothelial cell	638
subcutaneous fat cell	173
epicardial adipocyte	4.8k
adipocyte of epicardial fat of left ventricle	522

neuronal receptor cell

heart

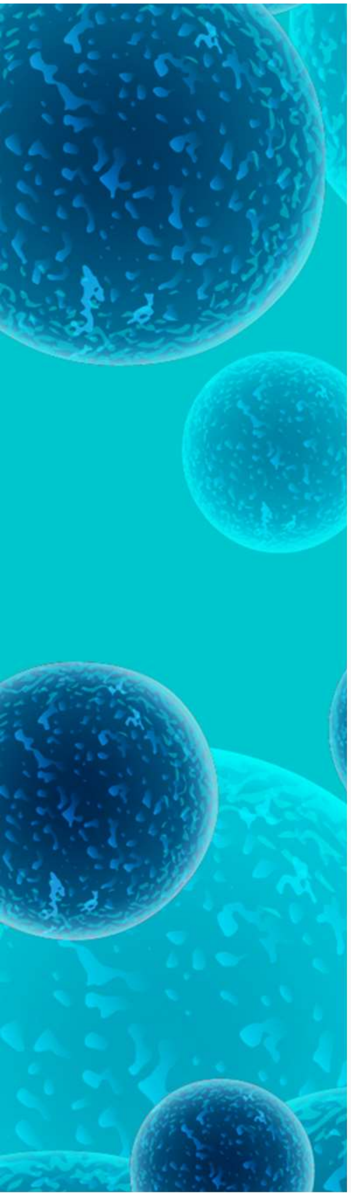
Open in CellGuide

Marker Genes BETA

+ Add to Dot Plot

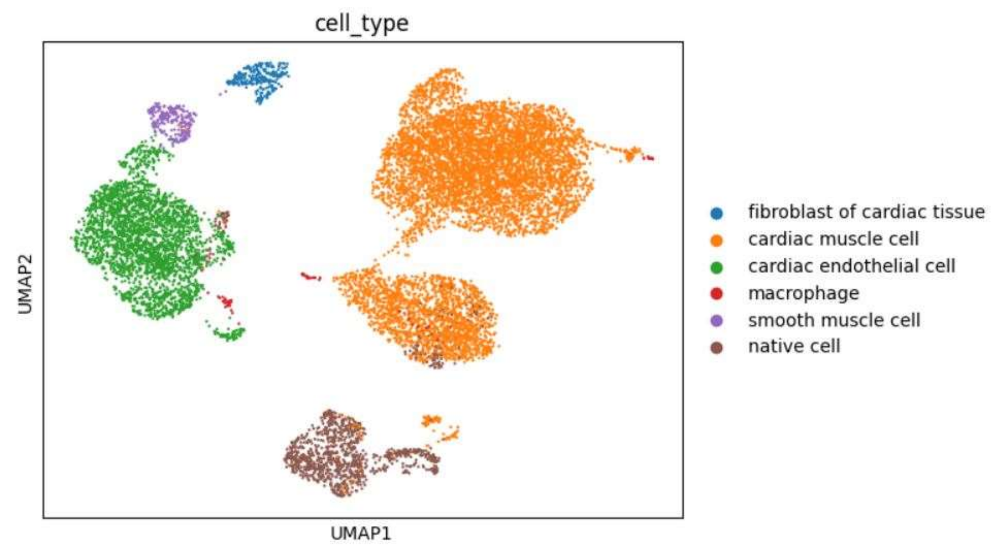
Copy — Low: <1 Medium: 1-2 High: >2

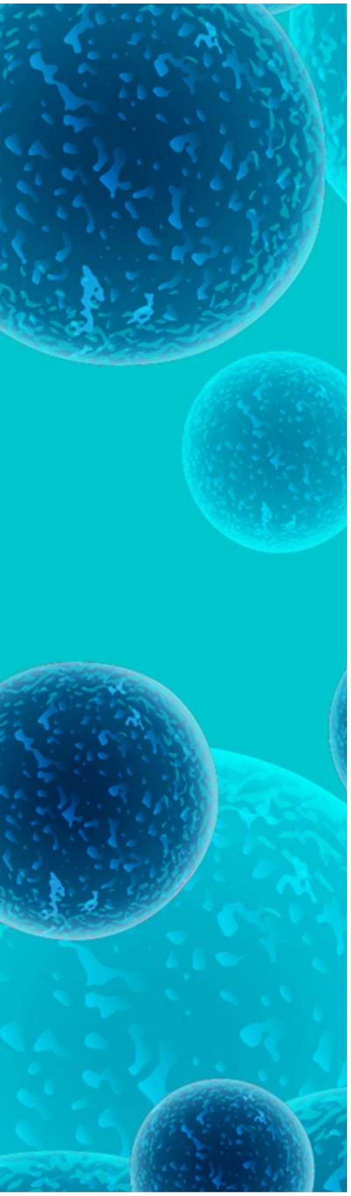
Gene	Marker Score
NRXN1	2.178
NRXN3	1.484
XKR4	1.412
ADGRB3	1.215
CDH19	1.132
ANK3	1.104
KIRREL3	1.059
CADM2	1.017
SORCS1	0.9242
RP11-141M1.3	0.9183
NCAM2	0.8501
DST	0.8395
FRMD5	0.8138



Dimensionality Reduction

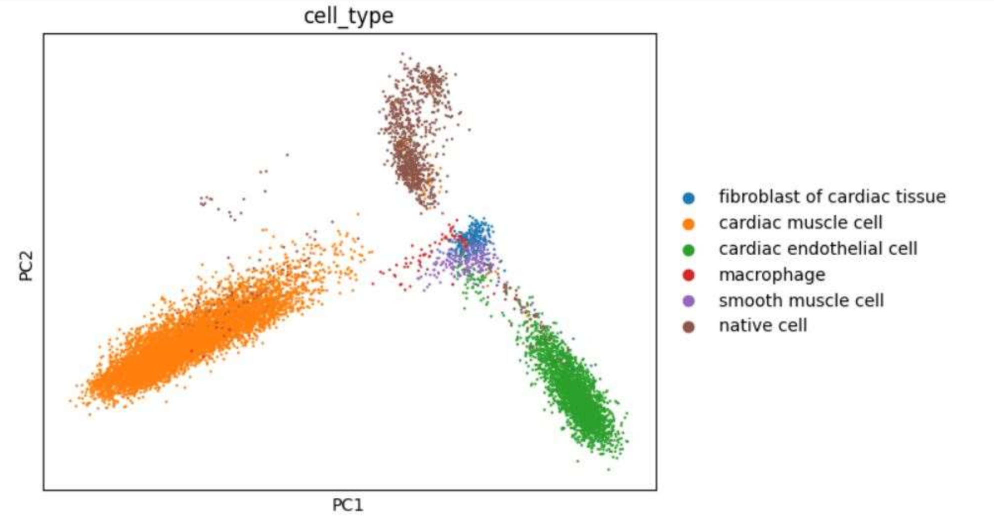
- UMAP (Uniform Manifold Approximation and Projection) - fairly flexible non-linear dimension reduction algorithm

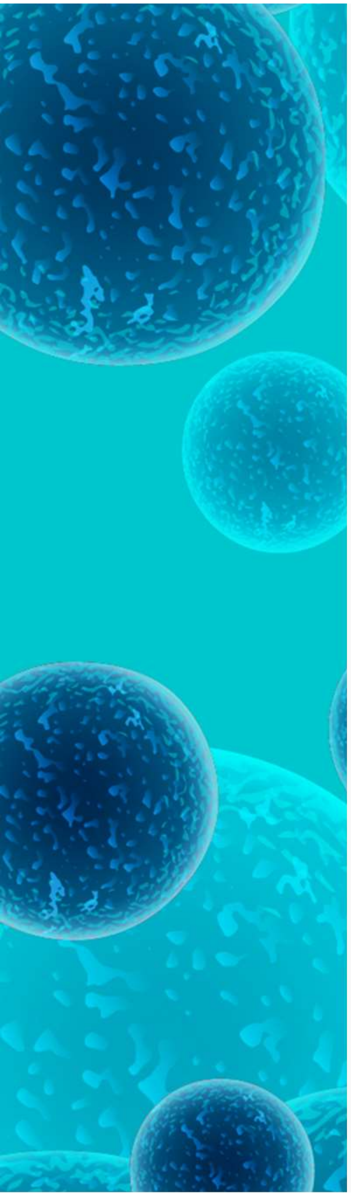




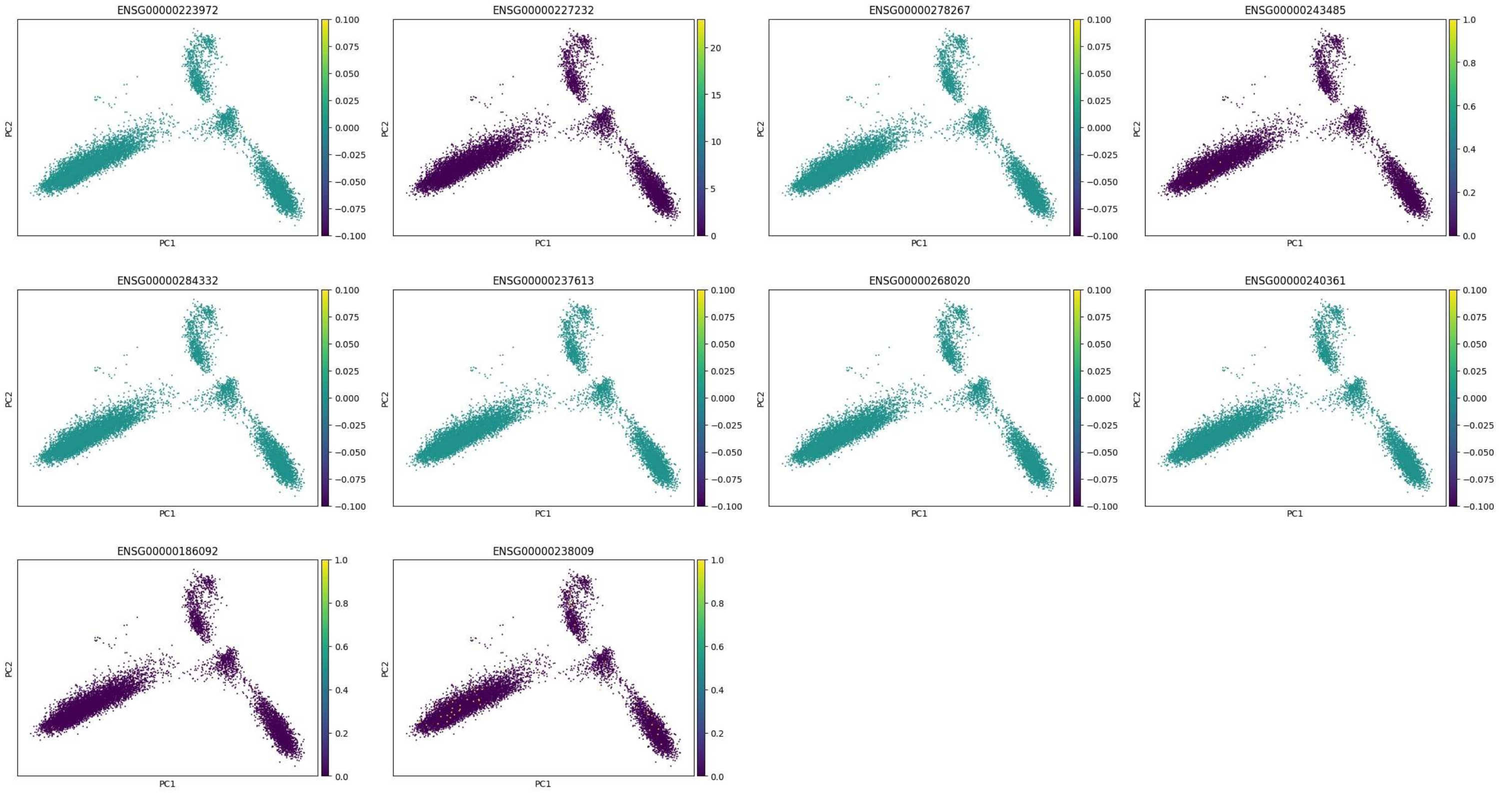
Dimensionality Reduction

- PCA - fairly flexible non-linear dimension reduction algorithm





Dimensionality Reduction


ENSG00000223972
PC1
PC2
ENSG00000227232
PC1
PC2
ENSG00000278267
PC1
PC2
ENSG00000243485
PC1
PC2
ENSG00000284332
PC1
PC2
ENSG00000237613
PC1
PC2
ENSG00000268020
PC1
PC2
ENSG00000240361
PC1
PC2
ENSG00000186092
PC1
PC2
ENSG00000238009
PC1
PC2