Key terms you have to study before workshop

PCA

Wilcoxon rank sum test

Differentially expressed genes

Leiden

Marker gene

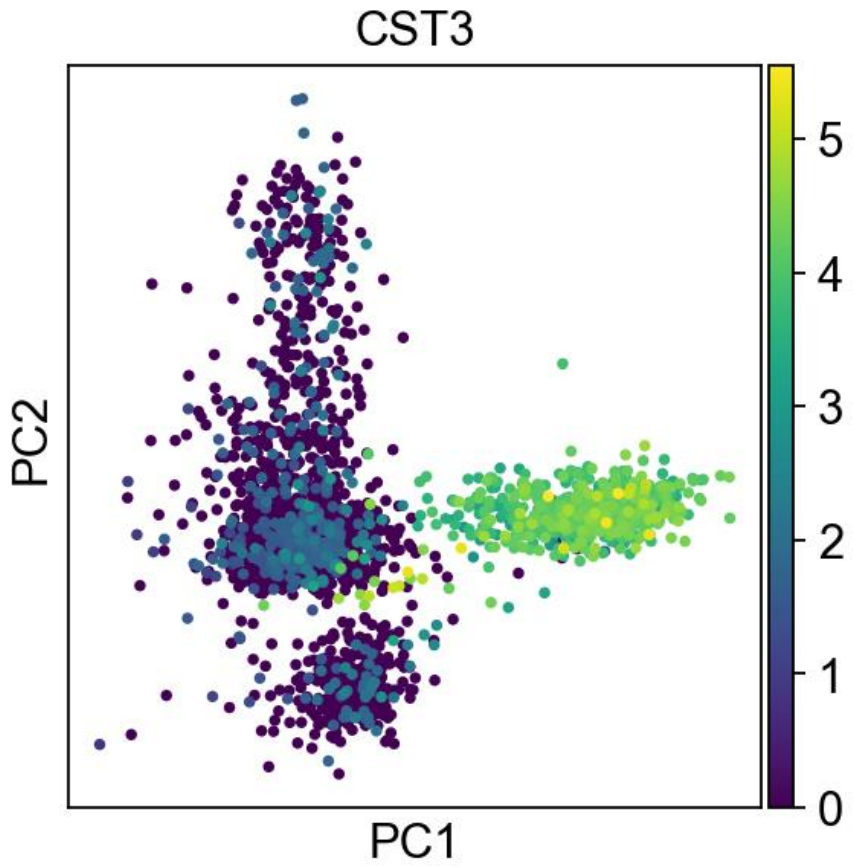
CST3 gene
cystatin C

Normal Function

The *CST3* gene provides instructions for making a protein called cystatin C. This protein is part of a family of proteins called cysteine protease inhibitors that help control several types of chemical reactions by blocking (inhibiting) the activity of certain enzymes. Cystatin C inhibits the activity of enzymes called cathepsins that cut apart other proteins in order to break them down.

Cystatin C is found in biological fluids, such as blood. Its levels are especially high in the fluid that surrounds and protects the brain and spinal cord (the cerebrospinal fluid or CSF).

```
sc.pl.pca(adata, color='CST3')
```



scanpy.pl.pca

```
scanpy.pl.pca(adata, *, color=None, mask_obs=None, gene_symbols=None,
use_raw=None, sort_order=True, edges=False, edges_width=0.1, edges_color='grey',
neighbors_key=None, arrows=False, arrows_kwds=None, groups=None, components=None,
dimensions=None, layer=None, projection='2d', scale_factor=None, color_map=None,
cmap=None, palette=None, na_color='lightgray', na_in_legend=True, size=None,
frameon=None, legend_fontsize=None, legend_fontweight='bold', legend_loc='right
margin', legend_fontoutline=None, colorbar_loc='right', vmax=None, vmin=None,
vcenter=None, norm=None, add_outline=False, outline_width=(0.3, 0.05),
outline_color=('black', 'white'), ncols=4, hspace=0.25, wspace=None, title=None,
show=None, save=None, ax=None, return_fig=None, marker='.',
annotate_var_explained=False, **kwargs)
```

[source]

Scatter plot in PCA coordinates.

Use the parameter `annotate_var_explained` to annotate the explained variance.

Parameters:

adata : *AnnData*

Annotated data matrix.

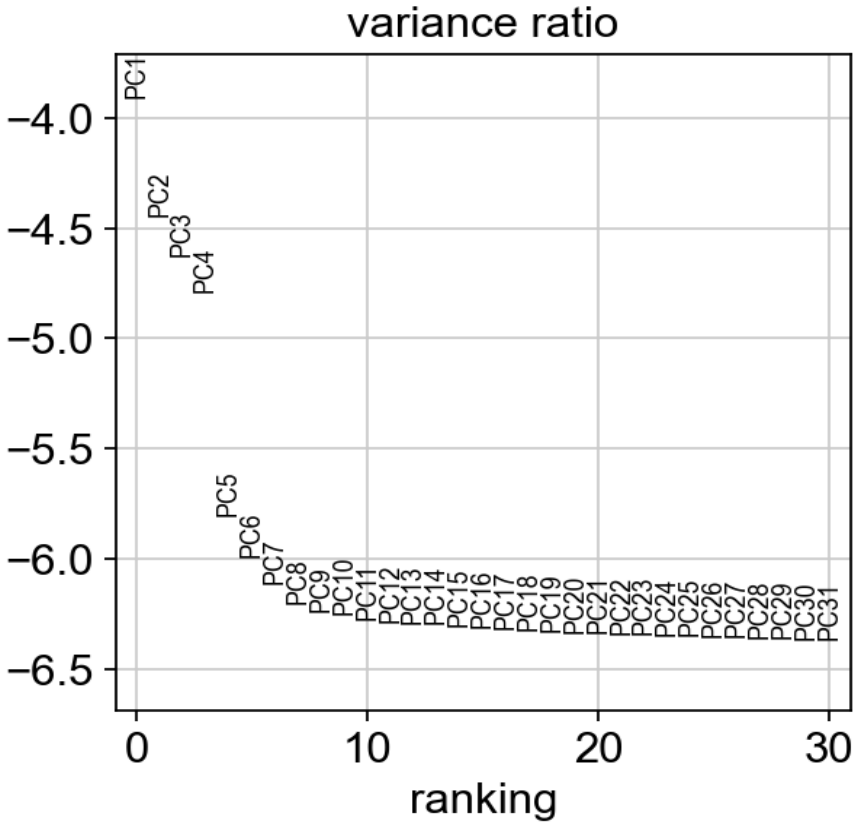
color : *str* / *Sequence*[*str*] / *None* (default: *None*)

Keys for annotations of observations/cells or variables/genes, e.g., `'ann1'` or `['ann1', 'ann2']`.

Principal component analysis

```
sc.pl.pca_variance_ratio(adata, log=True)
```

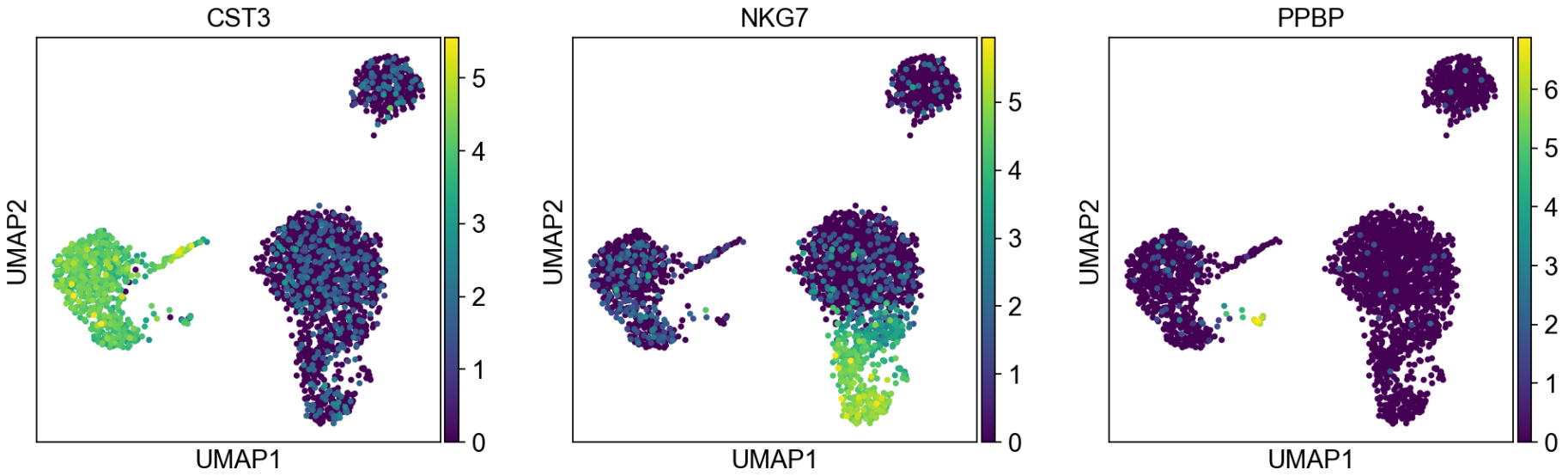
- Now that we've extracted the most "informative" cells and genes from our dataset, we can start the process of dimensional reduction by generating a list of principal components (PCs).
- Principal component analysis will reduce the number of columns (variable gene expression values) to set of PCs which explain the variance in our dataset.
- Here we're using a simple and effective method for choosing the PCs by plotting the variance ratio for each PC and choosing the last PC where the ratio starts to **"flatten" out.** For this particular dataset we've chosen 30 PCs, which will be passed to the "sc.pp.neighbors" function in the next step.



UMAP

```
sc.pl.umap(adata, color=['CST3', 'NKG7', 'PPBP'])
```

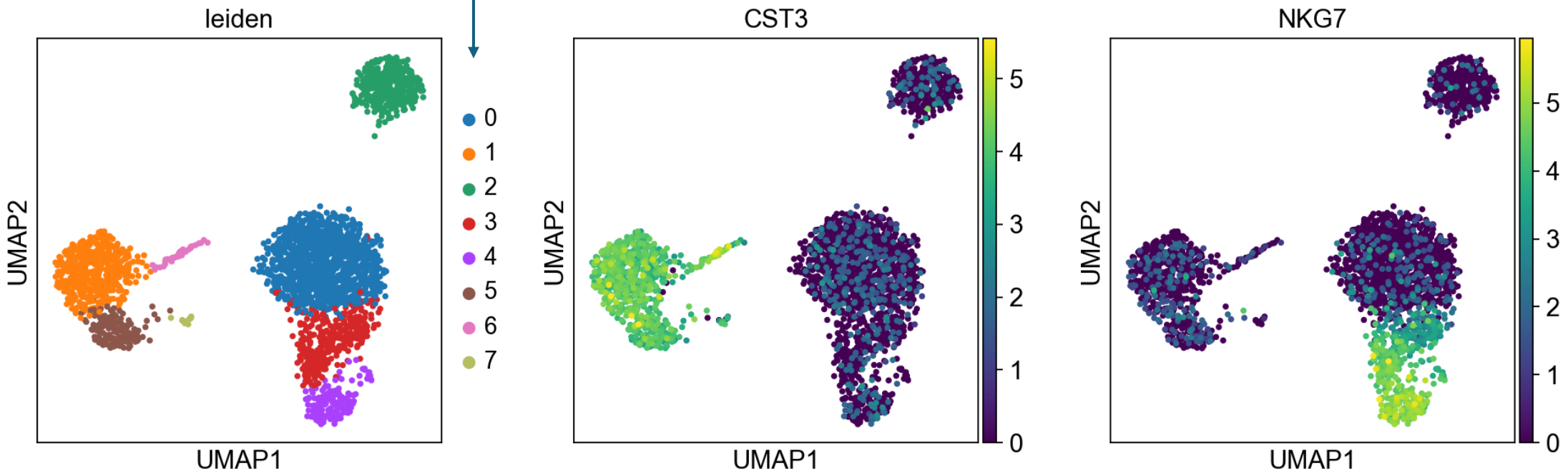
- The color scale ranges from purple (low or no expression) to yellow (high expression).



Which of these genes is most likely a potential marker?

```
sc.pl.umap(adata, color=['leiden'])
```

Leiden found 8 cluster



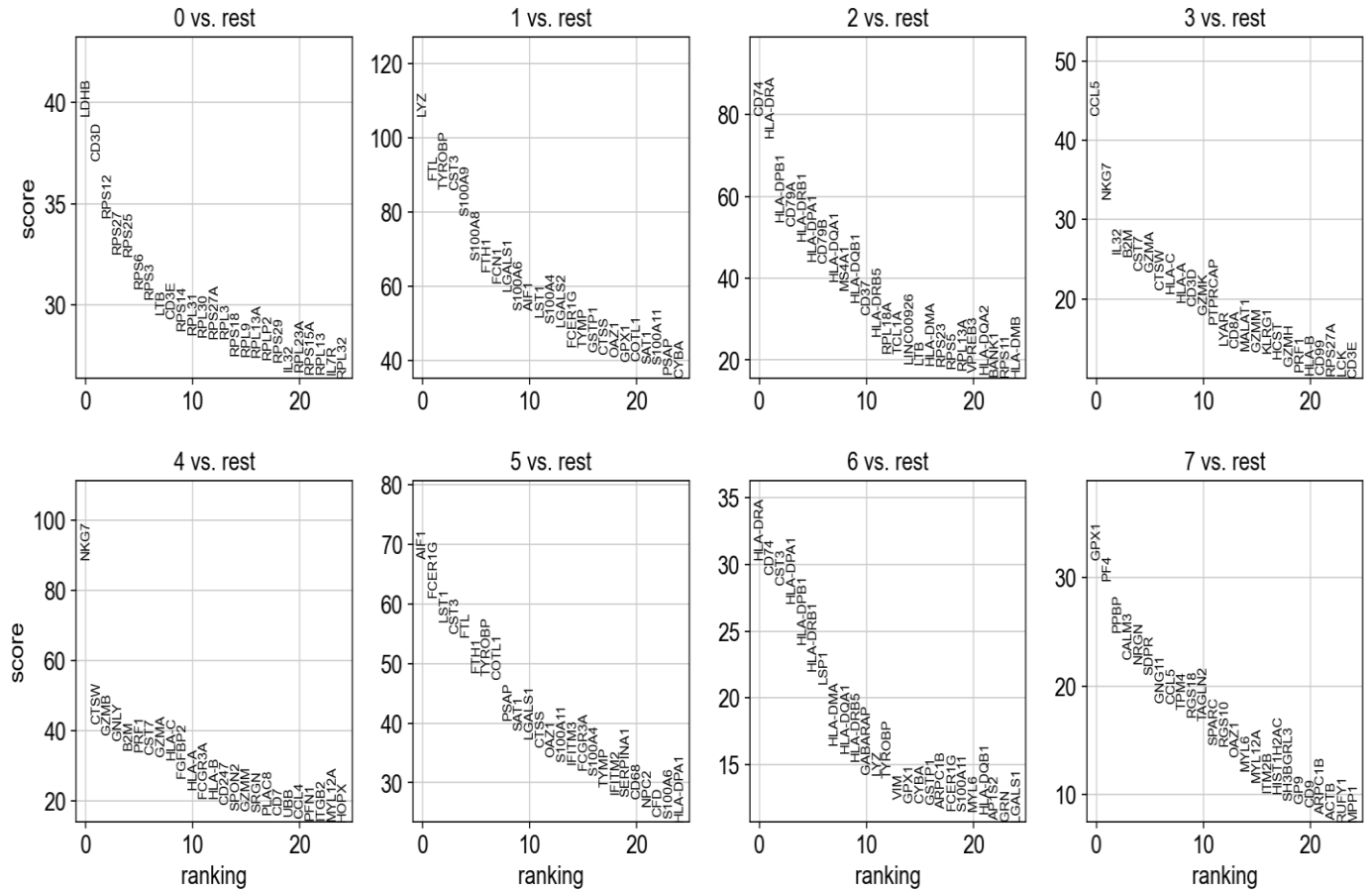
Colour points by discrete variable (Leiden clusters).

Leiden creates clusters by taking into account the number of links between cells in a cluster versus the overall expected number of links in the dataset.

The Leiden module has a resolution parameter which allows to determine the scale of the partition cluster and therefore the coarseness of the clustering. A higher resolution parameter leads to more clusters.

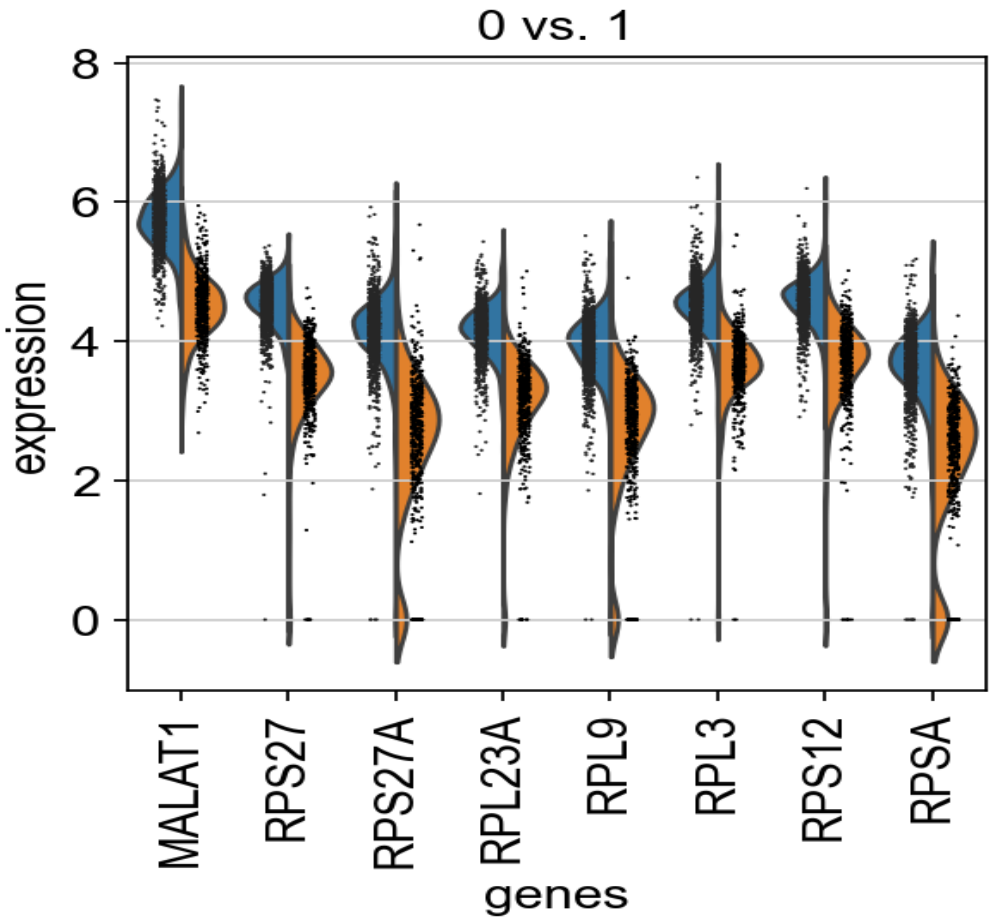
```
sc.tl.rank_genes_groups(adata, 'leiden', method='t-test')
sc.pl.rank_genes_groups(adata, n_genes=25, sharey=False)
```

After determining the appropriate number clusters, we'll perform a statistical test to find genes enriched in each cell population. For this example we'll use the simplest and quickest method, the t-test. Scanpy provides a number of different statistical tests which can be found here.



25 genes and leiden found 8 clusters

- marker_genes = ['IL7R', 'CD79A', 'MS4A1', 'CD8A', 'CD8B', 'LYZ', 'CD14','LGALS3', 'S100A8', 'GNLY', 'NKG7', 'KLRB1', 'FCGR3A', 'MS4A7', 'FCER1A', 'CST3', 'PPBP']

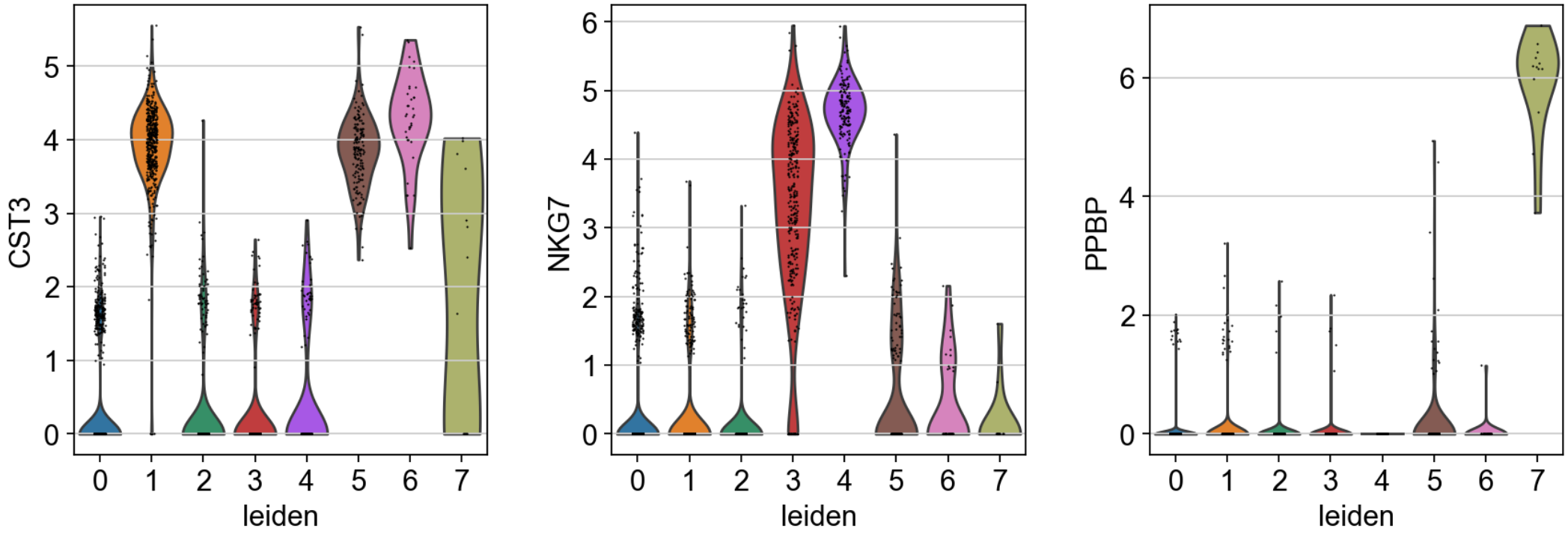


Let's get started with the known-marker based approach.

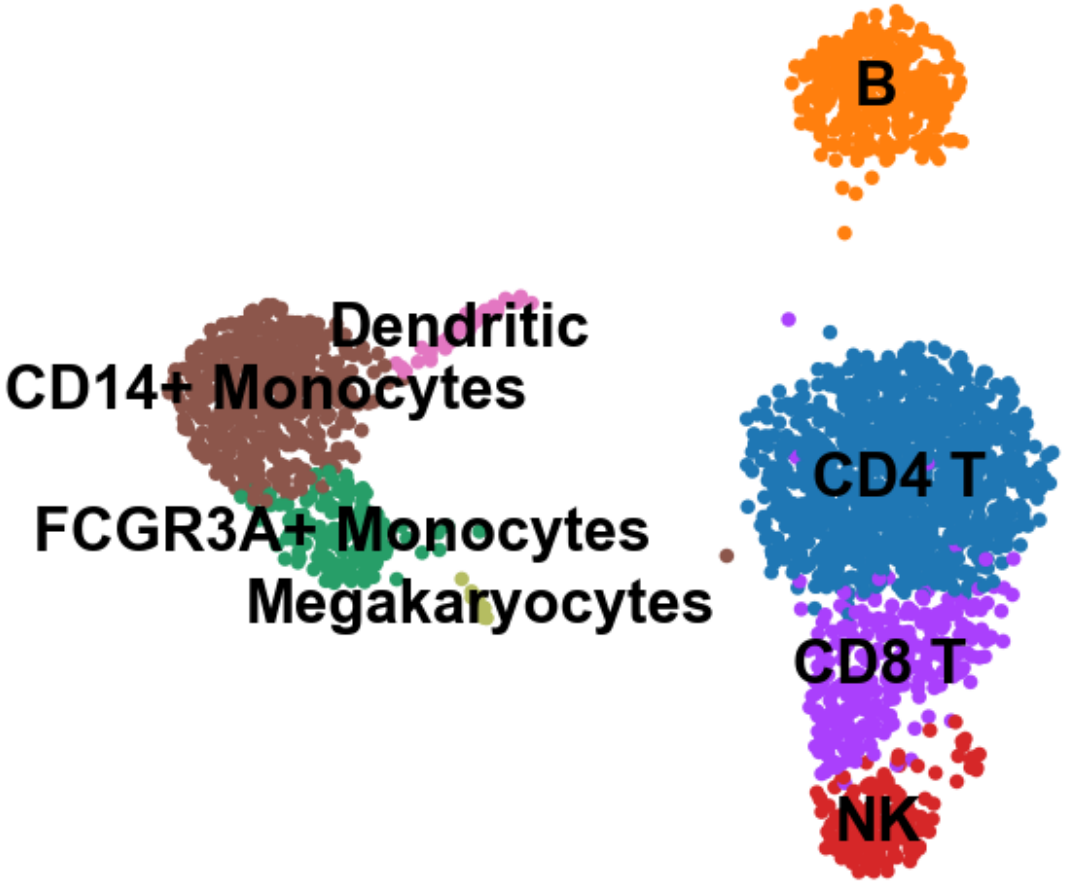
Peripheral blood mononuclear cells (PBMC) give selective responses to the immune system and are the major cells in the human body immunity. They contain several types of cells such as lymphocytes,monocytes or macrophages.

Louvain Group	Markers	Cell Type
0	IL7R	CD4 T cells
1	CD14, LYZ	CD14+ Monocytes
2	MS4A1	B cells
3	CD8A	CD8 T cells
4	GNLY, NKG7	NK cells
5	FCGR3A, MS4A7	FCGR3A+ Monocytes
6	FCER1A, CST3	Dendritic Cells
7	PPBP	Megakaryocytes

- compare a certain gene across groups

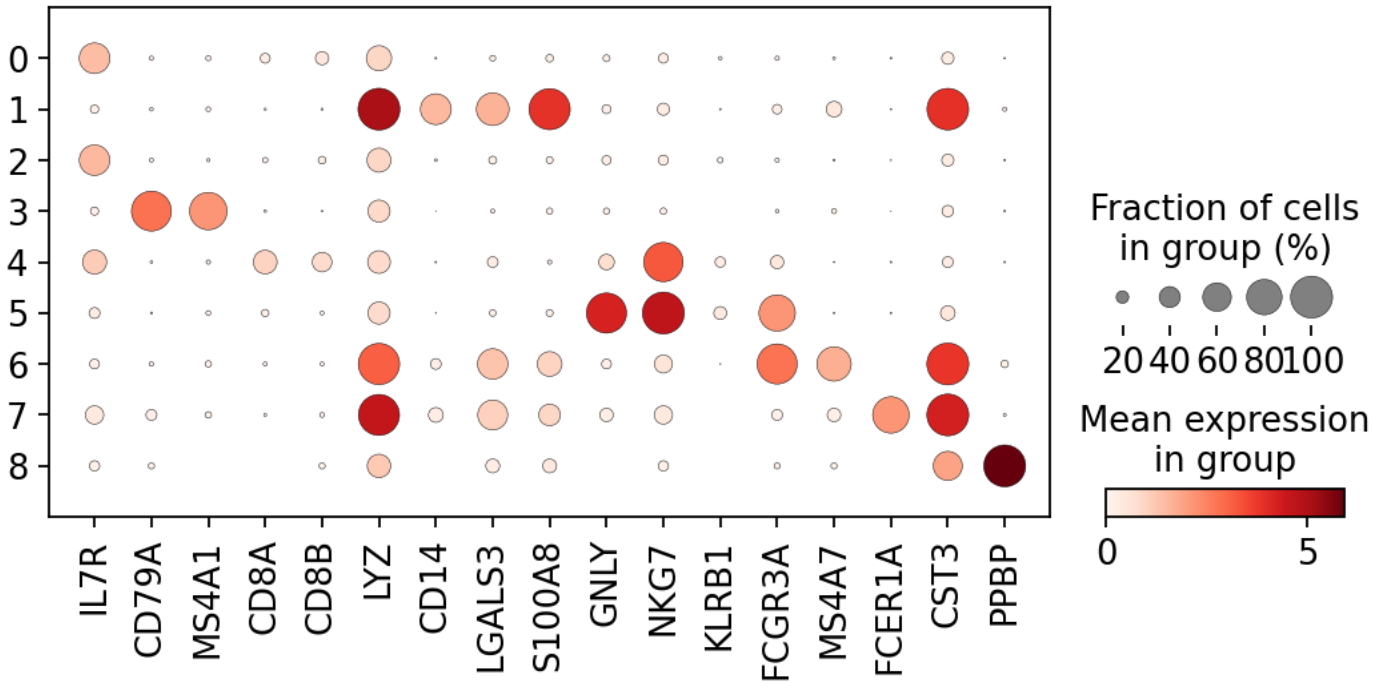


- A common goal of single cell RNA-seq analysis is to eventually classify all clusters into a cell "type".



As you can see, a lot of the differentially expressed genes are highly expressed in multiple clusters.

Let's take a look at cluster 7, which seems to have a set of relatively unique markers including PPBP and LYZ.



Useful links

- https://www.sc-best-practices.org/introduction/scrna_seq.html
- https://www.kaggle.com/code/aayush9753/3-normalization-pca-in-single-cell-rna-seq-data
- https://support.parsebiosciences.com/hc/en-us/articles/7704577188500-How-to-analyze-a-1-million-cell-data-set-using-Scanpy-and-Harmony