```
!pip install anndata
!pip install 'scanpy[leiden]'
!pip install python-igraph louvain
!pip install gseapy
```

```
⤓  Requirement already satisfied: anndata in /usr/local/lib/python3.10/dist-packages (0.10.9)
   Requirement already satisfied: array-api-compat!=1.5,>1.4 in /usr/local/lib/python3.10/dist-packages (from anndata) (1.9)
   Requirement already satisfied: exceptiongroup in /usr/local/lib/python3.10/dist-packages (from anndata) (1.2.2)
   Requirement already satisfied: h5py>=3.1 in /usr/local/lib/python3.10/dist-packages (from anndata) (3.11.0)
   Requirement already satisfied: natsort in /usr/local/lib/python3.10/dist-packages (from anndata) (8.4.0)
   Requirement already satisfied: numpy>=1.23 in /usr/local/lib/python3.10/dist-packages (from anndata) (1.26.4)
   Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from anndata) (24.1)
   Requirement already satisfied: pandas!=2.1.0rc0,!=2.1.2,>=1.4 in /usr/local/lib/python3.10/dist-packages (from anndata) (2.2
   Requirement already satisfied: scipy>1.8 in /usr/local/lib/python3.10/dist-packages (from anndata) (1.13.1)
   Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.10/dist-packages (from pandas!=2.1.0rc0,!=2.
   Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas!=2.1.0rc0,!=2.1.2,>=1.4-
   Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.10/dist-packages (from pandas!=2.1.0rc0,!=2.1.2,>=1.
   Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.8.2->pandas!=2.1
   Requirement already satisfied: scanpy[leiden] in /usr/local/lib/python3.10/dist-packages (1.10.3)
   Requirement already satisfied: anndata>=0.8 in /usr/local/lib/python3.10/dist-packages (from scanpy[leiden]) (0.10.9)
   Requirement already satisfied: h5py>=3.1 in /usr/local/lib/python3.10/dist-packages (from scanpy[leiden]) (3.11.0)
   Requirement already satisfied: joblib in /usr/local/lib/python3.10/dist-packages (from scanpy[leiden]) (1.4.2)
   Requirement already satisfied: legacy-api-wrap>=1.4 in /usr/local/lib/python3.10/dist-packages (from scanpy[leiden]) (1.4)
   Requirement already satisfied: matplotlib>=3.6 in /usr/local/lib/python3.10/dist-packages (from scanpy[leiden]) (3.7.1)
   Requirement already satisfied: natsort in /usr/local/lib/python3.10/dist-packages (from scanpy[leiden]) (8.4.0)
   Requirement already satisfied: networkx>=2.7 in /usr/local/lib/python3.10/dist-packages (from scanpy[leiden]) (3.3)
   Requirement already satisfied: numba>=0.56 in /usr/local/lib/python3.10/dist-packages (from scanpy[leiden]) (0.60.0)
   Requirement already satisfied: numpy>=1.23 in /usr/local/lib/python3.10/dist-packages (from scanpy[leiden]) (1.26.4)
   Requirement already satisfied: packaging>=21.3 in /usr/local/lib/python3.10/dist-packages (from scanpy[leiden]) (24.1)
   Requirement already satisfied: pandas>=1.5 in /usr/local/lib/python3.10/dist-packages (from scanpy[leiden]) (2.2.2)
   Requirement already satisfied: patsy in /usr/local/lib/python3.10/dist-packages (from scanpy[leiden]) (0.5.6)
   Requirement already satisfied: pynndescent>=0.5 in /usr/local/lib/python3.10/dist-packages (from scanpy[leiden]) (0.5.13)
   Requirement already satisfied: scikit-learn>=0.24 in /usr/local/lib/python3.10/dist-packages (from scanpy[leiden]) (1.5.2)
   Requirement already satisfied: scipy>=1.8 in /usr/local/lib/python3.10/dist-packages (from scanpy[leiden]) (1.13.1)
   Requirement already satisfied: seaborn>=0.13 in /usr/local/lib/python3.10/dist-packages (from scanpy[leiden]) (0.13.1)
   Requirement already satisfied: session-info in /usr/local/lib/python3.10/dist-packages (from scanpy[leiden]) (1.0.0)
   Requirement already satisfied: statsmodels>=0.13 in /usr/local/lib/python3.10/dist-packages (from scanpy[leiden]) (0.14.4)
   Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from scanpy[leiden]) (4.66.5)
   Requirement already satisfied: umap-learn!=0.5.0,>=0.5 in /usr/local/lib/python3.10/dist-packages (from scanpy[leiden]) (0.5
   Requirement already satisfied: igraph>=0.10 in /usr/local/lib/python3.10/dist-packages (from scanpy[leiden]) (0.11.6)
   Requirement already satisfied: leidenalg>=0.9.0 in /usr/local/lib/python3.10/dist-packages (from scanpy[leiden]) (0.10.2)
   Requirement already satisfied: array-api-compat!=1.5,>1.4 in /usr/local/lib/python3.10/dist-packages (from anndata>=0.8->sca
   Requirement already satisfied: exceptiongroup in /usr/local/lib/python3.10/dist-packages (from anndata>=0.8->scanpy[leiden])
   Requirement already satisfied: texttable>=1.6.2 in /usr/local/lib/python3.10/dist-packages (from igraph>=0.10->scanpy[leiden
   Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.6->scanpy[lei
   Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.6->scanpy[leiden]
   Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.6->scanpy[le
   Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.6->scanpy[le
   Requirement already satisfied: pillow>=6.2.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.6->scanpy[leiden
   Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.6->scanpy[lei
   Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.6->scanpy
   Requirement already satisfied: llvmlite<0.44,>=0.43.0dev0 in /usr/local/lib/python3.10/dist-packages (from numba>=0.56->scan
   Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas>=1.5->scanpy[leiden]) (2
   Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.10/dist-packages (from pandas>=1.5->scanpy[leiden])
   Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.10/dist-packages (from scikit-learn>=0.24->sca
   Requirement already satisfied: six in /usr/local/lib/python3.10/dist-packages (from patsy->scanpy[leiden]) (1.16.0)
   Requirement already satisfied: stdlib-list in /usr/local/lib/python3.10/dist-packages (from session-info->scanpy[leiden]) (0
   Requirement already satisfied: python-igraph in /usr/local/lib/python3.10/dist-packages (0.11.6)
   Requirement already satisfied: louvain in /usr/local/lib/python3.10/dist-packages (0.8.2)
   Requirement already satisfied: igraph==0.11.6 in /usr/local/lib/python3.10/dist-packages (from python-igraph) (0.11.6)
   Requirement already satisfied: texttable>=1.6.2 in /usr/local/lib/python3.10/dist-packages (from igraph==0.11.6->python-igra
   Collecting gseapy
     Downloading gseapy-1.1.3-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (11 kB)
```

```
from google.colab import drive
drive.mount('/content/drive')
```

```
⤓  Mounted at /content/drive
```

```
import os
main_dir = '/content/drive/MyDrive/'
senior_year_dir = '/content/drive/MyDrive/Senior Year/'
```

```
import scanpy as sc
```

```python
extract_dir = "/content/drive/MyDrive/Senior Year/Applied Data Science/"
adata = sc.read_h5ad(extract_dir + 'frogtail.h5ad')
print(adata)
```

```
AnnData object with n_obs × n_vars = 13199 × 31535
    obs: 'barcode_cells', 'cell', 'sample', 'DevelopmentalStage', 'DaysPostAmputation', 'cluster', 'X', 'Y', 'CellCyclePhase
/usr/local/lib/python3.10/dist-packages/anndata/_core/aligned_df.py:68: ImplicitModificationWarning: Transforming to str ind
  warnings.warn("Transforming to str index.", ImplicitModificationWarning)
```

```python
adata.obs.head()
```

	barcode_cells	cell	sample	DevelopmentalStage	DaysPostAmputation	cluster	X	Y
0	AAACCTGAGCTAGTTC.1	AAACCTGAGCTAGTTC.1	SIGAB5	st40	3	Erythrocyte 4	-6.395430	0.832323
1	AAACCTGGTGGGTCAA.1	AAACCTGGTGGGTCAA.1	SIGAB5	st40	3	Myeloid 1	-2.428271	13.826715
2	AAACCTGGTTTGTTGG.1	AAACCTGGTTTGTTGG.1	SIGAB5	st40	3	Beta ionocyte	-1.398049	-14.653897
3	AAACGGGGTCGGCATC.1	AAACGGGGTCGGCATC.1	SIGAB5	st40	3	Erythrocyte 4	-5.863765	0.489598
4	AAACGGGTCCTACAGA.1	AAACGGGTCCTACAGA.1	SIGAB5	st40	3	Goblet cell	2.106444	-6.594430

```python
adata.X
```

```
<13199x31535 sparse matrix of type '<class 'numpy.int64'>'
    with 29297098 stored elements in Compressed Sparse Row format>
```
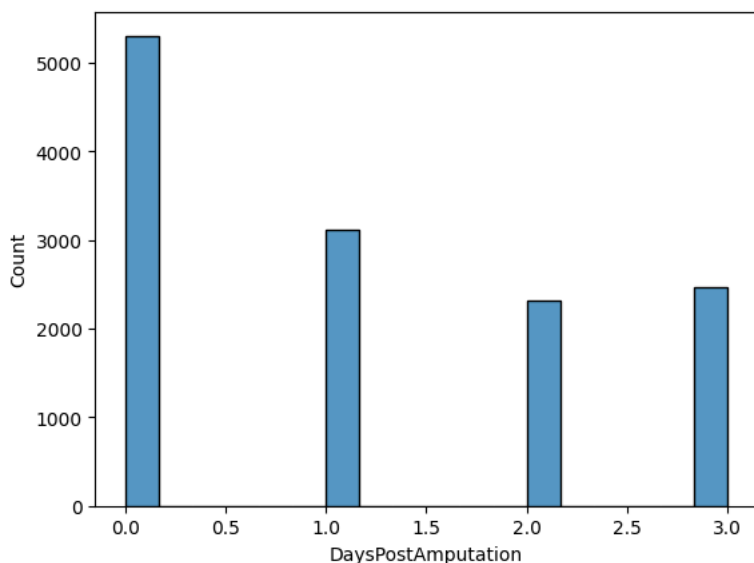
```python
adata.X.toarray()[:5, :20]
```

```
array([[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0],
       [0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0],
       [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]])
```

```python
import seaborn as sns
sns.histplot(adata.obs['DaysPostAmputation'])
```

```
<Axes: xlabel='DaysPostAmputation', ylabel='Count'>
```



```python
adata = adata[adata.obs['DaysPostAmputation'] == 0] #Subset on the time point 0
adata.layers['counts'] = adata.X.copy()
```

```
<ipython-input-10-88f00e64bca0>:2: ImplicitModificationWarning: Setting element `.layers['counts']` of view, initializing vi
  adata.layers['counts'] = adata.X.copy() # Copy the raw counts in a different layer
```

```
hvg_adata = sc.pp.log1p(adata, copy=True) #Log-normalize the data
sc.pp.highly_variable_genes(hvg_adata, n_top_genes=2000) #Select highly variable genes


#Scaling data
sc.pp.scale(hvg_adata)

#PCA on highly variable genes data
sc.tl.pca(hvg_adata)

#PCA diagnostics
sc.pl.pca_overview(hvg_adata, color='cluster', components=['1,2', '2,3'])  # Ensure 'cluster' is a valid annotation

#Variance ratio for the PCA
sc.pl.pca_variance_ratio(hvg_adata, log=True)  # Use the same object as above
```

cluster   cluster

Alpha ionocyte	Interneuron 2		Myeloid 1
Anterior notochord	Interneuron 3		Myeloid 2
Beta ionocyte	Interneuron 4		Myotome
Dermomyotome	Lymphoid 1 (Gata2-, Gata3+)		Posterior notochord
Differentiating myocyte	Lymphoid 2 (Cxcr6+)		ROCs
Differentiating neuron	Lymphoid 3		Satellite cell
Dopaminergic neurons	Lymphoid 4 (Gata2+, Gata3-)		Skeletal muscle
Epidermis	Lymphoid 5 (CD19+)		Small secretory cell
Erythrocyte 1	Lymphoid endothelial cells		Smooth muscle
Erythrocyte 2	Melanocyte		Spinal cord progenitor
Erythrocyte 3	Melanocyte precursor		Syndetome
Erythrocyte 4	Melanocyte stem cell		Vascular endothelial cell
Floor plate	Mesenchyme		Vulnerable Motor Neuron
Goblet cell	Motor neuron		
Interneuron 1			

PC1   PC2   PC3


variance ratio

variance ratio

ranking

```python
#Computing neighbors
sc.pp.neighbors(hvg_adata, n_neighbors=10, use_rep='X_pca')

#Leiden clustering
sc.tl.leiden(hvg_adata, resolution=0.5)

#UMAP
sc.tl.umap(hvg_adata)

#UMAP with Leiden clusters
sc.pl.umap(hvg_adata, color="leiden", title='Leiden Clusters')
```

⤷ <ipython-input-17-ca6fbfdc2d42>:5: FutureWarning: In the future, the default backend for leiden will be igraph instead of le

   To achieve the future defaults please pass: flavor="igraph" and n_iterations=2. directed must also be False to work with i
     sc.tl.leiden(hvg_adata, resolution=0.5)

Leiden Clusters
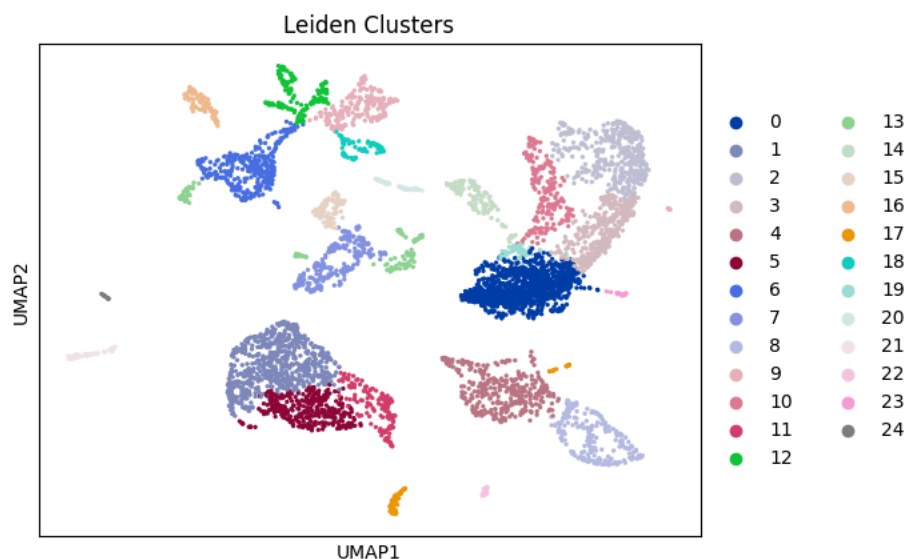
```python
from sklearn.metrics import adjusted_rand_score, silhouette_score

leiden_labels = hvg_adata.obs['leiden']

#Silhouette score
sil_score = silhouette_score(hvg_adata.obsm['X_umap'], leiden_labels)

print(f"Silhouette Score: {sil_score}")
```
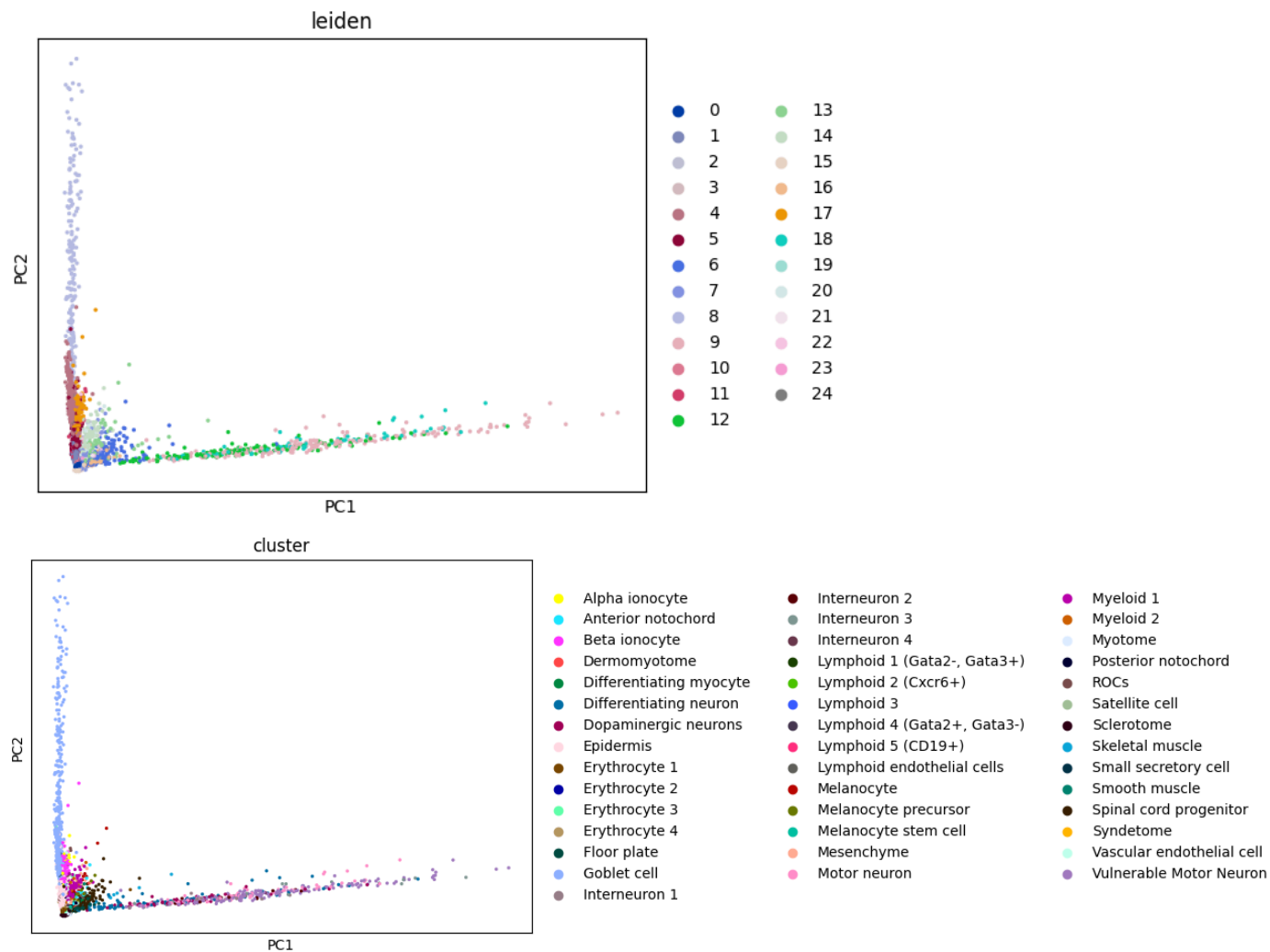
⤷ Silhouette Score: 0.3722538948059082

```python
sc.pl.pca(hvg_adata, color=['leiden']) # PC1/PC2 - Leiden colors
sc.pl.pca(hvg_adata, color=['cluster']) # Plotting PC1/PC2 - Original paper
```
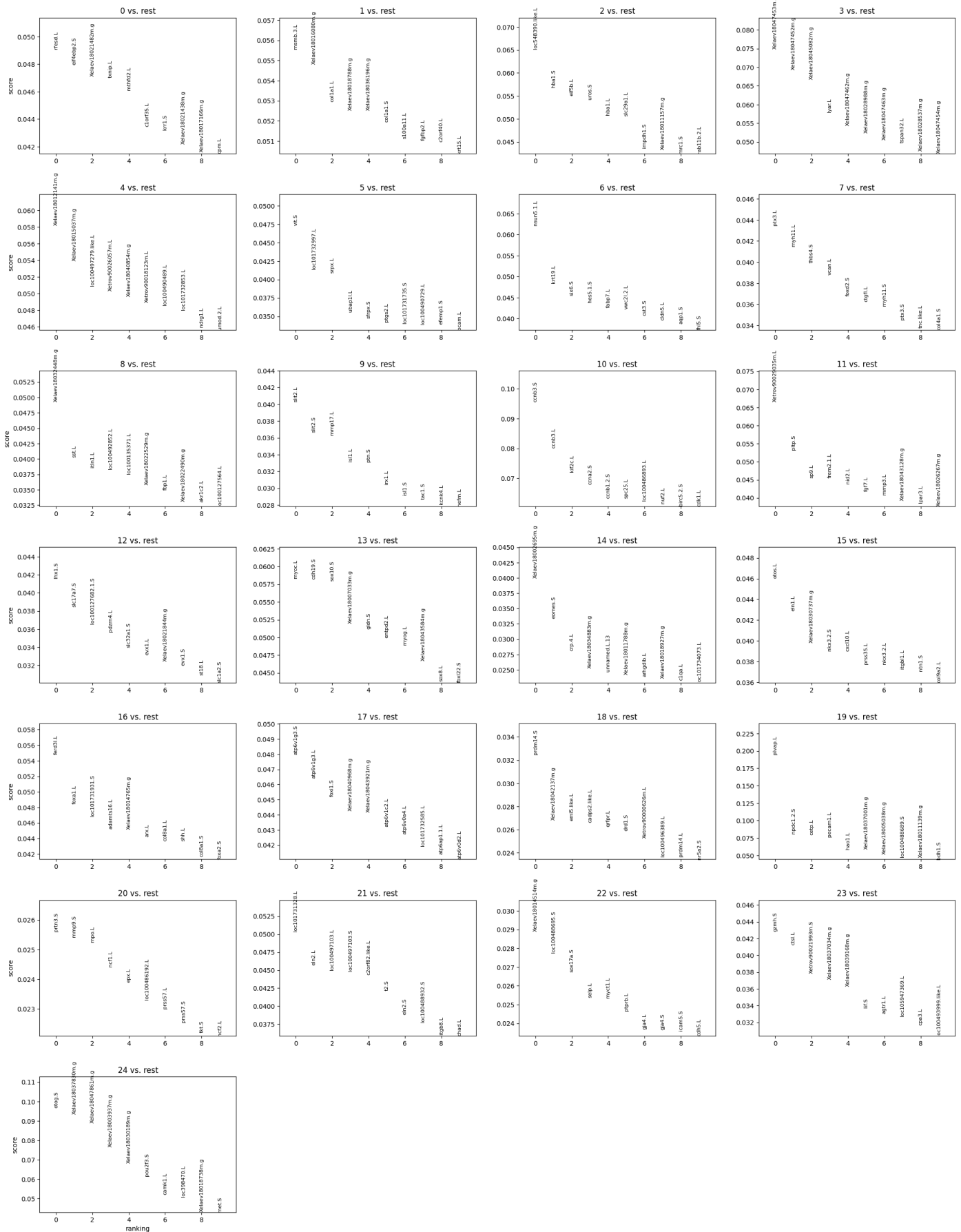
leiden

cluster

```
#Logistic regression to identify marker genes for the Leiden clusters
sc.tl.rank_genes_groups(hvg_adata, 'leiden', method='logreg')

#Top markers for each cluster
sc.pl.rank_genes_groups(hvg_adata, n_genes=10, sharey=False)

sc.pl.rank_genes_groups(hvg_adata, n_genes=10, sharey=False, save='_logreg_markers.png')

marker_genes = sc.get.rank_genes_groups_df(hvg_adata, group='0')
print(marker_genes)
```
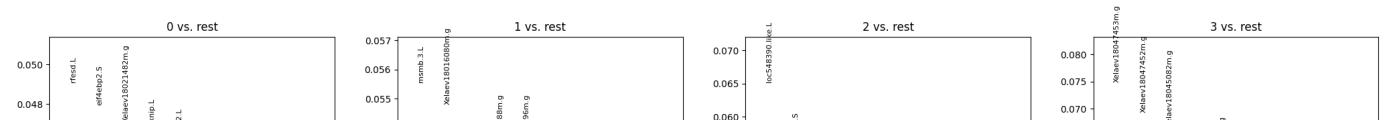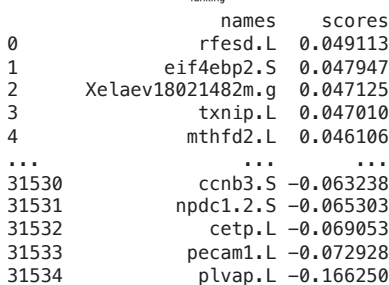
WARNING: saving figure to file figures/rank_genes_groups_leiden_logreg_markers.png

	names	scores
0	rfesd.L	0.049113
1	eif4ebp2.S	0.047947
2	Xelaev18021482m.g	0.047125
3	txnip.L	0.047010
4	mthfd2.L	0.046106
...	...	...
31530	ccnb3.S	-0.063238
31531	npdc1.2.S	-0.065303
31532	cetp.L	-0.069053
31533	pecam1.L	-0.072928
31534	plvap.L	-0.166250

```
    [31535 rows x 2 columns]
```

```python
#kNN-based clustering using Louvain
sc.pp.neighbors(hvg_adata, n_neighbors=10, use_rep='X_pca', method='umap')  # n_neighbors controls k in kNN

#Louvain clustering
sc.tl.louvain(hvg_adata, resolution=0.5)

# Visualize UMAP for kNN + Louvain clustering
sc.tl.umap(hvg_adata)
sc.pl.umap(hvg_adata, color="louvain", title='kNN Clustering with Louvain')

from sklearn.metrics import silhouette_score

#Silhouette score for kNN + Louvain clustering
louvain_labels = hvg_adata.obs['louvain']
knn_sil_score = silhouette_score(hvg_adata.obsm['X_umap'], louvain_labels)
print(f"Silhouette Score for kNN + Louvain Clustering: {knn_sil_score}")
```