In [1]:

```
#Load the libraries

import scanpy as sc
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
import scanpy.external as sce
import seaborn as sns
import anndata as ad
import squidpy as sq
import scvi
from sccoda.util import cell_composition_data as dat
from sccoda.util import data_visualization as viz
import warnings
warnings.filterwarnings("ignore")
from sccoda.util import comp_ana as mod
from scipy import io
from scipy.sparse import coo_matrix, csr_matrix
import os
import anndata
```

```
Global seed set to 0
```

In [2]:

```
# load sparse matrix:
X = io.mmread("D:Scanpy/immunocounts.mtx")
```

In [3]:

```
# create anndata object
adata = anndata.AnnData(
    X=X.transpose().tocsr()
)
```

In [4]:

```
# load cell metadata:
cell_meta = pd.read_csv("D:Scanpy/immunometadata.csv")
```

In [6]:

```
# load gene names:
with open("D:Scanpy/immunogenes.csv", 'r') as f:
    gene_names = f.read().splitlines()
```

In [7]:

```
# set anndata observations and index obs by barcodes, var by gene names
adata.obs = cell_meta
adata.obs.index = adata.obs['barcode']
adata.var.index = gene_names
```

In [9]:

```
adata.layers['counts'] = adata.X.copy()
sc.pp.normalize_total(adata, target_sum=1e4)
sc.pp.log1p(adata)
adata.raw = adata
```

In [10]:

```
adata.obs
```

Out[10]:

barcode	orig.ident	nCount_RNA	nFeature_RNA	patient_ID	ICB	Sex	Age	ICB_response	TKI_exposed	Stage	Biopsy_site	p
AAACCTGAGAATAGGG.p55	SeuratProject	1431	770	55	aPD1	M	57	PR	TKI	IV	Abdomen	
AAACCTGAGGCTAGGT.p55	SeuratProject	1797	865	55	aPD1	M	57	PR	TKI	IV	Abdomen	
AAACCTGCACTGTGTA.p55	SeuratProject	2071	987	55	aPD1	M	57	PR	TKI	IV	Abdomen	
AAACCTGCAGTCCTTC.p55	SeuratProject	682	368	55	aPD1	M	57	PR	TKI	IV	Abdomen	
AAACCTGGTAAATGTG.p55	SeuratProject	2915	1191	55	aPD1	M	57	PR	TKI	IV	Abdomen	
...	...	...	...	...	...	...	...	...	...	...	...	
TTGGAACGTGAGGGAG.p916	SeuratProject	2463	894	916	No_ICB	M	61	NaN	No_TKI	IV	Kidney	
TTGTAGGGTATGAAAC.p916	SeuratProject	7148	2281	916	No_ICB	M	61	NaN	No_TKI	IV	Kidney	
TTTACTGCACACATGT.p916	SeuratProject	5425	1964	916	No_ICB	M	61	NaN	No_TKI	IV	Kidney	
TTTGTCAAGAGCAATT.p916	SeuratProject	4650	1367	916	No_ICB	M	61	NaN	No_TKI	IV	Kidney	
TTTGTCAAGCGTTTAC.p916	SeuratProject	4298	1652	916	No_ICB	M	61	NaN	No_TKI	IV	Kidney	

33749 rows × 15 columns

In [13]:

```
adata.obs['patient_ID'] = adata.obs['patient_ID'].astype('category')
```

In [14]:

```
condition_key = "patient_ID"
```

In [15]:

```
sc.pp.highly_variable_genes(adata, flavor='seurat', n_top_genes=3000, layer='counts', subset=True, batch_key=condition_key)
```

In [16]:

```
scvi.model.SCVI.setup_anndata(adata, layer = "counts", batch_key = condition_key,
                              continuous_covariate_keys=['nCount_RNA', 'percent.mt'])
```

```
INFO     Using batches from adata.obs["patient_ID"]
INFO     No label_key inputted, assuming all cells have same label
INFO     Using data from adata.layers["counts"]
INFO     Successfully registered anndata object containing 33749 cells, 3000 vars, 8 batches, 1 labels, and 0
         proteins. Also registered 0 extra categorical covariates and 2 extra continuous covariates.
INFO     Please do not further modify adata until model is trained.
```

In [17]:

```
model = scvi.model.SCVI(adata, n_layers=2, n_latent=15)
```

In [18]:

```
scvi.data.view_anndata_setup(model.adata)
```

```
Anndata setup with scvi-tools version 0.14.6.
```

Data Summary

Data	Count
Cells	33749
Vars	3000
Labels	1
Batches	8
Proteins	0
Extra Categorical Covariates	0
Extra Continuous Covariates	2

SCVI Data Registry

Data	scvi-tools Location
X	adata.layers['counts']
batch_indices	adata.obs['_scvi_batch']
labels	adata.obs['_scvi_labels']
cont_covs	adata.obsm['_scvi_extra_continuous']

Label Categories

Source Location	Categories	scvi-tools Encoding
adata.obs['_scvi_labels']	0	0

Batch Categories

Source Location	Categories	scvi-tools Encoding
adata.obs['patient_ID']	55	0
	76	1
	90	2
	906	3
	912	4
	913	5
	915	6
	916	7

Extra Continuous Variables

Source Location	Range
adata.obs['nCount_RNA']	501 -> 96922
adata.obs['percent.mt']	0 -> 29.987452948557098864

In [19]:

```
model.train()
```

```
GPU available: False, used: False
TPU available: False, using: 0 TPU cores

Epoch 237/237: 100%|██████████| 237/237 [1:09:20<00:00, 17.56s/it, loss=576, v_num=1]
```

In [20]:

```
adata.obsm['X_scVI'] = model.get_latent_representation()
```

In [21]:

```
adata.layers['scvi_normalized'] = model.get_normalized_expression(library_size = 1e4)
```

In [22]:

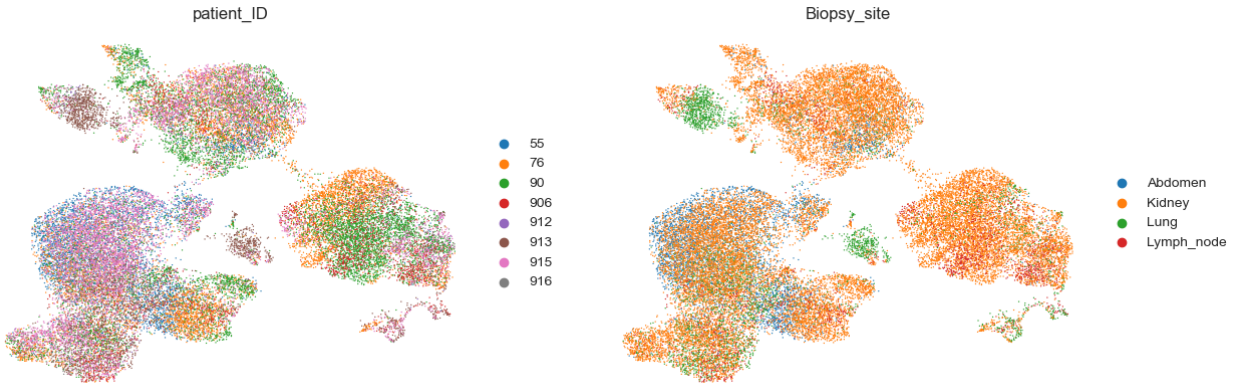
```
sc.pp.neighbors(adata, use_rep = 'X_scVI')
```

In [23]:

```
sc.tl.umap(adata)
```

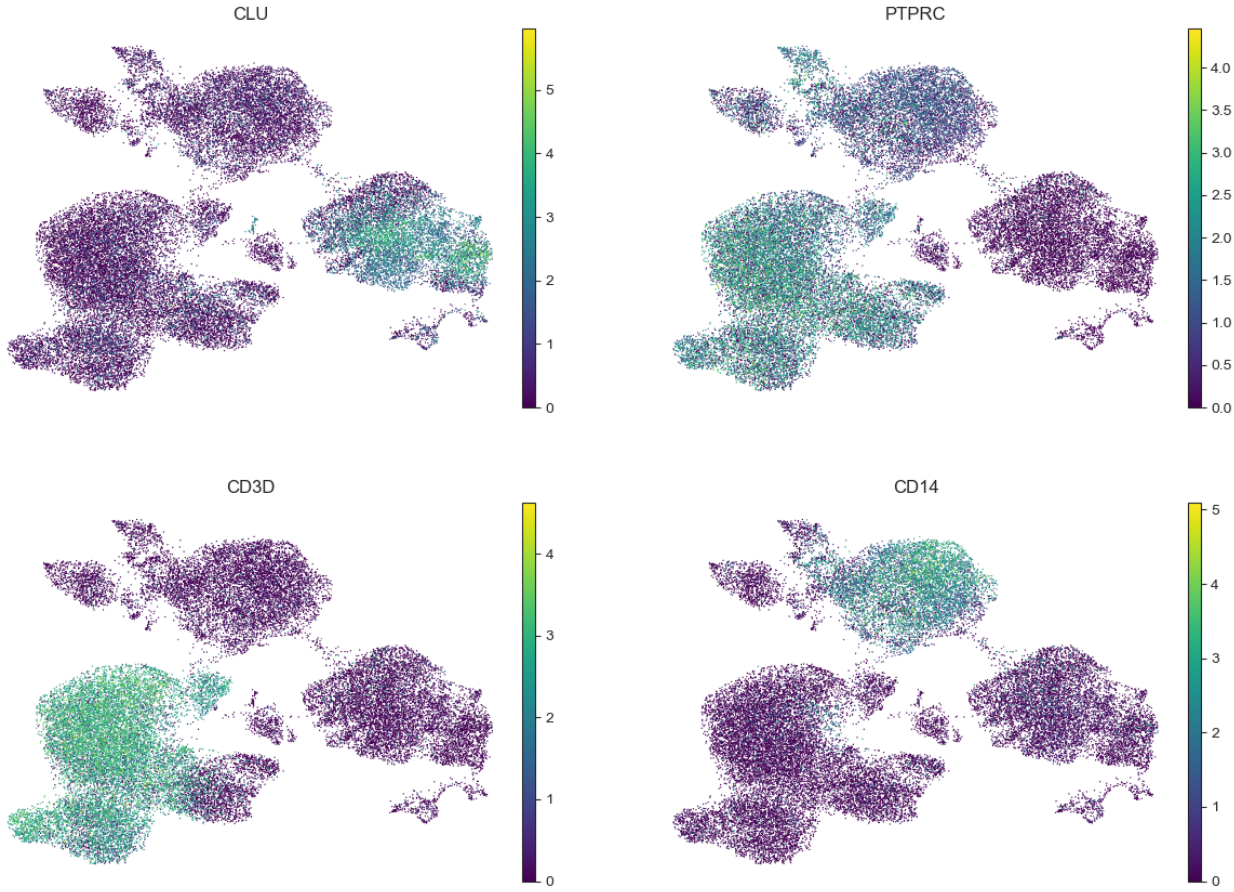
In [25]:

```
sc.pl.umap(adata, color = ['patient_ID', 'Biopsy_site'], frameon=False)
```



In [33]:

```
sc.pl.umap(adata, color=['CLU', 'PTPRC', 'CD3D', 'CD14'], cmap='viridis', frameon=False, ncols = 2)
```

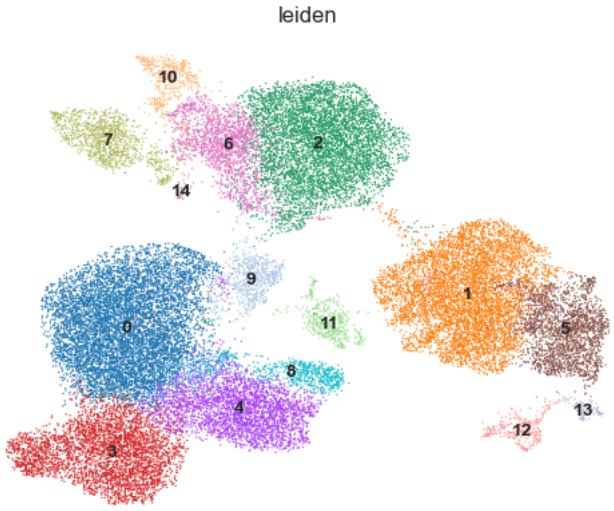


In [38]:

```
sc.tl.leiden(adata, resolution = 0.6)
```

In [39]:

```
sc.pl.umap(adata, color = 'leiden', frameon=False, legend_loc = 'on data')
```



In [40]:

```
sc.tl.rank_genes_groups(adata, 'leiden', method='wilcoxon')
```

In [41]:

```
markers = sc.get.rank_genes_groups_df(adata, None)
markers = markers[(markers.pvals_adj < 0.05) & (markers.logfoldchanges > .5)]
markers
```

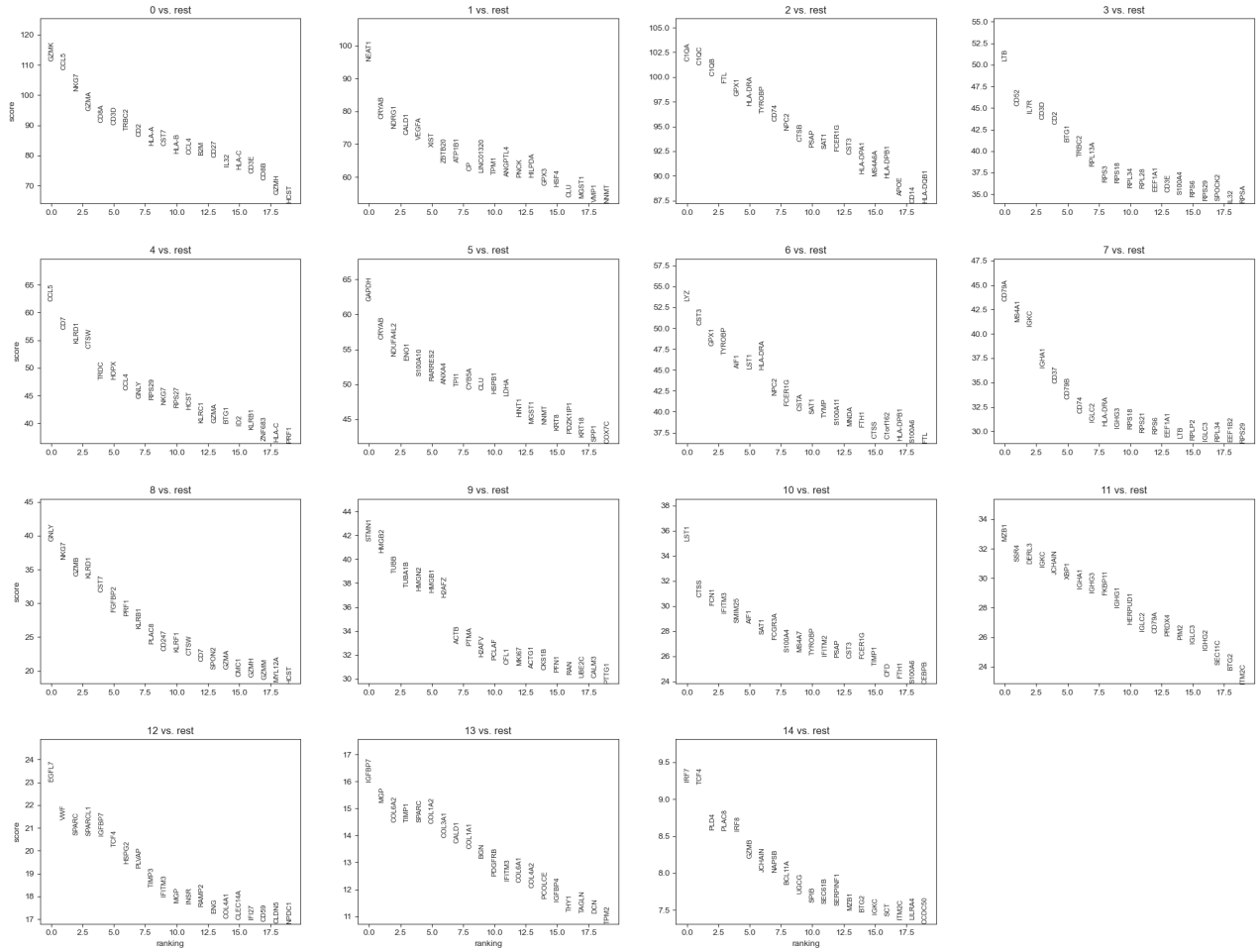
Out[41]:

	group	names	scores	logfoldchanges	pvals	pvals_adj
0	0	GZMK	111.260422	5.594204	0.000000	0.000000
1	0	CCL5	108.393097	4.818508	0.000000	0.000000
2	0	NKG7	101.391846	4.616215	0.000000	0.000000
3	0	GZMA	94.843018	4.009239	0.000000	0.000000
4	0	CD8A	90.991142	4.918987	0.000000	0.000000
...	...	...	...	...	...	...
418566	14	EEF1B2	3.801938	1.110772	0.000144	0.034888
418567	14	ERCC1	3.789707	2.281037	0.000151	0.036356
418568	14	HIST1H2AC	3.778048	3.457105	0.000158	0.037796
418569	14	TRAF4	3.766768	4.008862	0.000165	0.039231
418570	14	GNAS	3.720827	1.496311	0.000199	0.046735

21078 rows × 6 columns

In [42]:

```
sc.pl.rank_genes_groups(adata, n_genes=20, sharey=False)
```



In [43]:

```
adata.obs
```

Out[43]:

barcode	orig.ident	nCount_RNA	nFeature_RNA	patient_ID	ICB	Sex	Age	ICB_response	TKI_exposed	Stage	Biopsy_site	percer
AACCTGAGAATAGGG.p55	SeuratProject	1431	770	55	aPD1	M	57	PR	TKI	IV	Abdomen	3.63
AACCTGAGGCTAGGT.p55	SeuratProject	1797	865	55	aPD1	M	57	PR	TKI	IV	Abdomen	5.45
AACCTGCACTGTGTA.p55	SeuratProject	2071	987	55	aPD1	M	57	PR	TKI	IV	Abdomen	2.36
AACCTGCAGTCCTTC.p55	SeuratProject	682	368	55	aPD1	M	57	PR	TKI	IV	Abdomen	9.82
AACCTGGTAAATGTG.p55	SeuratProject	2915	1191	55	aPD1	M	57	PR	TKI	IV	Abdomen	2.40
...	...	...	...	...	...	...	...	...	...	...	...	
GGAACGTGAGGGAG.p916	SeuratProject	2463	894	916	No_ICB	M	61	NaN	No_TKI	IV	Kidney	3.65
TGTAGGGTATGAAAC.p916	SeuratProject	7148	2281	916	No_ICB	M	61	NaN	No_TKI	IV	Kidney	2.41
TTACTGCACACATGT.p916	SeuratProject	5425	1964	916	No_ICB	M	61	NaN	No_TKI	IV	Kidney	2.00
TTGTCAAGAGCAATT.p916	SeuratProject	4650	1367	916	No_ICB	M	61	NaN	No_TKI	IV	Kidney	3.03
TTGTCAAGCGTTTAC.p916	SeuratProject	4298	1652	916	No_ICB	M	61	NaN	No_TKI	IV	Kidney	4.00

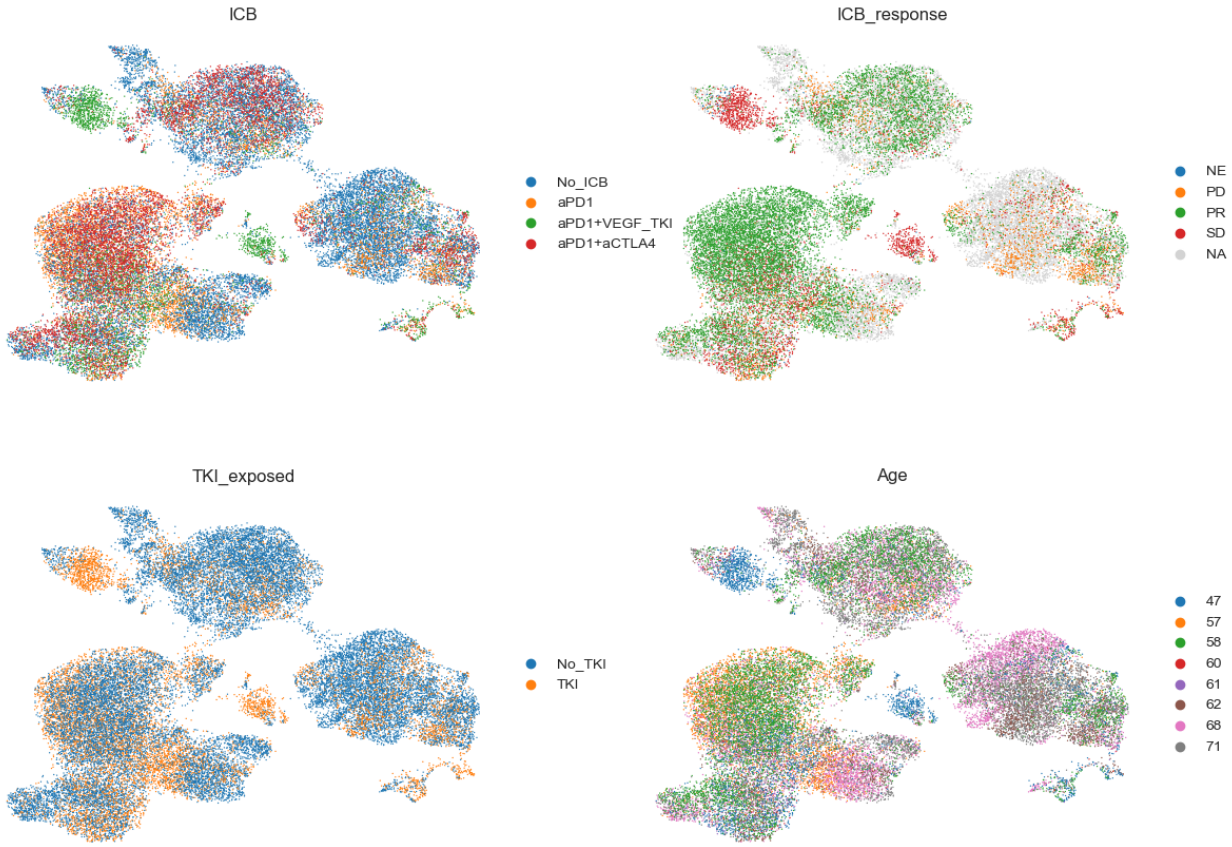
49 rows × 18 columns

In [44]:

```
adata.obs['Age'] = adata.obs['Age'].astype('category')
```

In [46]:

```
sc.pl.umap(adata, color = ['ICB', 'ICB_response', 'TKI_exposed', 'Age'], frameon=False, ncols = 2)
```



In [64]:

```
markers[markers.group=='1']
```

Out[64]:

	group	names	scores	logfoldchanges	pvals	pvals_adj
29890	1	NEAT1	95.279694	3.782570	0.000000	0.000000
29891	1	CRYAB	77.644592	3.681725	0.000000	0.000000
29892	1	NDRG1	74.695465	3.922263	0.000000	0.000000
29893	1	CALD1	72.885139	4.347495	0.000000	0.000000
29894	1	VEGFA	71.427032	4.576474	0.000000	0.000000
...	...	...	...	...	...	...
35148	1	TINAG	2.478877	1.979005	0.013180	0.049583
35149	1	RP11-640M9.2	2.478299	1.893292	0.013201	0.049651
35150	1	FAAP100	2.477372	1.089542	0.013235	0.049774
35152	1	FAHD2CP	2.476359	0.669088	0.013273	0.049903
35153	1	NGEF	2.475907	4.793962	0.013290	0.049960

4555 rows × 6 columns

In [65]:

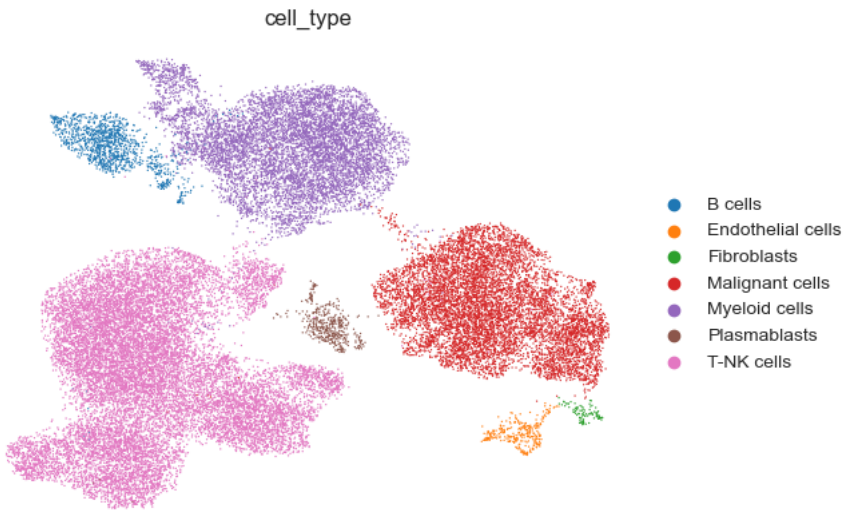
```
cell_type = {"0":"T-NK cells",
"1":"Malignant cells",
"2":"Myeloid cells",
"3":"T-NK cells",
"4":"T-NK cells",
"5":"Malignant cells",
"6":"Myeloid cells",
"7":"B cells",
"8":"T-NK cells",
"9":"T-NK cells",
"10":"Myeloid cells",
"11":"Plasmablasts",
"12":"Endothelial cells",
"13":"Fibroblasts",
"14":"B cells"
}
```

In [66]:

```
adata.obs['cell_type'] = adata.obs.leiden.map(cell_type)
```

In [67]:

```
sc.pl.umap(adata, color = ['cell_type'], frameon = False)
```



In [68]:

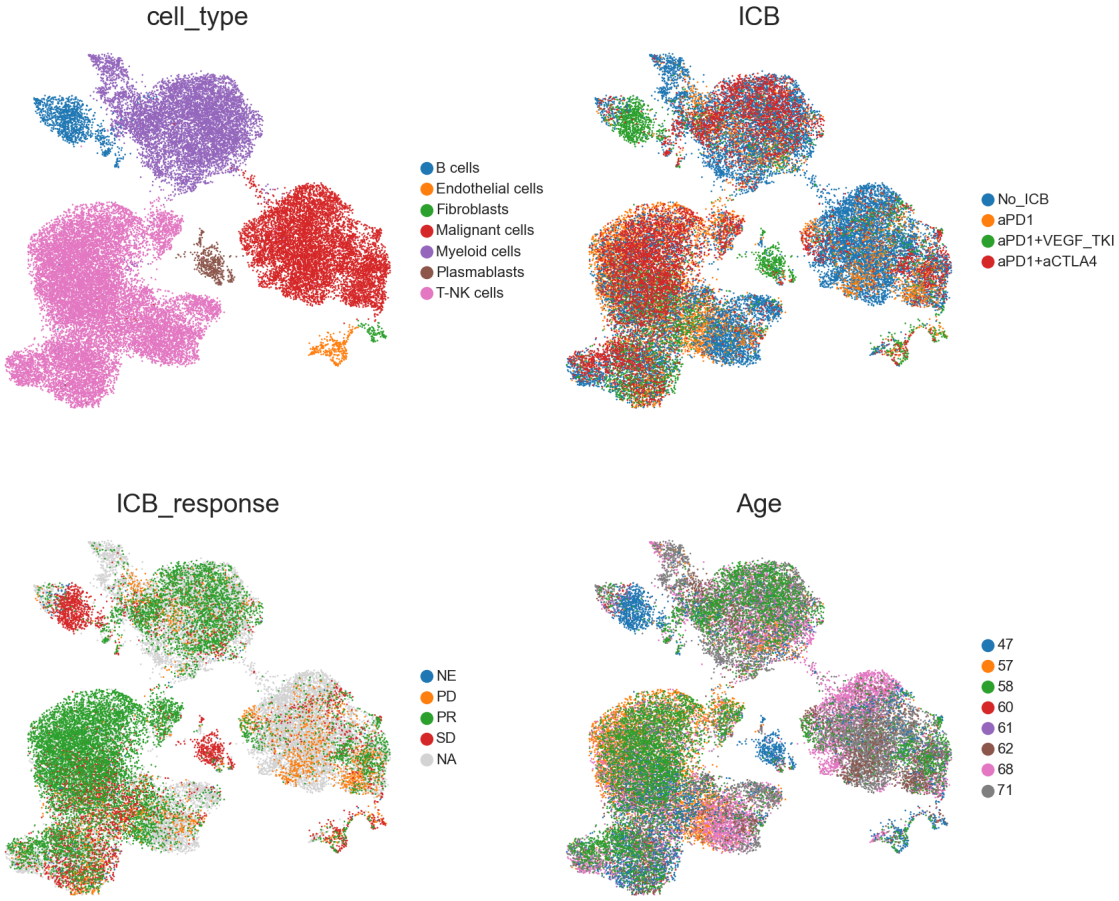
```
adata.uns['markers'] = markers
```

In [69]:

```
sc.set_figure_params(dpi=100)
```

In [73]:

```
sc.pl.umap(adata, color = ['cell_type','ICB', 'ICB_response', 'Age'], frameon=False, ncols = 2, legend_fontsize=8)
```



In [75]:

```
adata.obs.groupby(['patient_ID']).count()
```

Out[75]:

patient_ID	orig.ident	nCount_RNA	nFeature_RNA	ICB	Sex	Age	ICB_response	TKI_exposed	Stage	Biopsy_site	percent.mt	percent.hb	percent.rp
55	4689	4689	4689	4689	4689	4689	4689	4689	4689	4689	4689	4689	4689
76	7900	7900	7900	7900	7900	7900	0	7900	7900	7900	7900	7900	7900
90	8327	8327	8327	8327	8327	8327	0	8327	8327	8327	8327	8327	8327
906	2216	2216	2216	2216	2216	2216	2216	2216	2216	2216	2216	2216	2216
912	208	208	208	208	208	208	208	208	208	208	208	208	208
913	3422	3422	3422	3422	3422	3422	3422	3422	3422	3422	3422	3422	3422
915	6691	6691	6691	6691	6691	6691	6691	6691	6691	6691	6691	6691	6691
916	296	296	296	296	296	296	0	296	296	296	296	296	296

In [76]:

```
num_tot_cells = adata.obs.groupby(['patient_ID']).count()
num_tot_cells = dict(zip(num_tot_cells.index, num_tot_cells.leiden))
num_tot_cells
```

Out[76]:

```
{55: 4689,
 76: 7900,
 90: 8327,
 906: 2216,
 912: 208,
 913: 3422,
 915: 6691,
 916: 296}
```

In [79]:

```
cell_type_counts = adata.obs.groupby(['patient_ID', 'ICB', 'cell_type']).count()
cell_type_counts = cell_type_counts[cell_type_counts.sum(axis = 1) > 0].reset_index()
cell_type_counts = cell_type_counts[cell_type_counts.columns[0:5]]
cell_type_counts
```

Out[79]:

	patient_ID	ICB	cell_type	orig.ident	nCount_RNA
0	55	aPD1	B cells	3	3
1	55	aPD1	Endothelial cells	1	1
2	55	aPD1	Fibroblasts	1	1
3	55	aPD1	Malignant cells	100	100
4	55	aPD1	Myeloid cells	664	664
5	55	aPD1	Plasmablasts	4	4
6	55	aPD1	T-NK cells	3916	3916
7	76	No_ICB	B cells	54	54
8	76	No_ICB	Endothelial cells	55	55
9	76	No_ICB	Malignant cells	2489	2489

In [80]:

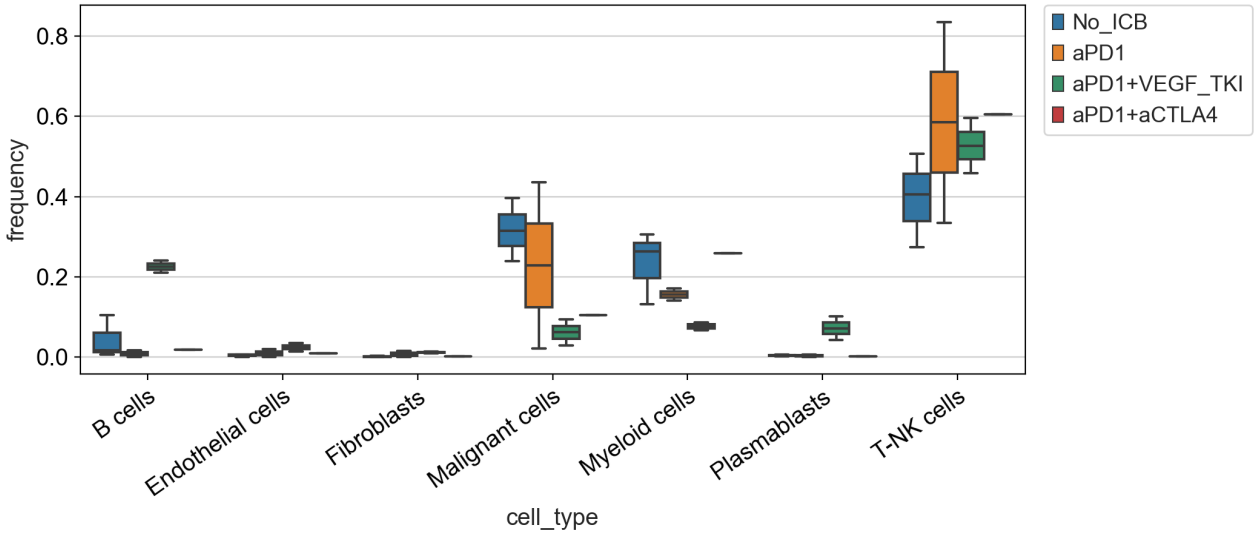
```
cell_type_counts['total_cells'] = cell_type_counts.patient_ID.map(num_tot_cells).astype(int)
cell_type_counts['frequency'] = cell_type_counts.nCount_RNA / cell_type_counts.total_cells
cell_type_counts
```

Out[80]:

	patient_ID	ICB	cell_type	orig.ident	nCount_RNA	total_cells	frequency
0	55	aPD1	B cells	3	3	4689	0.000640
1	55	aPD1	Endothelial cells	1	1	4689	0.000213
2	55	aPD1	Fibroblasts	1	1	4689	0.000213
3	55	aPD1	Malignant cells	100	100	4689	0.021327
4	55	aPD1	Myeloid cells	664	664	4689	0.141608
5	55	aPD1	Plasmablasts	4	4	4689	0.000853
6	55	aPD1	T-NK cells	3916	3916	4689	0.835146
7	76	No_ICB	B cells	54	54	7900	0.006835
8	76	No_ICB	Endothelial cells	55	55	7900	0.006962
9	76	No_ICB	Malignant cells	2489	2489	7900	0.315063

In [84]:

```
sc.set_figure_params(dpi=100)
import matplotlib.pyplot as plt
plt.figure(figsize = (10,4))
ax = sns.boxplot(data = cell_type_counts, x = 'cell_type', y = 'frequency', hue = 'ICB')
plt.xticks(rotation = 35, rotation_mode = 'anchor', ha = 'right')
plt.legend(bbox_to_anchor=(1.02, 1), loc='upper left', borderaxespad=0)
plt.show()
```



In [85]:

```
cell_type_counts = adata.obs.groupby(['patient_ID', 'ICB_response', 'cell_type']).count()
cell_type_counts = cell_type_counts[cell_type_counts.sum(axis = 1) > 0].reset_index()
cell_type_counts = cell_type_counts[cell_type_counts.columns[0:5]]
cell_type_counts
```

Out[85]:

	patient_ID	ICB_response	cell_type	orig.ident	nCount_RNA
0	55	PR	B cells	3	3
1	55	PR	Endothelial cells	1	1
2	55	PR	Fibroblasts	1	1
3	55	PR	Malignant cells	100	100
4	55	PR	Myeloid cells	664	664
5	55	PR	Plasmablasts	4	4
6	55	PR	T-NK cells	3916	3916
7	906	PD	B cells	38	38
8	906	PD	Endothelial cells	43	43
9	906	PD	Fibroblasts	33	33

In [86]:

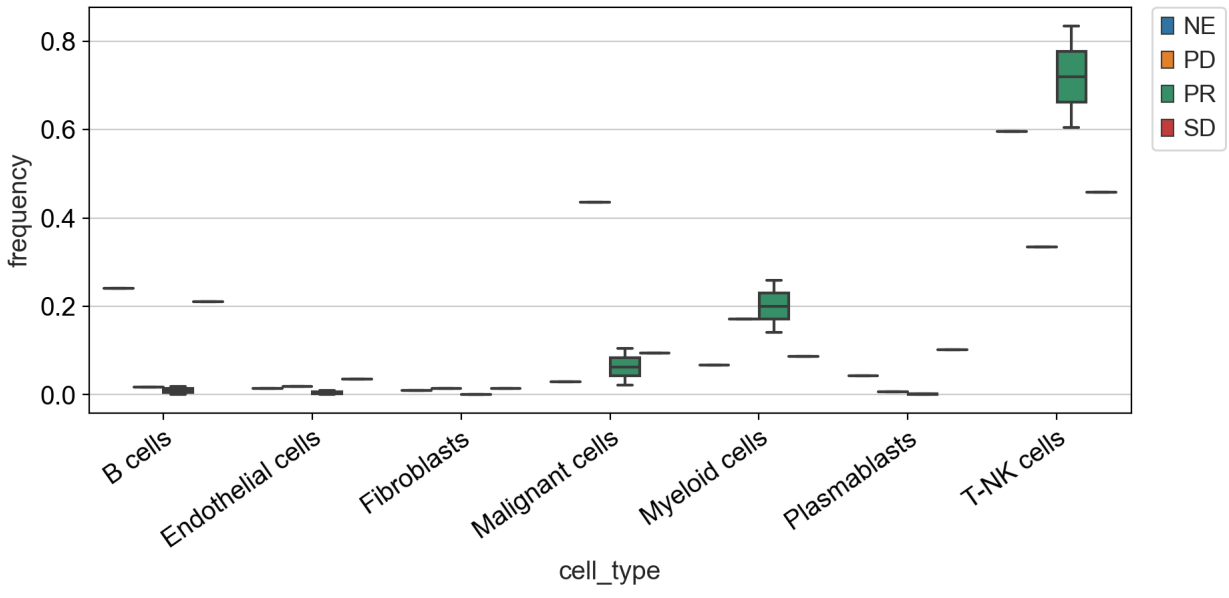
```
cell_type_counts['total_cells'] = cell_type_counts.patient_ID.map(num_tot_cells).astype(int)
cell_type_counts['frequency'] = cell_type_counts.nCount_RNA / cell_type_counts.total_cells
cell_type_counts
```

Out[86]:

	patient_ID	ICB_response	cell_type	orig.ident	nCount_RNA	total_cells	frequency
0	55	PR	B cells	3	3	4689	0.000640
1	55	PR	Endothelial cells	1	1	4689	0.000213
2	55	PR	Fibroblasts	1	1	4689	0.000213
3	55	PR	Malignant cells	100	100	4689	0.021327
4	55	PR	Myeloid cells	664	664	4689	0.141608
5	55	PR	Plasmablasts	4	4	4689	0.000853
6	55	PR	T-NK cells	3916	3916	4689	0.835146
7	906	PD	B cells	38	38	2216	0.017148
8	906	PD	Endothelial cells	43	43	2216	0.019404
9	906	PD	Fibroblasts	33	33	2216	0.014892

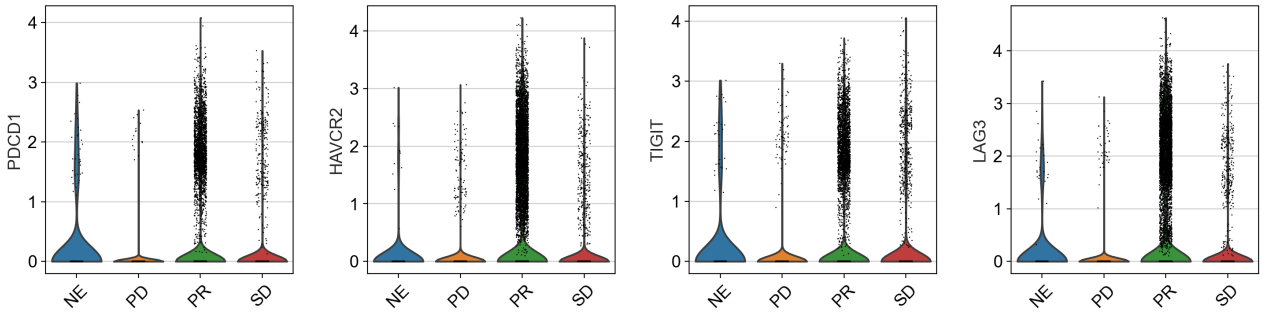
In [87]:

```
import matplotlib.pyplot as plt
plt.figure(figsize = (10,4))
ax = sns.boxplot(data = cell_type_counts, x = 'cell_type', y = 'frequency', hue = 'ICB_response')
plt.xticks(rotation = 35, rotation_mode = 'anchor', ha = 'right')
plt.legend(bbox_to_anchor=(1.02, 1), loc='upper left', borderaxespad=0)
plt.show()
```



In [90]:

```
sc.pl.violin(adata, ['PDCD1', 'HAVCR2', 'TIGIT', 'LAG3'], groupby='ICB_response', rotation=45)
```



In [91]:

```
adata.write_h5ad('D:/Scanpy/scImmuno.h5ad')
model.save('D:/Scanpy/scImmunomodel.model')
```