

BME-230B Spring 2019 HW 2 Question 2

James Casaletto, Andrew Davidson, Yuanqing Xue, Jim Zheng

- ref
 - [scanpy.tl.umap](https://icb-scanpy.readthedocs-hosted.com/en/stable/api/scanpy.tl.umap.html) (<https://icb-scanpy.readthedocs-hosted.com/en/stable/api/scanpy.tl.umap.html>)
 - [scanpy.api.pp.neighbors](https://icb-scanpy.readthedocs-hosted.com/en/stable/api/scanpy.api.pp.neighbors.html?highlight=neighbors) (<https://icb-scanpy.readthedocs-hosted.com/en/stable/api/scanpy.api.pp.neighbors.html?highlight=neighbors>)
 - [scanpy.pl.umap](https://icb-scanpy.readthedocs-hosted.com/en/stable/api/scanpy.pl.umap.html#scanpy.pl.umap) (<https://icb-scanpy.readthedocs-hosted.com/en/stable/api/scanpy.pl.umap.html#scanpy.pl.umap>)
 - [scanpy.tl.louvain](https://icb-scanpy.readthedocs-hosted.com/en/stable/api/scanpy.tl.louvain.html#scanpy.tl.louvain) (<https://icb-scanpy.readthedocs-hosted.com/en/stable/api/scanpy.tl.louvain.html#scanpy.tl.louvain>)
 - [GSEAPY: Gene Set Enrichment Analysis in Python](https://pypi.org/project/gseapy/). [pypi.org](https://pypi.org/project/gseapy/) (<https://pypi.org/project/gseapy/>)
 - [GSEAPY: Gene Set Enrichment Analysis in Python](https://gseapy.readthedocs.io/en/latest/introduction.html) [gseapy.readthedocs.io](https://gseapy.readthedocs.io/en/latest/introduction.html) (<https://gseapy.readthedocs.io/en/latest/introduction.html>)
 - [anndata](https://anndata.readthedocs.io/en/latest/anndata.AnnData.html) (<https://anndata.readthedocs.io/en/latest/anndata.AnnData.html>)
 - "uns" stands for unstructured data
 - "obs" are panda data frame observations
 - "obsm key-indexed multi-dimensional observations
 - [Hypergeometric distribution](https://en.wikipedia.org/wiki/Hypergeometric_distribution) (https://en.wikipedia.org/wiki/Hypergeometric_distribution)
 - [Hypergeometric Tests for Gene Lists](http://users.unimi.it/marray/2007/material/day4/Lecture7.pdf) (<http://users.unimi.it/marray/2007/material/day4/Lecture7.pdf>)

2.a. [10 pts] Turn in a copy of your code.

Your method should take in the 12-NNGs that you calculated above. Note that the pseudocode above returns solutions at all levels while your method should only produce a single level, the one that you obtain when no new communities are formed and the outer loop is exited. Although not required, we recommend using the iGraph or NetworkX module to develop the Louvain algorithm. You can use the provided `knn_to_graphModule.py` to create an `igraph` or `networkX` object of the k-NNG you previously made in 1.a.

Cluster both of the batches in the PBMC dataset using your own Louvain algorithm. Feel free to compare the results to running the Louvain algorithm provided by Scanpy (you can see examples of it in the scanpy tutorial under Clustering. Note that this scanpy louvain method is actually a slightly modified version that has an additional parameter called the resolution that can determine the coarseness of the community structure. Use resolution =1 to make it compatible with the original Louvain optimization function.

All louvain implementation is in separate notebook [BME-230b-Spring-2019-hw2_question2a_louvain.ipynb](#) ([BME-230b-Spring-2019-hw2_question2a_louvain.ipynb](#))

NOTE: for 2b, 2c, and 2d, we used Scanpy's implementation of Louvain

```
In [1]: from euclid_knn import KnnG
import gseapy as gp
import matplotlib.pyplot as plt
import numpy as np
import os

import pandas as pd
import scanpy.api as sc
import scanpy
print("scanpy.__version__:{}".format(scanpy.__version__))

import scipy.special
import scipy.stats as stats

scanpy.__version__:1.4
```

2.b. [5 pts]

Turn in a UMAP plot of the combined dataset as you did in question #1, but this time, color the cells by their Louvain cluster assignments determined for each cell within each batch as a different color in each plot [2 pts: UMAP]. Also report the modularity of the partition you obtained on the combined dataset [3 pts: Reporting Modularity].

```
In [2]: %%time
anndata = sc.read("PBMC.merged.h5ad")

CPU times: user 4.2 s, sys: 134 ms, total: 4.33 s
Wall time: 1.53 s
```

```
In [3]: # run our implementation of nearest neighbors and update anndata
KnnG(anndata, n_neighbors=12, runPCA=True, nPC=50)
```

```
Out[3]: <euclid_knn.KnnG at 0x7f763c1d4b00>
```

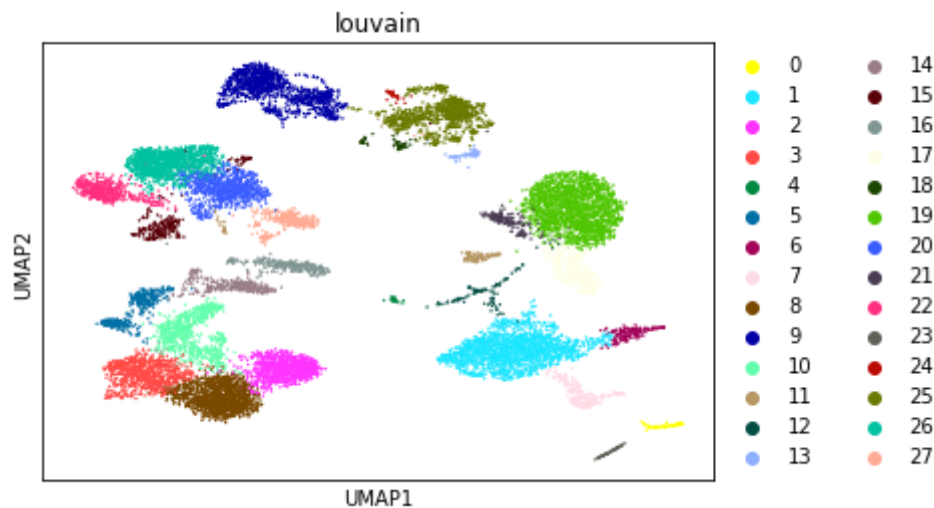
```
In [4]: %%time
# running Scanpy's version of Louvian
scanpy.tl.louvain(anndata,
                  flavor='igraph',
                  directed=False,
                  use_weights=True)

CPU times: user 7.09 s, sys: 53.6 ms, total: 7.14 s
Wall time: 1.35 s
```

```
In [5]: scanpy.tl.umap(anndata)
```

```
In [6]: plt.figure(figsize=(10,10))
scanpy.pl.umap(anndata, color=["louvain"])
```

<Figure size 720x720 with 0 Axes>



2.c. [5 pts]

Turn in a table that lists each cluster and its best-matching cell type annotation. The table should contain the cluster number and its best matching cell-type annotation based on the hypergeometric analysis.

```
In [12]: import hw2q2
```

```
In [13]: cellCountsByClusterId, cellTypesInClusters = hw2q2.createCountsDict(anndata)
```

```
In [16]: annotations = hw2q2.bestAnnotation(anndata, cellTypesInClusters, cellCounts)
          annotations.sort_values(by=['clusterId'])
```

```
Out[16]:
```

	clusterId	Cell type	p-value
0	0	Plasmacytoid dendritic cell	0.000000e+00
1	1	Monocyte_CD14	0.000000e+00
2	2	CD8 T cell	3.223022e-11
3	3	CD4 T cell	0.000000e+00
4	4	Hematopoietic stem cell	0.000000e+00
5	5	CD8 T cell	1.956746e-11
6	6	Monocyte_FCGR3A	0.000000e+00
7	7	Monocyte_CD14	0.000000e+00
8	8	CD4 T cell	7.790102e-12
9	9	B cell	0.000000e+00
10	10	CD8 T cell	2.847245e-11
11	11	Monocyte_CD14	3.811396e-12
12	12	Megakaryocyte	1.195422e-11
13	13	B cell	0.000000e+00
14	14	NK cell	0.000000e+00
15	15	CD8 T cell	0.000000e+00
16	16	NK cell	0.000000e+00
17	17	Monocyte_FCGR3A	0.000000e+00
18	18	B cell	0.000000e+00
19	19	Monocyte_CD14	0.000000e+00
20	20	CD4 T cell	1.010458e-11
21	21	Monocyte_CD14	0.000000e+00
22	22	CD8 T cell	0.000000e+00
23	23	Plasmacytoid dendritic cell	0.000000e+00
24	24	B cell	0.000000e+00
25	25	B cell	0.000000e+00
26	26	CD4 T cell	1.193878e-11
27	27	CD8 T cell	5.398015e-12

2.d. [5 pts]

Turn in a list of top 5 pathways for each cluster in each dataset. You should use the gene expression signature of each cluster to find an associated pathway. A gene signature for a cluster represents the gene expression levels for a characteristic cell that is a member of the cluster. Use

the centroid μ_i of the i th cluster as the signature. Compute the centroids for each cluster in each dataset. You will next derive a gene-signature based annotation for each cluster using these centroids. Use a list of Gene Ontology Biological Process categories (provided in the Resources section at the top of this homework) and your signatures to perform an all-against-all Gene Set Enrichment Analysis (GSEA). Turn in a table that lists the top 5 pathways for each cluster

```
In [18]: clusterSigs = hw2q2.calculateClusterSignatures(anndata)
```

```
In [19]: %%time
pathways = hw2q2.rankPathWays(anndata, clusterSigs, topN=5)
```

```
CPU times: user 2min 46s, sys: 8.38 s, total: 2min 54s
Wall time: 9min 33s
```

```
In [20]: # https://stackoverflow.com/a/35693013/4586180
# display data frame with out index column
from IPython.display import display, HTML
display(HTML(pathways.to_html(index=False)))
```

Term	nes	cluster id
cellular protein metabolic process (GO:0044267)	1.333285	0
Fc-epsilon receptor signaling pathway (GO:0038...	1.323810	0
antigen processing and presentation of exogeno...	1.325697	0
antigen processing and presentation of peptide...	1.352975	0
positive regulation of phosphorylation (GO:004...	1.361454	0
neutrophil degranulation (GO:0043312)	2.059123	1
neutrophil activation involved in immune respo...	2.078532	1
neutrophil mediated immunity (GO:0002446)	2.107963	1
cellular protein metabolic process (GO:0044267)	1.827333	1
granulocyte chemotaxis (GO:0071621)	1.694330	1
T cell activation (GO:0042110)	2.452248	2
complement activation, classical pathway (GO:0...	1.699152	2
transmembrane receptor protein tyrosine kinase...	1.610634	2
enzyme linked receptor protein signaling pathw...	1.477404	2
response to cytokine (GO:0034097)	1.284005	2
regulation of protein kinase B signaling (GO:0...	1.595247	3
T cell receptor signaling pathway (GO:0050852)	1.599044	3
tumor necrosis factor-mediated signaling pathw...	1.802821	3
positive regulation of GTPase activity (GO:004...	1.579765	3
positive regulation of protein kinase B signal...	1.567524	3
positive regulation of gene expression (GO:001...	1.605392	4
positive regulation of nucleic acid-templated ...	1.758239	4
positive regulation of cellular biosynthetic p...	1.674370	4
negative regulation of nucleic acid-templated ...	1.671713	4
regulation of transcription, DNA-templated (GO...	1.664383	4
protein oligomerization (GO:0051259)	1.430223	5
cellular defense response (GO:0006968)	1.400160	5
regulation of immune response (GO:0050776)	1.720955	5
positive regulation of hydrolase activity (GO:...	1.436126	5
inflammatory response (GO:0006954)	1.403699	5
neutrophil degranulation (GO:0043312)	1.616853	6
neutrophil mediated immunity (GO:0002446)	1.622177	6
neutrophil activation involved in immune respo...	1.653849	6

Term	nes	cluster id
antigen processing and presentation of peptide...	1.571275	6
antigen processing and presentation of exogeno...	1.577903	6
antigen processing and presentation of exogeno...	1.855222	7
antigen processing and presentation of peptide...	1.814602	7
cellular response to interferon-gamma (GO:0071...	1.794303	7
interferon-gamma-mediated signaling pathway (G...	1.770483	7
antigen processing and presentation of exogeno...	1.757600	7
T cell activation (GO:0042110)	2.312814	8
complement activation, classical pathway (GO:0...	1.697793	8
positive regulation of protein kinase B signal...	1.363084	8
regulation of acute inflammatory response (GO:...	1.272369	8
humoral immune response mediated by circulatin...	1.323808	8
B cell receptor signaling pathway (GO:0050853)	2.208003	9
positive regulation of B cell activation (GO:0...	2.012516	9
positive regulation of lymphocyte activation (...)	2.003641	9
antigen receptor-mediated signaling pathway (G...	2.173027	9
humoral immune response mediated by circulatin...	2.036609	9
regulation of GTPase activity (GO:0043087)	1.827392	10
positive regulation of GTPase activity (GO:004...	1.794914	10
positive regulation of hydrolase activity (GO:...	1.781682	10
T cell activation (GO:0042110)	1.671555	10
regulation of immune response (GO:0050776)	1.576390	10
antigen processing and presentation of exogeno...	1.784130	11
antigen processing and presentation of exogeno...	1.792342	11
T cell receptor signaling pathway (GO:0050852)	1.593702	11
interferon-gamma-mediated signaling pathway (G...	1.539432	11
antigen processing and presentation of peptide...	1.751178	11
response to molecule of bacterial origin (GO:0...	1.388828	12
positive regulation of leukocyte chemotaxis (G...	1.406809	12
platelet degranulation (GO:0002576)	1.433550	12
muscle contraction (GO:0006936)	1.454998	12
regulated exocytosis (GO:0045055)	1.519884	12
regulation of immune effector process (GO:0002...	1.737337	13
interferon-gamma-mediated signaling pathway (G...	1.728635	13
regulation of complement activation (GO:0030449)	1.712084	13
antigen processing and presentation of exogeno...	1.698735	13

Term	nes	cluster id
regulation of B cell activation (GO:0050864)	1.699109	13
positive regulation of hydrolase activity (GO:...	1.609459	14
positive regulation of ERK1 and ERK2 cascade (...)	1.568272	14
regulation of GTPase activity (GO:0043087)	1.565476	14
positive regulation of GTPase activity (GO:004...	1.551697	14
regulation of immune response (GO:0050776)	1.640011	14
T cell activation (GO:0042110)	1.695259	15
negative regulation of cytokine production (GO...	1.488782	15
regulation of GTPase activity (GO:0043087)	1.562767	15
positive regulation of GTPase activity (GO:004...	1.595306	15
positive regulation of hydrolase activity (GO:...	1.605712	15
positive regulation of hydrolase activity (GO:...	1.529969	16
positive regulation of GTPase activity (GO:004...	1.494077	16
response to interferon-gamma (GO:0034341)	1.414437	16
regulation of GTPase activity (GO:0043087)	1.538661	16
cellular defense response (GO:0006968)	1.458698	16
neutrophil mediated immunity (GO:0002446)	1.625153	17
antigen processing and presentation of exogeno...	1.510699	17
regulated exocytosis (GO:0045055)	1.472575	17
neutrophil degranulation (GO:0043312)	1.611826	17
neutrophil activation involved in immune respo...	1.624471	17
regulation of protein processing (GO:0070613)	1.774148	18
Fc-gamma receptor signaling pathway involved i...	1.743001	18
Fc-epsilon receptor signaling pathway (GO:0038...	1.743197	18
regulation of complement activation (GO:0030449)	1.753462	18
humoral immune response mediated by circulatin...	1.806558	18
extracellular matrix organization (GO:0030198)	1.516295	19
positive regulation of NF-kappaB transcription...	1.551196	19
neutrophil degranulation (GO:0043312)	1.796621	19
neutrophil mediated immunity (GO:0002446)	1.819045	19
neutrophil activation involved in immune respo...	1.837040	19
positive regulation of gene expression (GO:001...	1.616762	20
regulation of transcription, DNA-templated (GO...	1.638333	20
regulation of transcription from RNA polymeras...	1.693520	20
regulation of programmed cell death (GO:0043067)	1.712579	20
tumor necrosis factor-mediated signaling pathw...	1.716800	20

Term	nes	cluster id
neutrophil degranulation (GO:0043312)	1.710914	21
neutrophil activation involved in immune respo...	1.706809	21
neutrophil mediated immunity (GO:0002446)	1.696067	21
cellular response to lipopolysaccharide (GO:00...	1.500996	21
regulation of cytokine production (GO:0001817)	1.481354	21
regulation of cellular macromolecule biosynthe...	1.459647	22
T cell activation (GO:0042110)	2.110025	22
regulation of cell cycle (GO:0051726)	1.500324	22
transmembrane receptor protein tyrosine kinase...	1.597967	22
enzyme linked receptor protein signaling pathw...	1.713461	22
interferon-gamma-mediated signaling pathway (G...	1.337556	23
regulation of B cell proliferation (GO:0030888)	1.345112	23
regulation of protein phosphorylation (GO:0001...	1.357005	23
positive regulation of phosphorylation (GO:004...	1.408704	23
Fc receptor signaling pathway (GO:0038093)	1.381194	23
regulation of B cell activation (GO:0050864)	2.035837	24
complement activation, classical pathway (GO:0...	1.948536	24
phagocytosis, engulfment (GO:0006911)	1.922344	24
positive regulation of lymphocyte activation (...)	1.886077	24
B cell receptor signaling pathway (GO:0050853)	1.918055	24
regulation of protein activation cascade (GO:2...	1.937766	25
regulation of immune effector process (GO:0002...	1.954491	25
regulation of complement activation (GO:0030449)	1.973642	25
humoral immune response mediated by circulatin...	1.981524	25
regulation of humoral immune response (GO:0002...	2.039666	25
regulation of transcription from RNA polymeras...	1.395068	26
T cell activation (GO:0042110)	1.466689	26
protein complex assembly (GO:0006461)	1.678697	26
regulation of cell cycle (GO:0051726)	1.349217	26
regulation of cellular macromolecule biosynthe...	1.411476	26
regulation of transcription, DNA-templated (GO...	1.473158	27
negative regulation of transcription, DNA-temp...	1.490239	27
positive regulation of nucleic acid-templated ...	1.535971	27
regulation of transcription from RNA polymeras...	1.599694	27
negative regulation of transcription from RNA ...	1.725175	27

