

資料科學導論期末專題

降維工具實作 - 以基因表現量資料為例

組名：科科

組員：生科110 林耿弘、周治瑗、蔡秉勳

Brief introduction of the problem

Motivations:

本報告主要使用UMAP此降維工具，該工具為近幾年生物資訊領域非常熱門的降維方法，因此希望藉由此期末專題來學習UMAP的原理與實作。並且找尋一個感興趣的生物議題來探討。

Problem Statement:

Are nuclear pore complexes (NPCs) tissue-specific?

近期小組成員找到一篇感興趣的論文，主要是描述本次分析對象NPC具有組織特異性(Tissue-specific)。

Published: 23 October 2012

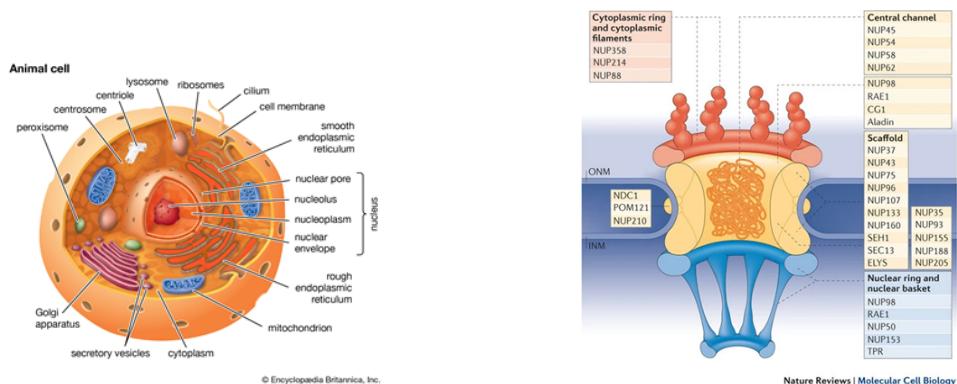
Nuclear pore complex composition: a new regulator of tissue-specific and developmental functions

Marcela Raices & Maximiliano A. D'Angelo 

Nature Reviews Molecular Cell Biology 13, 687–699 (2012) | Cite this article

2214 Accesses | 186 Citations | 3 Altmetric | Metrics

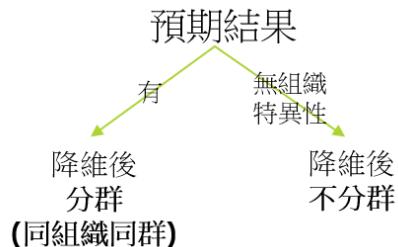
NPC為細胞核的核孔構造，主要功能為控制物質的進出，總共有31種不同的核孔蛋白所組成。



在過去科學家普遍認為NPC具有高度保守性，亦即在人體不同組織的細胞中的NPC中構造一樣，然而，近幾年來有越來越多研究發現NPC不像過去所認為的如此恆定

不變，有多個證據指出NPC有組織上的特異性，也就是說在不同組織細胞中的NPC構造可能不一樣。

因此，使用31NPC genes的基因表現量資料，將每筆31維度的資料進行降維(31D -> 2D)，並進行可視化，然後看sample是否與同組織的samples形成同一Cluster，和不同組織的Clusters分離。



若有分群，在一定程度上可以說NPC有組織特異性，否則無組織特異性。

Data description and preprocessing

本報告使用The Genotype-Tissue Expression (GTEx)資料庫資料，該資料庫為收錄正常組織的基因表達資料，在本研究中下載Gene TPMs資料集：

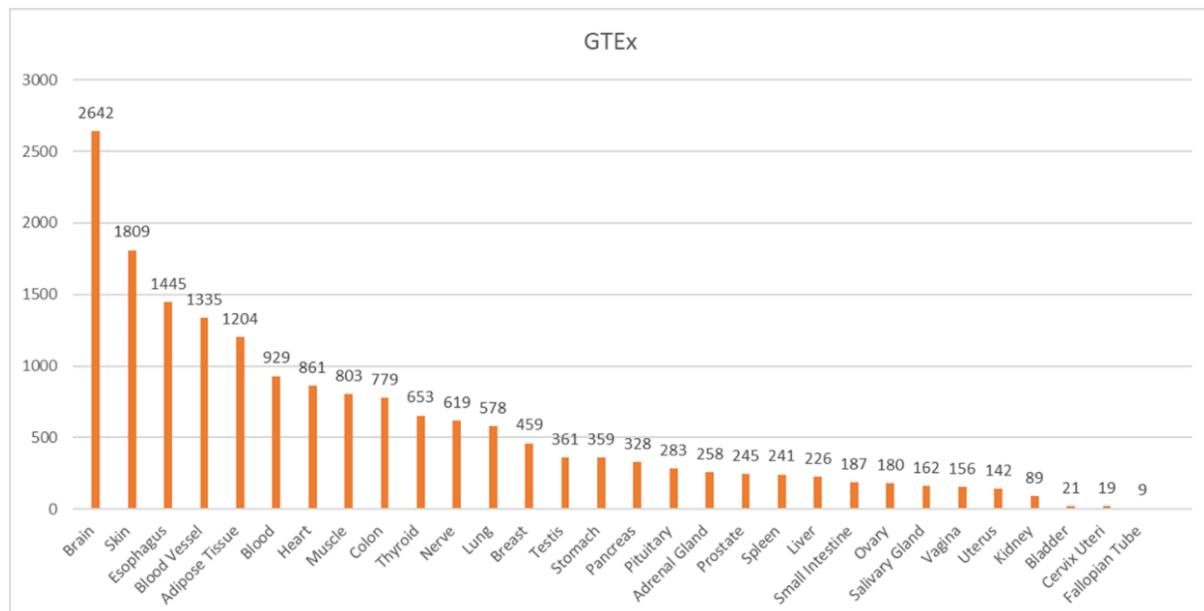
The screenshot shows the GTEx Portal interface. At the top, there's a navigation bar with links for Home, Datasets, Expression, QTLs & Browsers, Sample Data, and Documentation. Below the navigation bar, a message encourages users to take a survey: "Please take our very brief user survey to help us plan new features for the GTEx Portal: <http://bit.ly/2021glexsurvey>". The main content area is titled "RNA-Seq Data". It displays a table with the following columns: Description, Name, and Size. The table lists several files related to Gene TPMs analysis:

Description	Name	Size
Gene read counts.	GTEX_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_reads.gct.gz	875M
Gene TPMs.	GTEX_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_tpm.gct.gz	1.6G
Median gene-level TPM by tissue. Median expression was calculated from the file GTEX_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_tpm.gct.gz.	GTEX_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_median_tpm.gct.gz	6.7M
Exon-exon junction read counts.	GTEX_Analysis_2017-06-05_v8_STARv2.5.3a_junctions.gct.gz	4.0G
Transcript read counts.	GTEX_Analysis_2017-06-05_v8_RSEMv1.3.0_transcript_expected_count.gct.gz	4.4G
Transcript TPMs.	GTEX_Analysis_2017-06-05_v8_RSEMv1.3.0_transcript_tpm.gct.gz	3.5G
Exon read counts.	GTEX_Analysis_2017-06-05_v8_RNASeQCv1.1.9_exon_reads.parquet	11G

		Sample 1	Sample 2		
Name	Description	GTEX-1117F-0226-SM-5GZZ7	GTEX-1117F-0003-SM-5DWSB	...	Sample 17382
ENSG00000 223972.5	DDX11L1	0	0	...	0.02522
ENSG00000 227232.5	WASH7P	8.764	3.861	...	2.167
...
ENSG00000 210196.2	MT-TP	0	6.226	...	0.6413

Gene number: 56200

下載Gene TPMs後打開該檔案為上方表格形式文字檔，因為樣本無對應的組織，因此需再下載另一檔案有詳列各個樣本所對應的組織，將兩個檔案整合後，可以得到以下個個組織的樣本數：



總共有30種組織，17382個樣本數。

接著將我們所需要的相關31個NPC基因給取出：

	1	2	3	4	5	6	7	8	9	10	...	22	23	24	25	26	27	28	29	30	31
0																					
Muscle	22.440	21.020	27.23	12.500	7.274	14.900	25.380	12.110	7.602	15.11	...	5.570	3.067	9.301	7.289	5.098	23.55	21.510	2.437	11.100	4.451
Muscle	34.900	19.710	40.72	11.690	9.197	22.330	31.920	11.340	8.941	28.91	...	5.521	3.149	16.580	12.360	8.075	30.78	23.760	8.515	10.270	1.362
Muscle	9.916	5.068	11.47	5.559	3.262	9.701	9.638	4.853	3.289	20.17	...	2.354	1.545	5.905	4.374	2.282	9.34	8.357	1.310	5.002	3.307
Muscle	15.030	7.153	16.64	9.087	6.092	13.290	15.710	5.733	7.948	20.80	...	4.102	2.685	6.954	6.914	2.847	15.46	13.670	2.543	5.461	5.133
Muscle	33.680	14.540	26.52	10.910	11.150	21.830	29.890	8.330	8.930	15.87	...	4.080	3.345	16.450	12.700	6.291	31.37	24.120	6.673	10.050	1.898
...	
Ovary	70.570	25.500	15.81	28.880	16.930	27.060	35.440	17.900	17.870	46.92	...	9.478	7.695	19.500	24.460	23.720	33.54	54.820	9.526	25.080	2.798
Ovary	32.750	29.040	23.77	54.790	27.450	29.820	35.600	37.010	28.790	91.07	...	6.690	9.283	21.390	33.520	25.860	39.23	65.080	10.330	26.630	3.904
Ovary	51.210	20.160	18.07	34.830	17.200	26.180	32.190	22.920	14.470	56.43	...	7.520	9.168	16.680	26.540	16.380	30.49	44.980	8.804	17.250	1.283
Ovary	36.780	24.500	19.53	33.970	27.460	34.090	33.960	25.140	25.960	112.90	...	7.303	6.738	25.820	27.240	26.620	35.62	55.150	6.955	22.910	5.288

下一步，進行標準化：

為了減少樣本間的誤差，我們將每個樣本的31個NPCs基因總表現量總和等於1，如下圖所示：

	Gene 1	Gene 2	...	Gene31		Gene 1	Gene 2	...	Gene31		
Sample 1	22.4	21	...	4.4	Normalization	Sample 1	0.05	0.047	...	0.01	Sum = 1
Sample 2	34.9	19.7	...	1.3		Sample 2	0.064	0.036	...	0.002	
...		
Sample 17381	51.2	20.1	...	1.2		Sample 17381	0.041	0.027	...	0.005	
Sample 17382	36.7	24.5	...	5.2		Sample 17382	0.032	0.032	...	0.001	

因此，我們把樣本差異注重在基因間的比例上。

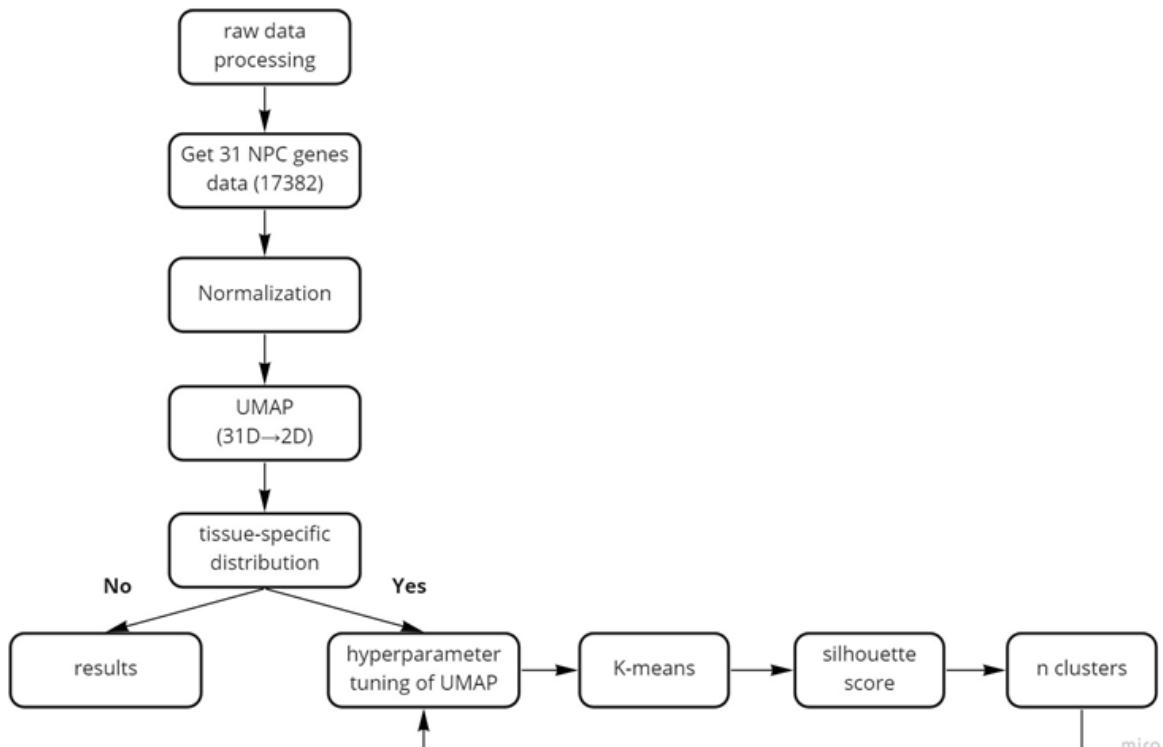
我們將標準化的資料作為我們分析的對象。

Insights discovered from the data

e.g., Data distribution and preliminary analysis

e.g., Which attributes have higher correlation with the prediction target?

Methodology details



分析步驟可分為六項：

1. umap()將31維資料降成2維

umap()的可調參數共有四種:n_neighbors代表number of neighbors, 指附近有幾個資料點為鄰居, 可用於平衡保留局部或全局結構, 預設為15且範圍為2~200, 因此該實驗假設2, 10, 20, 50, 100, 200共六種進行實作;min_dist代表minimum distance, 指降維後資料點間的最小距離, 可決定資料分布的緊密性, 預設為0.1且範圍為0.1~0.99, 因此該實驗假設0.0, 0.1, 0.25, 0.5, 0.8, 0.99共六種進行實作;n_components指欲降成幾維的資料, 預設為2且通常會設置成三以下以方便視覺化, 該實驗維持預設值; metric指資料點間的距離適用何種距離算法, 該實驗維持預設的歐式距離。

2. 降維結果繪製散布圖

將降維結果的兩維依據xy軸繪製成散布圖, 資料點依據30種不同組織標上不同的顏色, 用以觀察同一組織的資料點分佈情形, 若有聚集可推斷該組織具有特異性。

3. 降維資料找出最適合分群數目

為確保有最佳的聚類效果, 使用silhouette coefficient(輪廓係數法)決定同群間緊密、異群間遠離的分佈, 度量最適合的分群數目。

4. 降維資料使用Kmeans分群

降維資料使用scikit-learn套件中的kmeans演算法, 指定k值為前步驟的分群數目, 分群結果依據xy軸繪製成散布圖, 資料點依據不同分群編號上色, 並於每群的中心點標出該群編號。

5. 分群結果和組織分布比對重疊性

繪製一個dataframe, 每一列代表一種組織共有30列, 欄位包含main cluster, cluster size, tissue size, percentage, main cluster是主要分群編號, cluster size主要分群編號的資料筆數, tissue size是該組織的資料筆數, percentage是 $\frac{\text{cluster size}}{\text{tissue size}}$ 代表該組織正確分群的比率。

6. 依據最佳結果調整umap超參數

前步驟的30筆percentage資料繪製盒鬚圖, 並於圖上標出中位數(median)與第三個四分位數(Q3)值, 最高的Q3值代表該筆降維資料的四分之一分群結果和組織分佈重疊性最高。

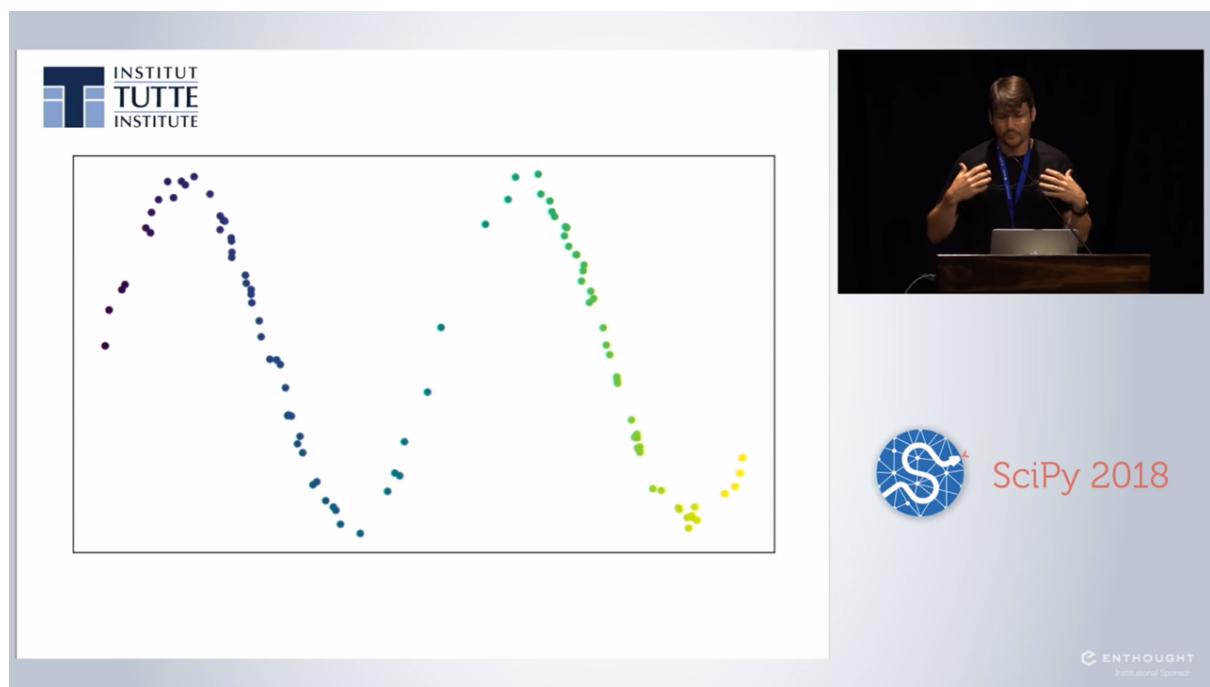
Principle of UMAP

UMPA(Uniform Manifold Approximation and Projection), 是一種非線性的降維工具, 於2018由McInnes, L., Healy, J., Saul, N., & Großberger, L.等科學家發表於*Open Source Software*。在原本生物資訊領域裡, 另一降維工具—TSNE為生物學家們最常使用的非線性降維工具, 但2018年誕生的UMAP很好的利用了TSNE的相關概念, 並結合了作者自己的想法後, 使其能更好的保留資料全局的樣貌, 並且UMAP普遍運行的時間比TSNE要更短, 使其成為生物資訊領域裡降維工具的當紅炸子雞。

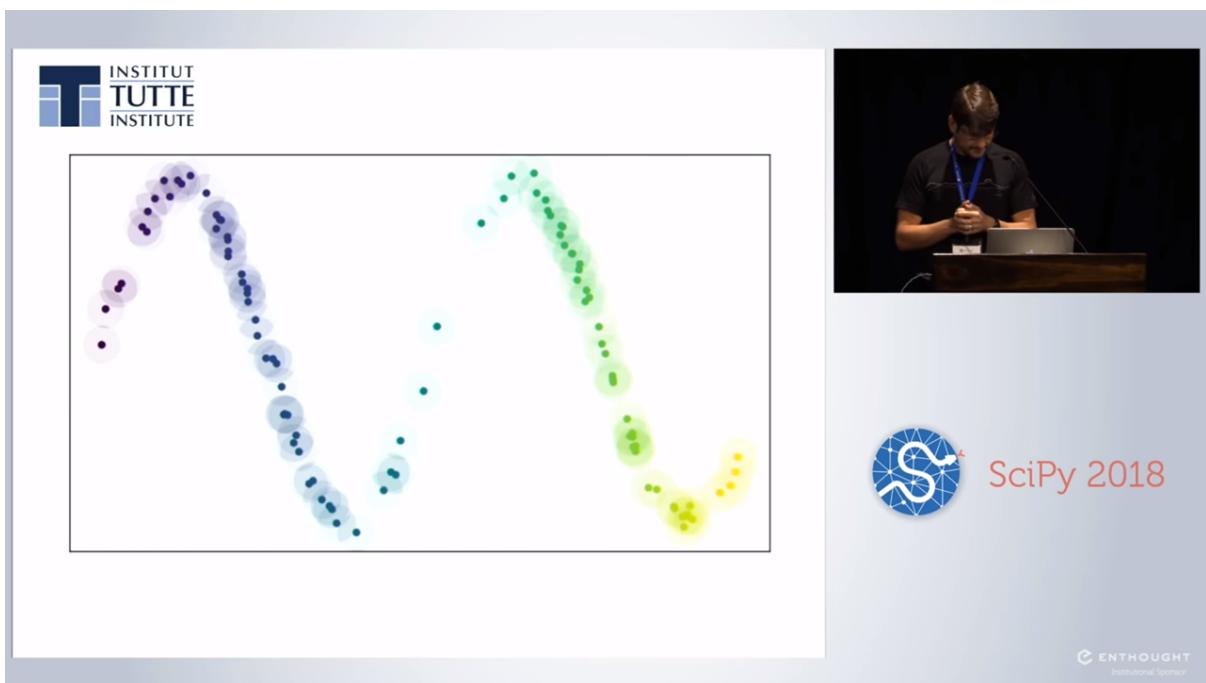
非線性的降維工具有一共同的策略:

1. 試著在高維度的資料中建立資料點之間的關係(資料點與資料點的距離)
2. 想辦法在低維度呈現高維度的資料點之間的關係

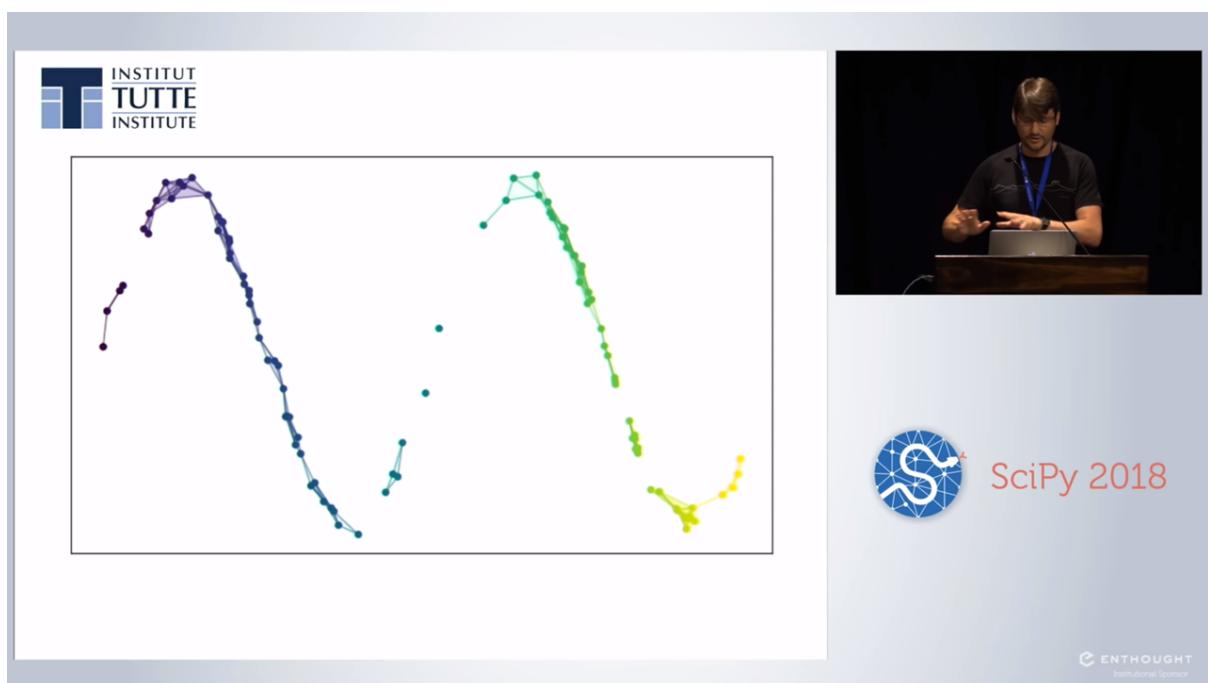
UMAP也同樣運用了這個策略, 而其精華在於如何建立資料點與資料點之間的關聯。以下會根據原作者給予的範例來進行講解。



首先這是原始的資料分布(二維), 可以看出有較集中也有較疏鬆的部分。第一步是找出點與點之間的關係(意即找出誰是誰的鄰居)。



於是乎利用一個直觀的方式— ϵ -neighborhood, 在點的周圍畫一個半徑為定值的圓，而圓之間有交互的，抑或直接包含住資料點的，代表他們有關係(彼此是鄰居)，最後會得到以下結果：



可以發現的是，集中區域點與點之間的關係比較有被建立，而疏鬆地區的則沒辦法，這樣會導致降維時疏鬆區域的部分無法被保留，進而導致只有集中區域的資料點再

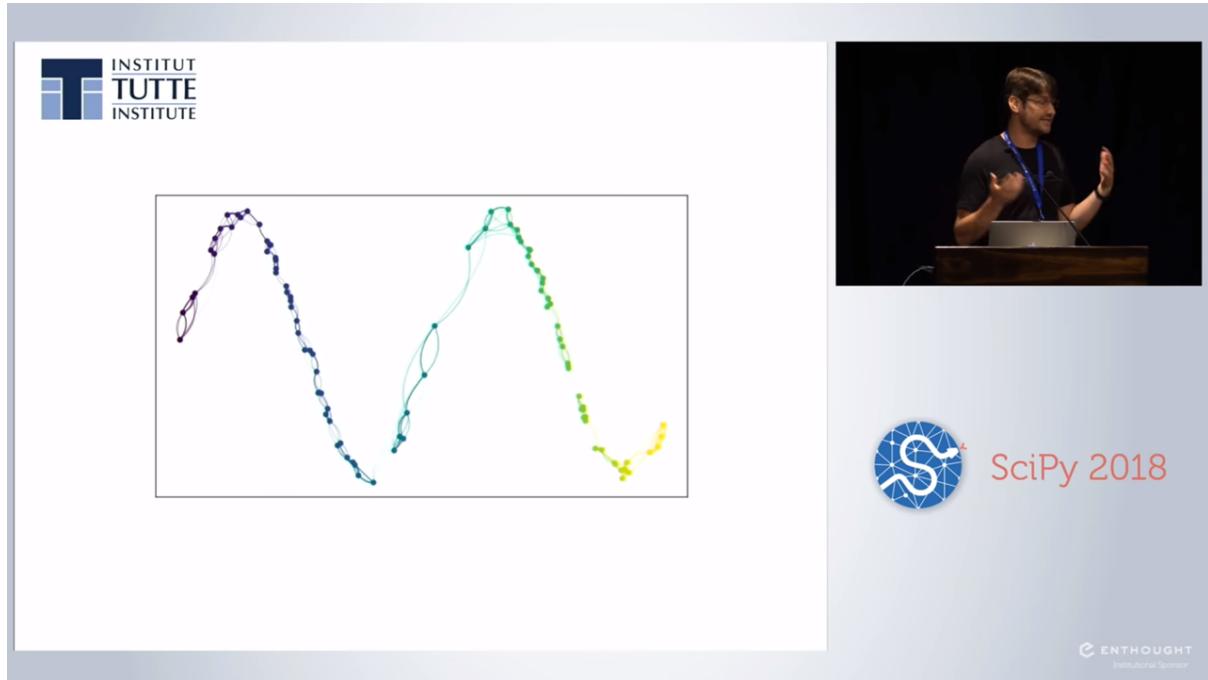
降維時能被叫好的呈現，代表這個降維只能保留局部構造，而整體的結構則無法留下，作者就是要克服這件事情，所以作者做了一個新的預設：



作者假定每個點一定要有一個關聯(代表最少每個點都要有一個鄰居)，根據這個設定，作者對於每個點找鄰居的範圍，起點從利用與最近鄰居的距離當成半徑所畫成的圓開始，並再額外選定一個範圍，從與最近鄰居距離所構成的圓往外開始計算與範圍內資料點的關係。作者利用了SNE的想法：把資料點的距離帶入高斯函術後會形成機率分布，而在機率分布的裡的機率值則為彼此關係的資訊，所以從以上可以理解UMAP在高維裡建立資料點關係的公式：

$$p_{j|i} = \exp\left(-\frac{\|x_i - x_j\|^2 - \rho_i}{\sigma_i}\right)$$

其中 $|x_i - x_j|$ 為點*i*與點*j*的距離，*i*為欲計算關係的資料點，*j*為範圍內其他可能與*i*有關係的資料點， ρ_i 為點*i*與最近鄰居的距離。 σ_i 為決定範圍(有多少點*j*)的參數，其利用KNN的概念，決定k而影響 σ_i 。利用此方法會建立出以下資料點之間的關係：

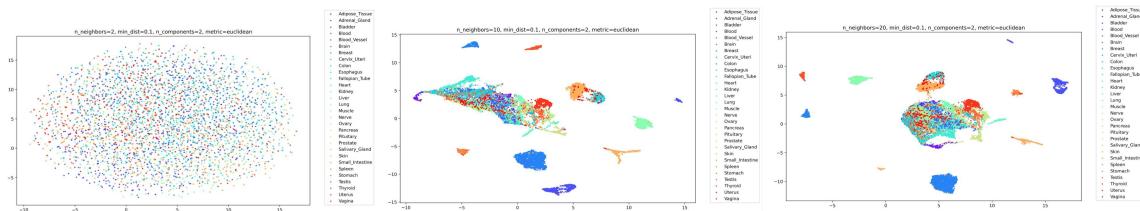


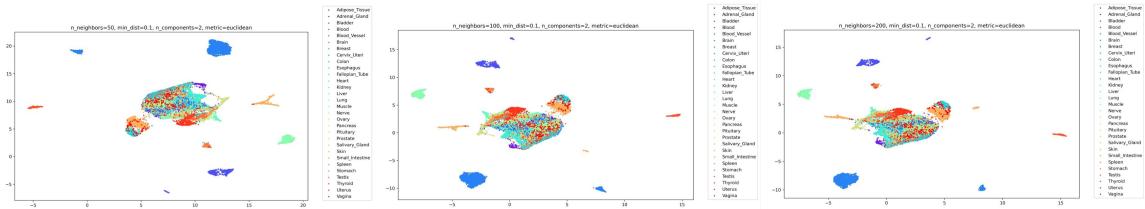
可以發現，資料點之間的關聯性更好了，尤其是在比較疏鬆的區域，這代表在降維時UMAP比起TSNE或其他工具，可以更好的保留全局結構，來完成作者的目標：Uniform Manifold。

總的來說，UMAP最精華的部份是變化了找鄰居的起始點與尋找的範圍來更好的使高維資料點較缺乏的部分不會再降維時被忽略，其接下來的步驟與TSNE相差不多，同樣都是用T-distributed的方式避免降維後資料點會有過於集中的現象，並利用Binary cross entropy loss最佳化gradient descent，所以UMAP可以說是繼承了TSNE精華的同時，又創造出了自己獨特優勢的一面。

Evaluation and Results

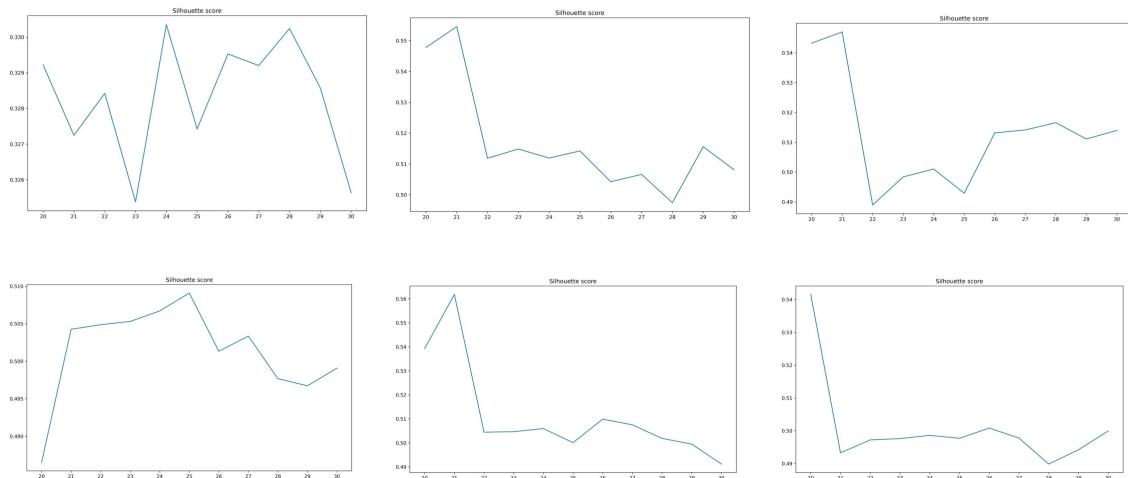
如下圖，`umap()`中`min_dist`維持預設的0.1，由左至右再由上至下`n_neighbors`依序設為2, 10, 20, 50, 100, 200，xy軸分別代表二維資料的數值，同一種組織的資料點標為同一種顏色。可得知設為2時並無分組效果，而設為10以上便有明顯的分組效果，尤其越靠近圖四周的資料點越明顯有聚集的情形，就單純觀察`umap`結果散佈圖而言，我們推斷設為10以上的數字即可。





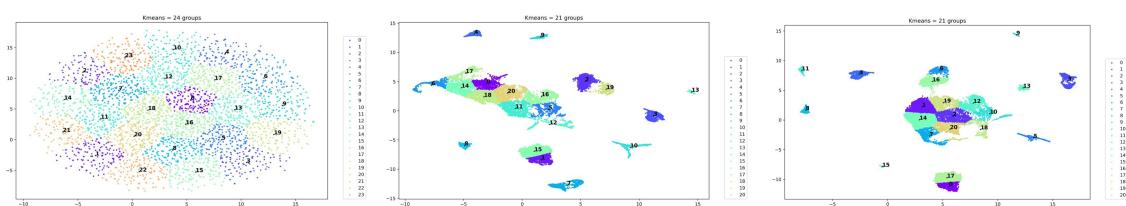
更換n_neighbors的umap結果圖

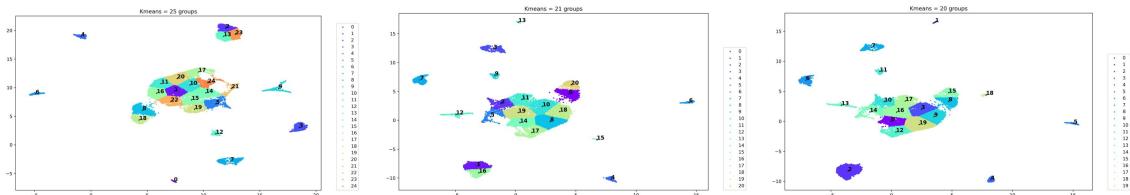
降維後的二維資料用silhouette score()計算最適合的K值，如下圖，參數順序依照上圖，x軸為K值設成20~30之間，因為預計最高分群數目不超過原有的30種組織，y軸為silhouette score的值，範圍為0.1至1.0，silhouette score最大者最適合K值，可得大部分K值落在20和21，代表該二維資料適合分成20或21群。



更換n_neighbors的silhouette score結果圖

接著，用最佳K值指定KMeans()的n_clusters，如下圖，參數順序依照上圖，xy軸分別代表二維資料的數值，同一群的資料點標為同一種顏色，並於該群的中心點標上三角形符號與該分群編號，可得散佈在圖四周位置的資料點分群明顯。





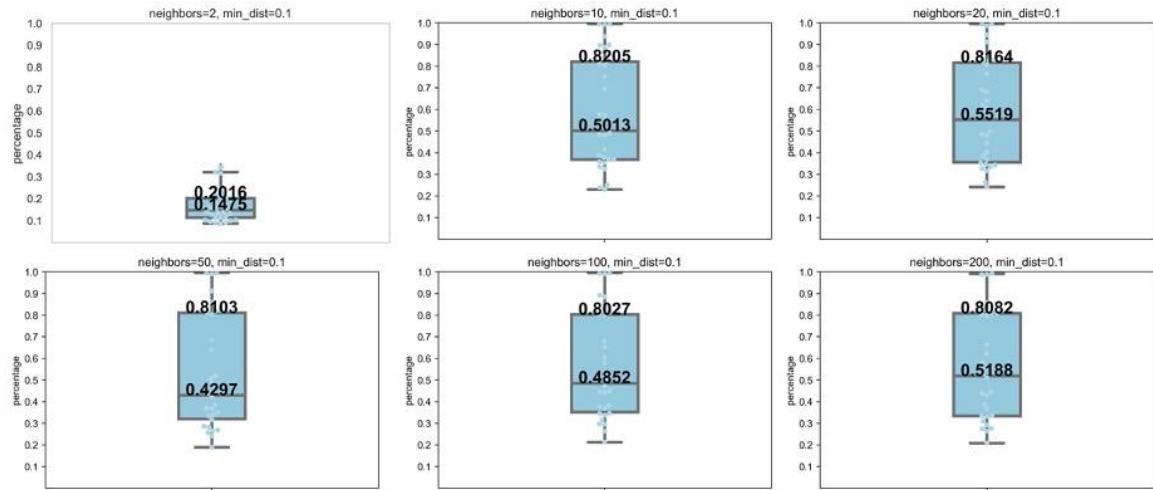
更換n_neighbors的kmeans結果圖

為了評估分群結果和組織是否有高度重疊性，使用percentage dataframe保存欲畫盒鬚圖的percentage數值，下圖為n_neighbors為10、min_dist為0.1時的percentage dataframe。

	tissue	main_cluster	cluster_size	tissue_size	percentage
0	Muscle	0	796	803	0.991283
1	Blood_Vessel	6	478	1335	0.358052
2	Heart	8	497	861	0.577236
3	Uterus	3	46	142	0.323944
4	Vagina	6	39	156	0.250000
5	Breast	6	225	459	0.490196
6	Skin	1	1255	1809	0.693753
7	Salivary_Gland	5	63	162	0.388889
8	Brain	6	1279	2642	0.484103
9	Adrenal_Gland	2	209	258	0.810078
10	Thyroid	8	612	653	0.937213
11	Lung	5	240	578	0.415225
12	Spleen	1	240	241	0.996851
13	Pancreas	0	294	328	0.896341
14	Esophagus	8	533	1445	0.368858
15	Colon	2	179	779	0.229782
16	Small_Intestine	1	66	187	0.352941
17	Prostate	6	118	245	0.481633
18	Testis	1	359	361	0.994460
19	Nerve	5	510	619	0.823910
20	Blood	0	748	929	0.805167
21	Pituitary	5	251	283	0.886926
22	Liver	5	203	226	0.898230
23	Kidney	0	67	89	0.752809
24	Cervix_Uteri	3	7	19	0.368421
25	Fallopian_Tube	2	3	9	0.333333
26	Bladder	6	5	21	0.238095
27	Adipose_Tissue	9	617	1204	0.512458
28	Stomach	1	206	359	0.573816
29	Ovary	5	68	180	0.377778

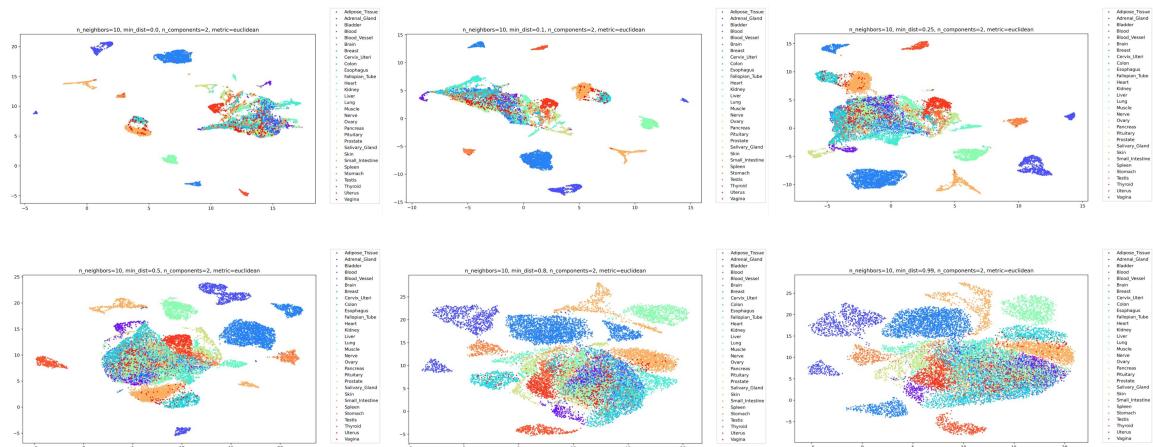
percentage dataframe

30筆percentage資料繪製盒鬚圖，如下圖，參數順序比照前面的圖，y軸為percentage數值，範圍從0.0至1.0，淺藍色點為資料點，下方黑色數字代表中位數，上方黑色數字代表Q3，六張圖中可得知n_neighbors為10的Q3值是0.8205且最高，表示共有四分之一的資料分群和組織吻合度達0.8205，因此我們推斷n_neighbors為10是最適合的。



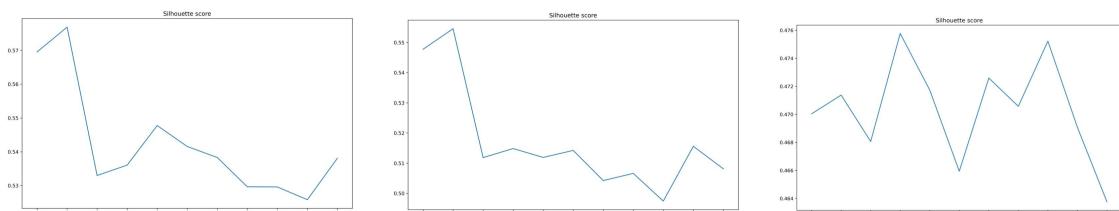
更換n_neighbors的percentage盒鬚圖

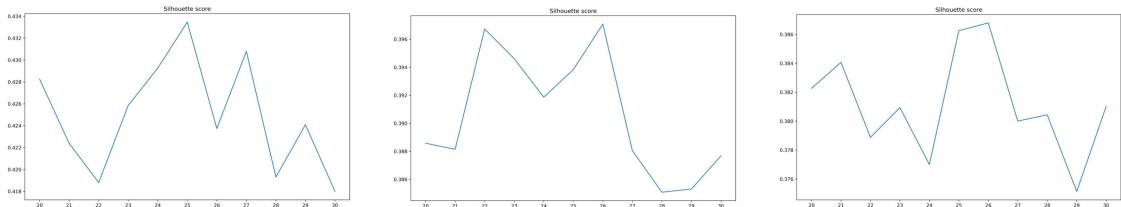
再著，固定n_neighbors為10後，min_dist設成0.0, 0.1, 0.25, 0.5, 0.8, 0.99，圖順序一樣是由左至右由上至下，下圖中可觀察到min_dist設為0.1的資料點不同群之間的距離較遠，同群的資料點較緊密，因此單純觀察我們推斷設為0.1是最適合的。



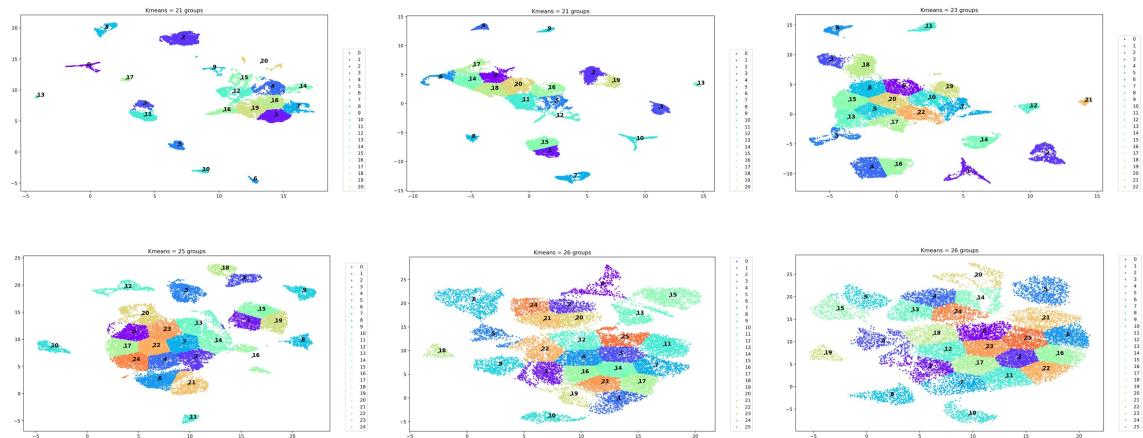
更換min_dist的umap結果圖

接著，依照上述步驟找出最佳K值做分群，並且畫出kmeans結果圖，觀察可發現min_dist設為0.8以上的分群結果較不準確，因為分佈在同一區的資料點被分在不同群內，相對而言min_dist設為0.1的分群結果較準確。



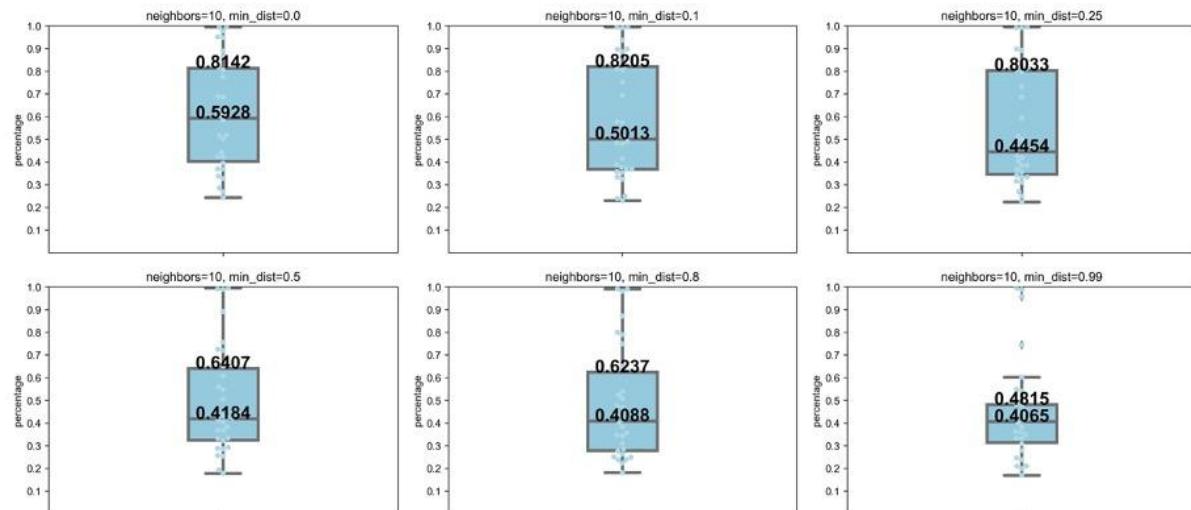


更改min_dist的silhouette score結果圖



更改min_dist的kmeans結果圖

最後，就percentage盒鬚圖中可得知Q3值最高的依舊是min_dist為0.1的0.8205，因此我們認為n_neighbors設成10、min_dist設成0.1是最適合的umap參數。



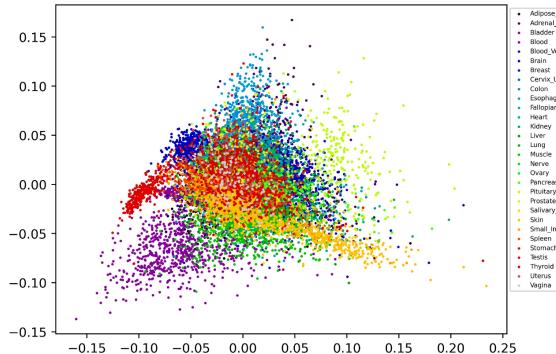
更改min_dist的percentage盒鬚圖

Conclusions and novelty

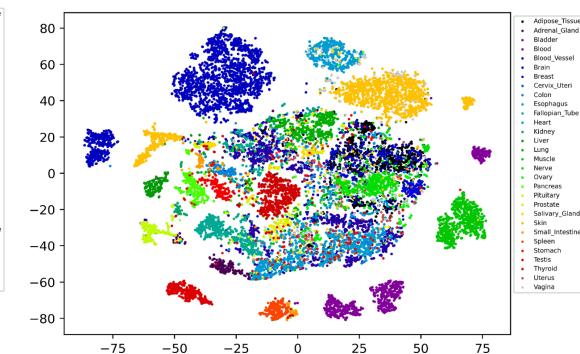
組織特異性代表分群結果和組織分佈有很高的重疊性，就實驗結果而言是部分有組織特異性而部分沒有，若將閾值定為0.8205則表示有七種組織具有組織特異性：

脾臟、睪丸、肌肉、甲狀腺、肝臟、胰臟、腦下垂體、神經，其中脾臟、睪丸、肌肉的重疊性高達0.99以上，具有高度組織特異性。

另外，我們觀察到降維資料中間區域具有密集的資料點，調整參數後也無法將其分開或是準確分群，因此嘗試了其他降維工具PCA和t-SNE，結果如下：



PCA降維結果圖



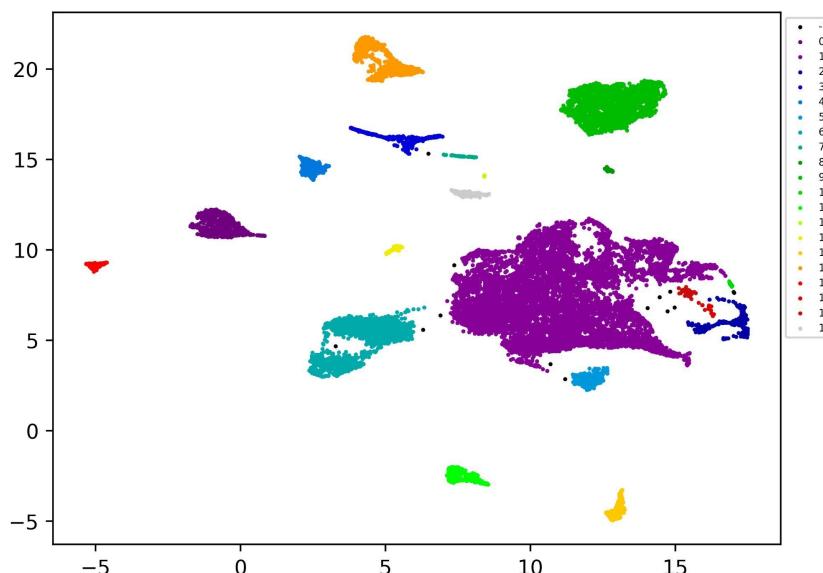
t-SNE降維結果圖

從PCA降維結果圖中可發現資料點密集集中無法分群，可能是因為PCA為線性降維工具，對於非線性高維資料的處理不適合，從t-SNE將為結果圖中可得四周的資料點明顯分群，但依舊無法將中心的密集區分開，故同樣無法達到我們預期的效果。

Extension

在報告後，我們參考老師的建議，除了使用k-means來分群外，也使用了**DBSCAN**來進行分群。

超參數: `eps=0.3, min_samples=3, metric='euclidean'`.



DBSCAN Result

討論：黑色(-1)為離群樣本，結果總共分為20群，與原預期的30群有些落差，主要是因為圖中紫色的cluster為較無組織特異性的樣本，但無可否認的是有10多個clusters彼此遠離，在DBSCAN有良好的分離，亦即這些群集有可能就是具有相當程度的組織特異性。

因此，在DBSCAN中的判斷下，結果可以推得NPCs的確還是部分組織有組織特異性，部分沒有。

The contribution of each team member

Name	Contributions
蔡秉勳	NPC基因表現量資料處理、PCA、t-SNE、DBSCAN延伸分析
林耿弘	umap原理研究、umap工具的數學意義
周治瑗	NPC表現量資料的umap實作、kmeans分群、組織特異性的結果分析