Single-cell Bioinformatics Analysis

Mengrui Zhang, Magdy Alabady
The Georgia Genomics and Bioinformatics Core (GGBC)

8/23/2019

1. Introduction

In this analysis, we will perform the analysis of single-cell RNAseq data on Cat. The data were from 10x genomics analysis pipeline by using "Cellranger counts" with the cat reference genome. The R packages we use in this analysis will be "Monocle3" and "Seurat".

We plan to perform the following:

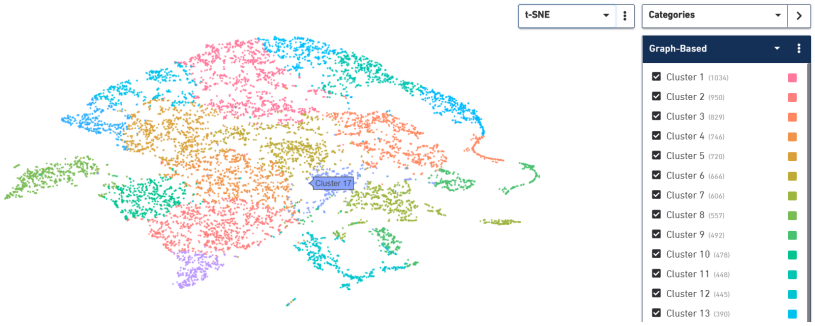
1. Preprocessing the data using Principal components analysis (PCA).
2. Cluster cells using UMAP and Louvain methods.
3. Construct cell trajectories.
4. Oder cells in chronological order.
5. Pseudo time analysis.
6. Significant gene analysis.

2. Data set analysis

2.1 Raw data analysi using "Cellranger" and "10x Loupe Cell Browser"

In this section, for the first step, we run an analysis pipeline using "Cellranger count" by 10x genomics. The reference genome was created using "Felis_catus_9.0" database. The result can be shown by using the "10x Loupe Cell Browser". The following 2 figure is an example of "10x Loupe Cell Browser" interface.

This figure shows the clustering result (K-mean) using "10x Loupe Cell Browser".



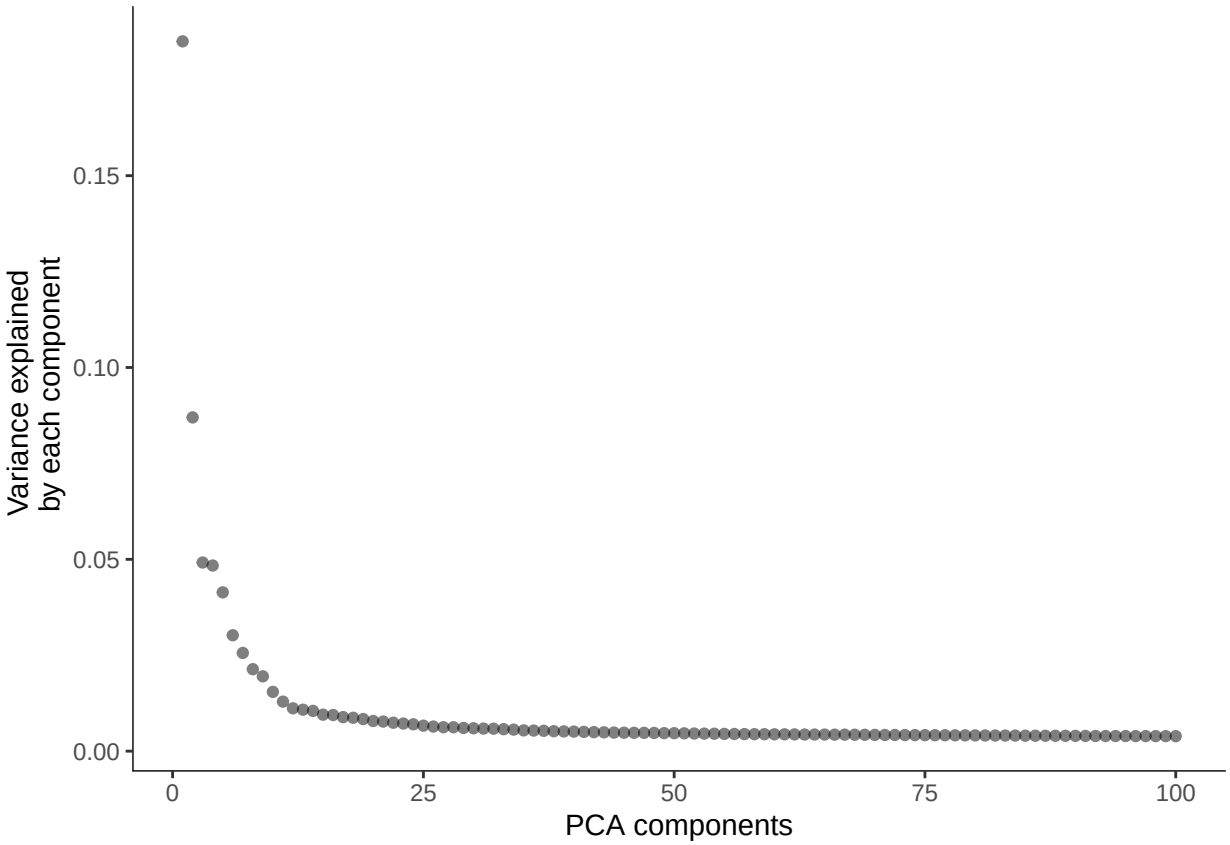
The above figure shows the clustering result from "Loupe" using the t-SNE method. There are more detailed results inside of the "10x Loupe Cell Browser", the software has a great interactive interference for users.

2.2 Load dataset into R package "Monocle3" and preprocessing with PCA

The output of Cellranger analysis pipeline has three files: "barcodes", "features" and "matrix". The "barcodes" file contains all the cell names and the "feature" file contain the genes and gene short names. The "matrix" file includes the counts with the location of gene and cell. We load this data into Rstudio.

```
## class: cell_data_set
## dim: 26373 10000
## metadata(1): cds_version
## assays(1): counts
## rownames(26373): ENSFCAG00000041896 ENSFCAG00000011704 ...
##   ENSFCAG00000032078 ENSFCAG00000032079
## rowData names(2): id gene_short_name
## colnames(10000): AAACCCAAGCTGTTCA-1 AAACCCAAGGTAGGCT-1 ...
##   TTTGTTGTCTCCTGAC-1 TTTGTTGTCTTAATCC-1
## colData names(2): barcode Size_Factor
## reducedDimNames(0):
## spikeNames(0):
```

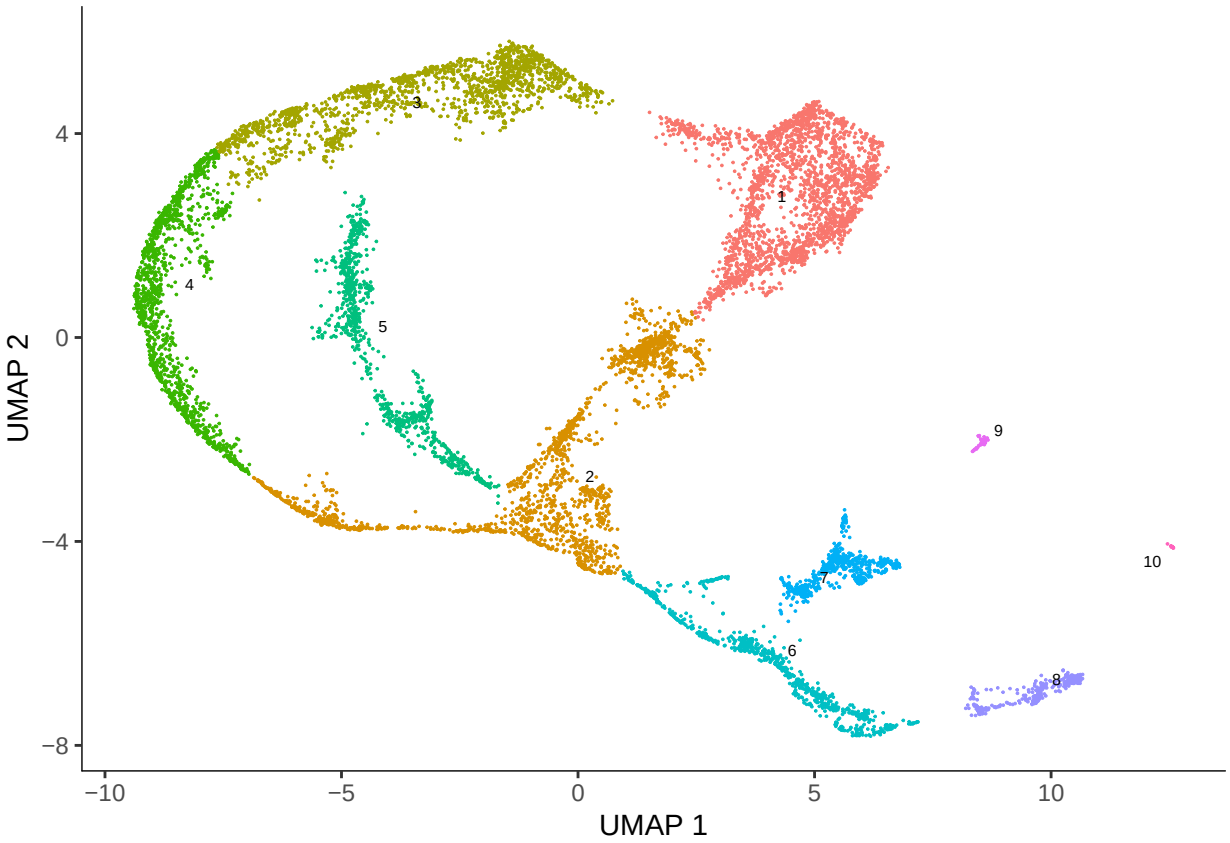
The above is a summary of the R object for this dataset. There is a total of 10000 cells and 26373 genes. We preprocessing the data into 100 dimensions for our later analysis. This preprocessing step is to normalize the data, to use Principal Components Analysis (the standard for RNA-seq) and to remove any batch effects.



Here we plot the PCA result with 100 dimensions. We can see the importance of component decrease as the dimension increase. We choose 100 dimensions for our data should be enough.

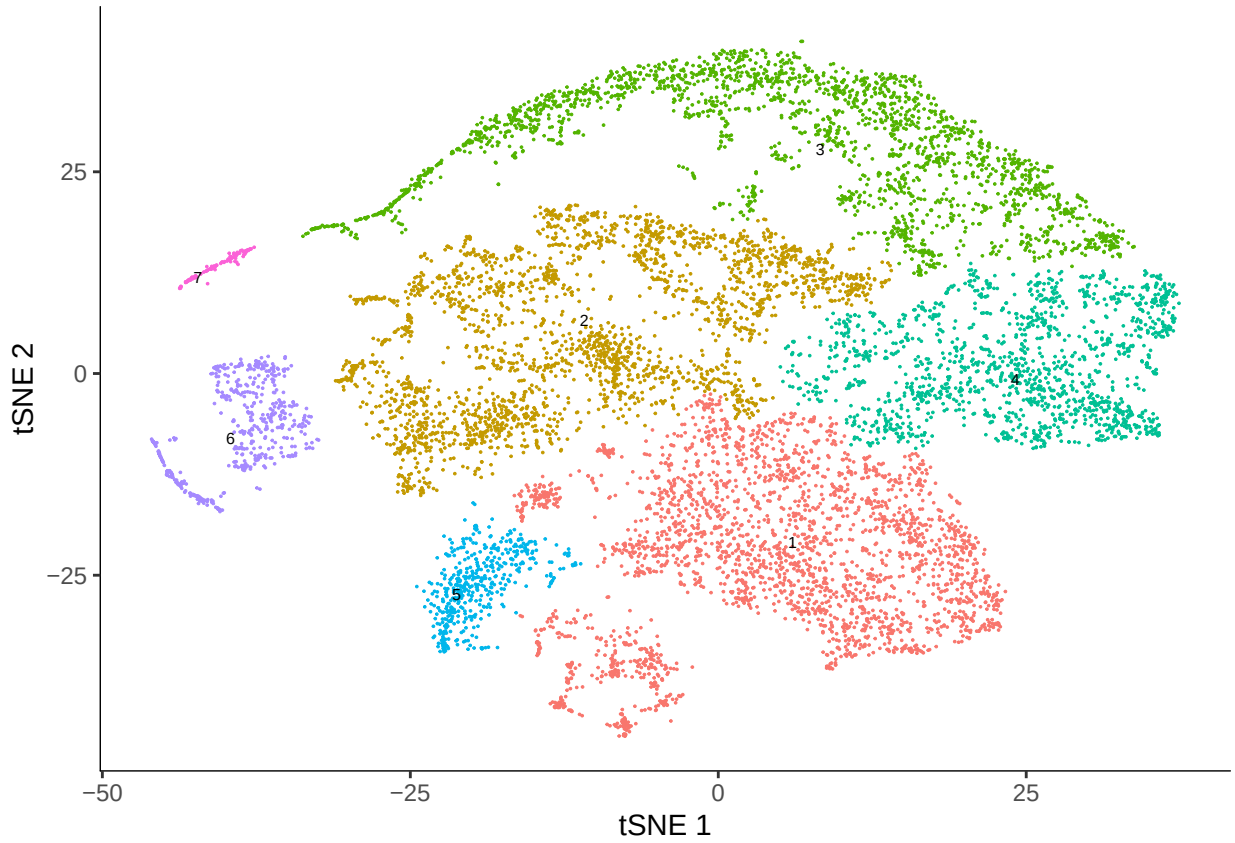
2.3 Cluster cells using UMAP and Louvain methods.

We cluster the cells using Uniform Manifold Approximation and Projection (UMAP). There is a total of 10 clusters.



2.4 cluster the cells using tSNE

We cluster the cells using T-distributed Stochastic Neighbor Embedding(tSNE).



This figure shows the clustering result using "t-SNE" method.

2.5. Find Marker gene for each cluster.

Now we run a test to find out what genes makes them different from one another. Then we can rank markers genes according to "pseudo R2" value. For each cluster, we choose top 3 genes according to the "pseudo R2" value. The following table shows a total of 30 genes for all 10 clusters by UMAP.

Table 1: Table continues below

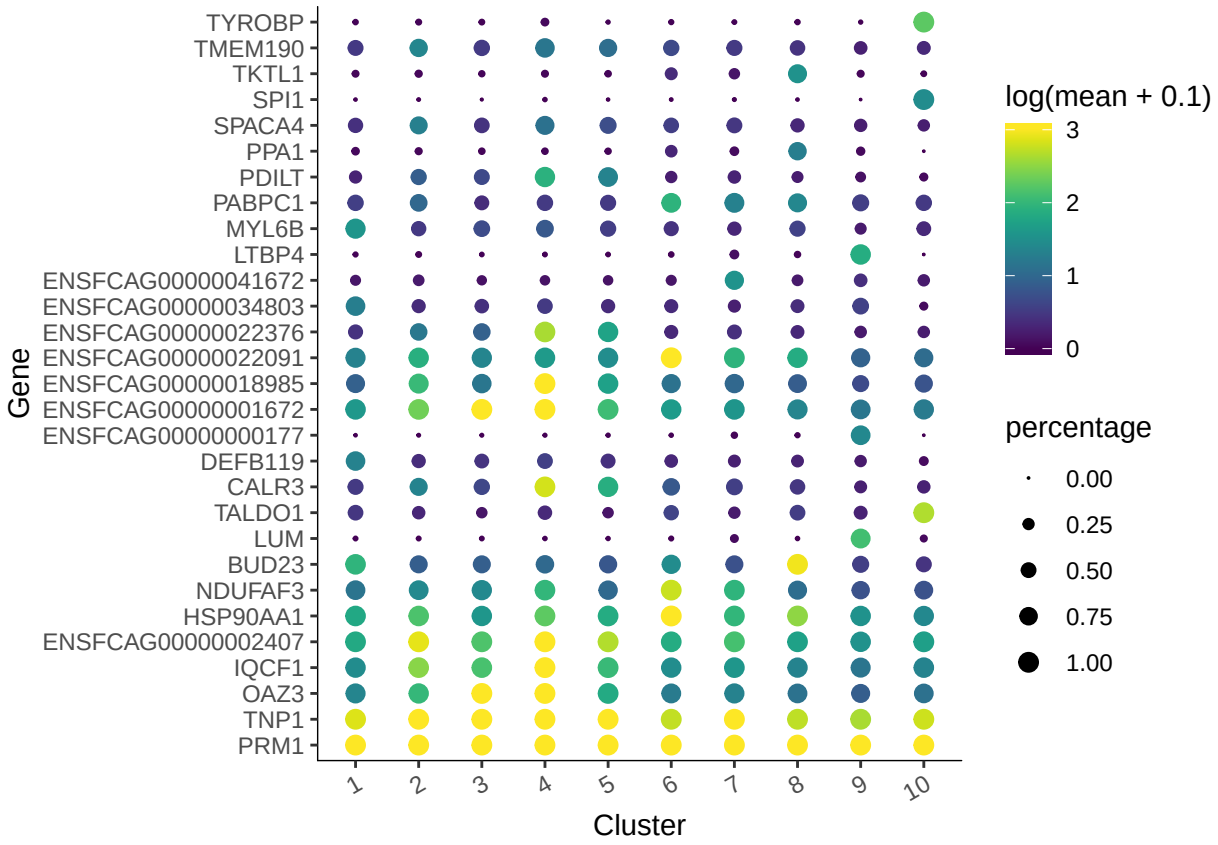
gene_id	gene_short_name	cell_group	marker_score
ENSFCAG00000000177	ENSFCAG00000000177	9	0.7868
ENSFCAG00000003764	LUM	9	0.7692
ENSFCAG00000001036	LTBP4	9	0.8202
ENSFCAG00000011392	TKTL1	8	0.4749
ENSFCAG00000028251	PPA1	8	0.4262
ENSFCAG00000004748	BUD23	8	0.3515
ENSFCAG00000041672	ENSFCAG00000041672	7	0.4333
ENSFCAG00000009843	NDUFAF3	7	0.2344
ENSFCAG00000001874	PABPC1	7	0.2281
ENSFCAG00000009843	NDUFAF3	6	0.2909
ENSFCAG00000000814	HSP90AA1	6	0.3503
ENSFCAG00000022091	ENSFCAG00000022091	6	0.3135
ENSFCAG00000036928	CALR3	5	0.3789
ENSFCAG00000022376	ENSFCAG00000022376	5	0.3672

gene_id	gene_short_name	cell_group	marker_score
ENSFCAG00000018103	PDILT	5	0.3666
ENSFCAG00000002407	ENSFCAG00000002407	4	0.3139
ENSFCAG00000036860	TNP1	4	0.3331
ENSFCAG00000009142	IQCF1	4	0.2999
ENSFCAG00000010979	OAZ3	3	0.3354
ENSFCAG00000001672	ENSFCAG00000001672	3	0.2991
ENSFCAG00000041147	PRM1	3	0.3076
ENSFCAG00000018985	ENSFCAG00000018985	2	0.18
ENSFCAG00000044046	TMEM190	2	0.1573
ENSFCAG00000014461	SPACA4	2	0.1759
ENSFCAG00000002531	SPI1	10	0.9521
ENSFCAG00000007593	TALDO1	10	0.7372
ENSFCAG00000006617	TYROBP	10	0.9218
ENSFCAG00000034803	ENSFCAG00000034803	1	0.3392
ENSFCAG00000014863	MYL6B	1	0.3697
ENSFCAG00000039915	DEFB119	1	0.3613

Table 2: Table continues below

mean_expression	fraction_expressing	specificity	pseudo_R2
5.838	0.8727	0.9015	0.7525
14.53	0.8909	0.8634	0.7646
9.556	0.9545	0.8592	0.8503
2.948	0.7978	0.5953	0.4971
2.064	0.7368	0.5785	0.4188
15.88	0.9945	0.3534	0.5408
9.633	0.8175	0.53	0.4467
11.31	0.9782	0.2396	0.2533
5.283	0.8924	0.2556	0.1922
14.78	0.9909	0.2936	0.3613
42.6	1	0.3503	0.6828
22.95	0.9974	0.3144	0.4995
12.42	0.9434	0.4017	0.3813
9.642	0.9456	0.3883	0.3873
5.853	0.8757	0.4186	0.3489
51.79	1	0.3139	0.7892
178.4	1	0.3331	0.8589
29.89	1	0.2999	0.7817
30.34	0.9994	0.3356	0.6881
34.17	1	0.2991	0.6917
494.9	1	0.3076	0.6821
7.694	0.9042	0.1991	0.06178
3.445	0.7274	0.2162	0.07765
2.909	0.6976	0.2522	0.07934
7.108	1	0.9521	0.9363
27.9	1	0.7372	0.9256
17.99	1	0.9218	0.957
4.615	0.8594	0.3947	0.3184
6.589	0.924	0.4001	0.3858
5.101	0.8537	0.4232	0.3288

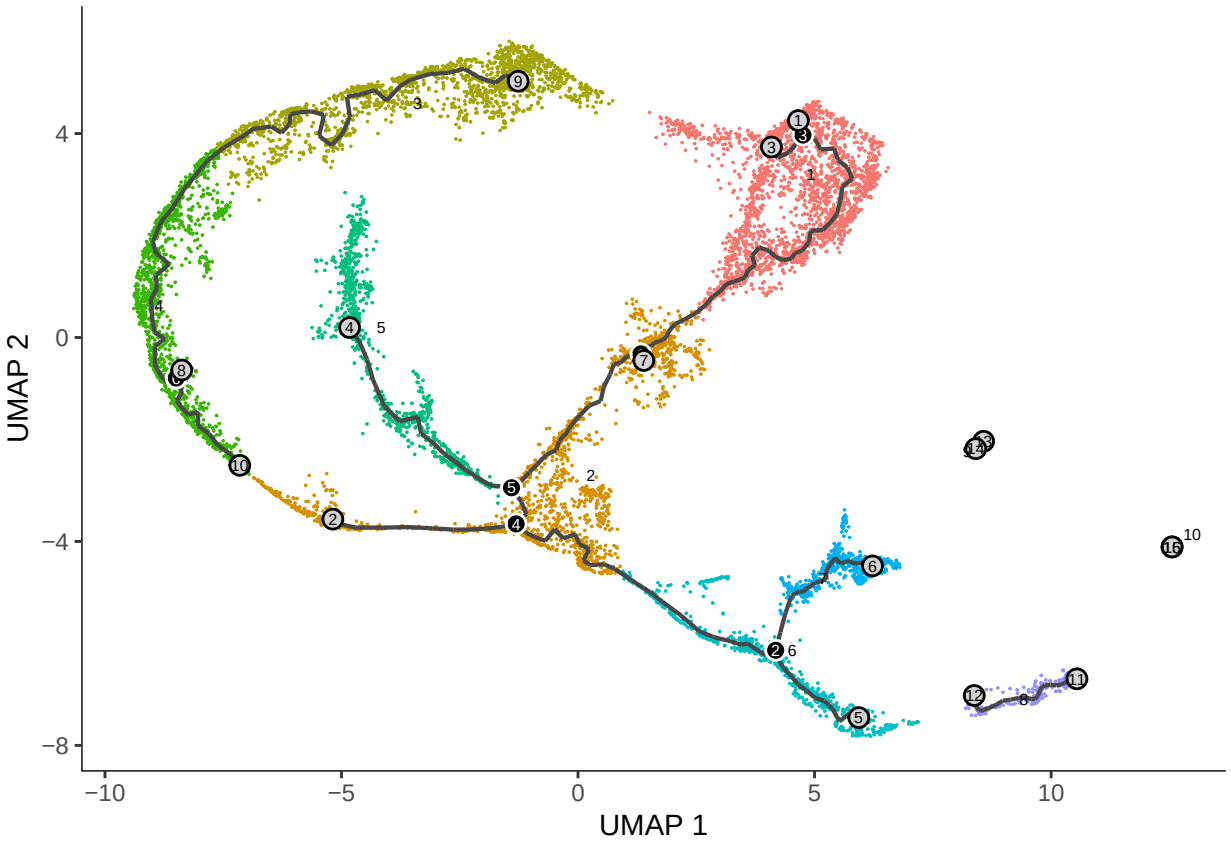
marker_test_p_value	marker_test_q_value
3.701e-112	9.762e-107
5.569e-114	1.469e-108
6.784e-127	1.789e-121
2.862e-154	7.549e-149
1.432e-129	3.775e-124
3.607e-168	9.512e-163
4.96e-187	1.308e-181
5.768e-105	1.521e-99
1.136e-79	2.996e-74
4.856e-164	1.281e-158
4.6e-319	1.213e-313
2.224e-229	5.866e-224
2.039e-186	5.379e-181
1.776e-189	4.684e-184
3.65e-170	9.627e-165
0	0
0	0
0	0
0	0
0	0
0	0
2.196e-40	5.793e-35
2.058e-50	5.427e-45
1.73e-51	4.561e-46
2.934e-57	7.739e-52
1.273e-56	3.356e-51
1.663e-58	4.385e-53
4.588e-223	1.21e-217
8.8e-273	2.321e-267
1.325e-230	3.493e-225



The above figure is a heatmap shows the significant genes and their corresponding clusters.

2.6 Construct trajectories

The trajectories of cells can be learned by using "monocle3" package. The trajectories were calculated based on the UMAP cluster method.



From the above figure, we can see that there are about 2 major trajectories on the graph, one trajectory on the left goes from top to bottom and one trajectory on the center which has around 5 branches. There are also a few small trajectories which a little far from the major clusters.

2.7 Oder cells in chronological order

In this section, We first use the R package "Seurat" basically to create a cell-gene data matrix and perform normalization. The following table is the significant genes in each of the developmental stages.

Table 4: Table continues below

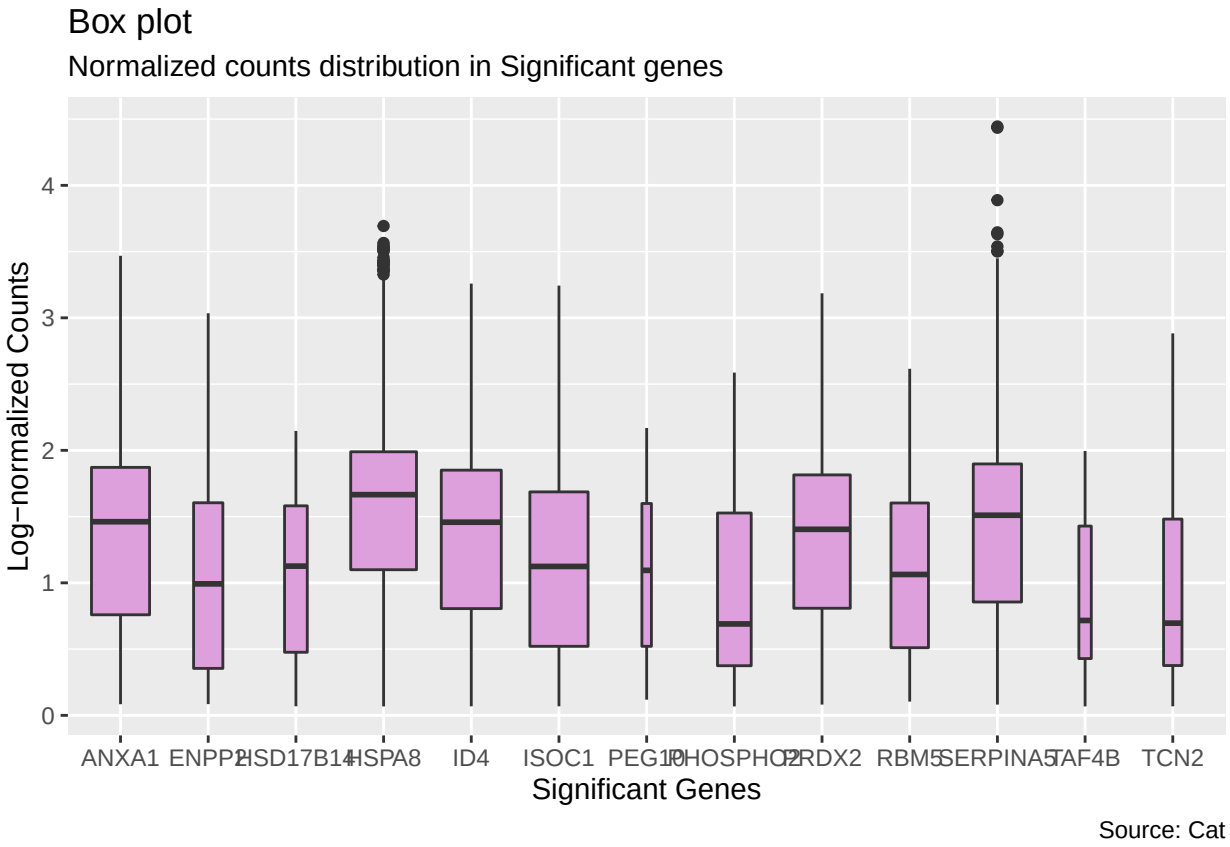
Stage	Order
Spermatogonial Stem Cells (SSC's)-Stem Progenitor	1
early undifferentiated Spermatogonia-early Signaling	2
late undifferentiated	3
Spermatogonia-Proliferation/Differentiation	
Progenitor-early Differentiation-Leptotene/Zygotene	4
Progenitor-early Differentiation-Leptotene/Zygotene	4.2
Progenitor-early Differentiation-Leptotene/Zygotene	4.3
Zygotene/Pachytene	5
Pachytene/Diplotene	6
early round Spermatids	7
late round Spermatids	8
perhaps Sertoli/peritubular or epithelial cells	9
perhaps Immun/Macrophages/Perivascular	10

Genes
ID4- HSPA8- TAF4B- PEG10- RBM5- TCN2- ISOC1- HSD17B14- PHOSPHO2- ENPP2- PRDX2- ANXA1- SERPINA5- PAX7-
PIWIL1- TAF4B- PIWIL2- RBM5- SERPINA5- MX1- C7-
GFRA1- PIWIL4- ZBTB16- TKTL1- HSPA8- DMRT1- PPA1- KIT- TAF4B- PIWIL2- RBM5- CTD1- ISOC1- NMT2- PHOSPHO2- HNRNPH3- SMC3-
STRA8- BMI1- Dmc1- PIWIL1- PKTL1- PIWIL4- HSPA8- PPA1- HORMAD1- PIWIL2- RBM5- TCN2- CTD1- NMT2- MEIOB- TEX101- CETN3- PHOSPHO2- MLLT10- HNRNPH3- SYCP1- RAD21L1- SMC3-
STRA8- BMI1- Dmc1- PIWIL1- PKTL1- PIWIL4- HSPA8- PPA1- HORMAD1- PIWIL2- RBM5- TCN2- CTD1- NMT2- MEIOB- TEX101- CETN3- PHOSPHO2- MLLT10- MLLT10- SSSCA1- LYAR- HNRNPH3- SYCP1- MLH1- SMC3-
SYCP3- PIWIL1- PGK2- ACR- BMI1- DUSP6- PAX7- HORMAD1- LDHC- SPO11- MYBL1- PIWIL2- RBM5- YBX1- ASB9- ISOC1- NMT2- MEIOB- TEX101- PHOSPHO2- TPPP3- SPACA1- IQCF1- SSSCA1- LYAR- MLH1-
SYCP3- PGK2- ACR- PRM1- BMI1- SALL4- PAX7- MYBL1- TNP1- CA2- YBX1- TEKT5- ASB9- ISOC1- NMT2- TEX101- TPPP3- C17orf98- PRSS37- TP53TG5- FSCN3- OSBP2- ODF3L2- SPACA1- IQCF1- SSSCA1- ANXA1-
PGK2- ACR- PRM1- SALL4- PAX7- MYBL1- TNP1- CA2- SAXO1- YBX1- TEKT5- ASB9- PRSS37- PRSS58- TP53TG5- FSCN3- OSBP2- ODF3L2- IQCF1-
PRM1- PAX7- ASB9- PRSS37- PRSS58- OSBP2- ODF3L2-
TCN2- TPPP3- PTN- DCN- LUM- LTBP4- MGP- ANXA1- MX1- EGR1- COL3A1- C7- MX1- EGR1-

In this analysis, we select cells that are in the first stage (Spermatogonial Stem Cells (SSC's)-Stem Progenitor) and then construct the pseudo time analysis. By using the significant genes, we can find significant cells that contain significant genes. We will make a cutoff point here to subset the cells. The next step is to find the nearest nodes(points) on the trajectories for the significant cells. We order those nodes and construct the pseudo time analysis.

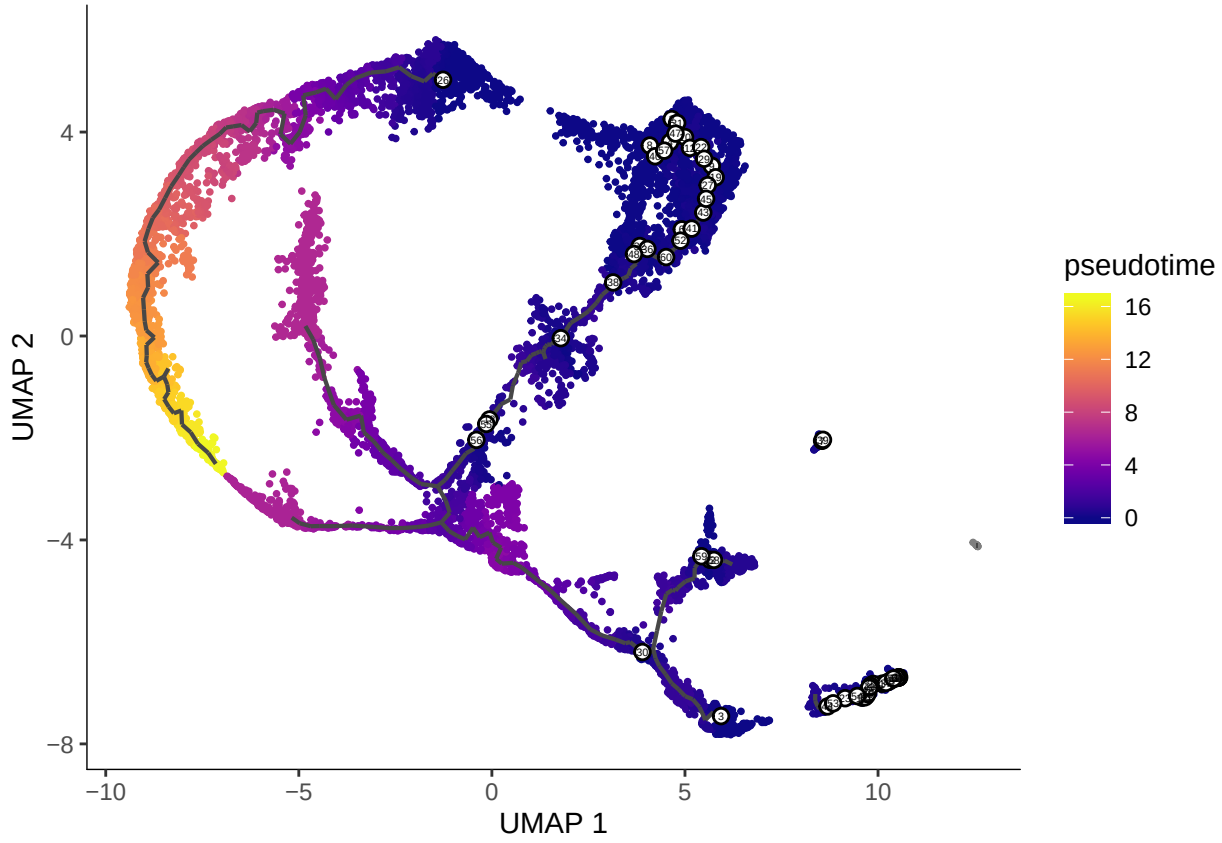
The above figure includes all the gene short names at the stage: Spermatogonial Stem Cells (SSC's)-Stem Progenitor. We choose the starting points based on these genes.

	gene_names
1	ID4
2	HSPA8
3	TAF4B
4	PEG10
5	RBM5
6	TCN2
7	ISOC1
8	HSD17B14
9	PHOSPHO2
10	ENPP2
11	PRDX2
12	ANXA1
13	SERPINA5



The figure shows a box plot of all the significant genes in stage 1 and their counts after log-normalization. Now, we choose a cutoff point as 2.8 (can be discussed).

2.8 Pseudo time analysis



Here is the pseudo time analysis figure with the starting points as the significant genes from stage 1 (Spermatogonial Stem Cells (SSC's)-Stem Progenitor).