Reimplementing and Improving in RaceID

Pei Lin, 66672977

School of Information Science and Technology
ShanghaiTech University

May 14, 2019

Outline

Outline

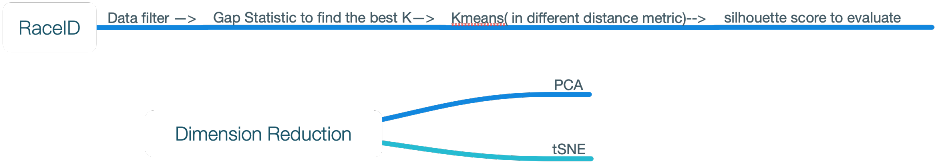
Introduction

Clustering analysis has been widely applied to single cell RNA-sequencing data to discover cell types and cell states.

RaceID is published in Nature 2014 using the data from mouse intestinal which helps to discover the rare cell type.

Introduction to RaceID

RaceID is programed by R and the workflow is summarized below:


RaceID
Data filter —>
Gap Statistic to find the best K—>
Kmeans(in different distance metric)-->
silhouette score to evaluate
Dimension Reduction
PCA
tSNE

Introduction to My Work

What I have done:

- **Reimplementing by Python.**
- **Improvement by other papers.**
- **Parallel programing.**
- **Work in another data set.**

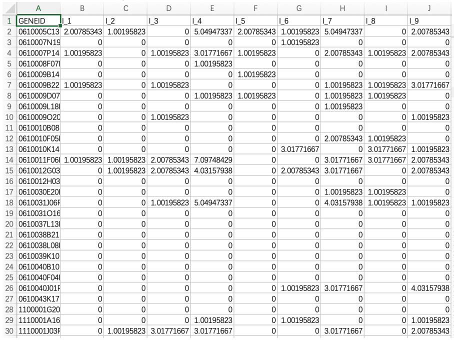
Outline

Data Type

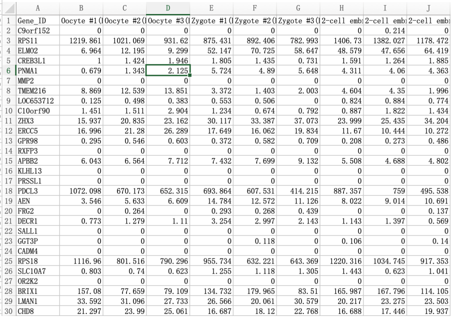
From:

1.Single-cell messenger RNA sequencing reveals rare intestinal cell types. (origin dataset)

2.Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells.



	A	B	C	D	E	F	G	H	I	J
1	GENEID	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9
2	0610005C13	2.00785343	1.00195823	0	5.04947337	2.00785343	1.00195823	5.04947337	0	2.00785343
3	0610007N19	0	0	0	0	0	1.00195823	0	0	0
4	0610007P14	1.00195823	0	1.00195823	3.01771667	1.00195823	0	2.00785343	1.00195823	2.00785343
5	0610008F07	0	0	0	1.00195823	0	0	0	0	0
6	0610009B14	0	0	0	0	1.00195823	0	0	0	0
7	0610009B22	1.00195823	0	1.00195823	0	0	0	1.00195823	1.00195823	3.01771667
8	0610009D07	0	0	0	1.00195823	1.00195823	0	1.00195823	1.00195823	0
9	0610009L18	0	0	0	0	0	0	1.00195823	0	0
10	0610009O20	0	0	1.00195823	0	0	0	0	0	1.00195823
11	0610010B08	0	0	0	0	0	0	0	0	0
12	0610010F05	0	0	0	0	0	0	2.00785343	1.00195823	0
13	0610010K14	0	0	0	0	0	3.01771667	0	3.01771667	1.00195823
14	0610011F06	1.00195823	1.00195823	2.00785343	7.09748429	0	0	3.01771667	3.01771667	2.00785343
15	0610012G03	0	1.00195823	2.00785343	4.03157938	0	2.00785343	3.01771667	0	2.00785343
16	0610012H03	0	0	0	0	0	0	0	0	0
17	0610030E20	0	0	0	0	0	0	1.00195823	1.00195823	0
18	0610031J06	0	0	1.00195823	5.04947337	0	0	4.03157938	1.00195823	1.00195823
19	0610031O16	0	0	0	0	0	0	0	0	0
20	0610037L13	0	0	0	0	0	0	0	0	0
21	0610038B21	0	0	0	0	0	0	0	0	0
22	0610038L08	0	0	0	0	0	0	0	0	0
23	0610039K10	0	0	0	0	0	0	0	0	0
24	0610040B10	0	0	0	0	0	0	0	0	0
25	0610040F04	0	0	0	0	0	0	0	0	0
26	0610040J01	0	0	0	0	0	1.00195823	3.01771667	0	4.03157938
27	0610043K17	0	0	0	0	0	0	0	0	0
28	1100001G20	0	0	0	0	0	0	0	0	0
29	1110001A16	0	0	0	1.00195823	0	1.00195823	0	0	1.00195823
30	1110001J03	0	1.00195823	3.01771667	3.01771667	0	0	3.01771667	0	2.00785343



	A	B	C	D	E	F	G	H	I	J
1	Gene_ID	Oocyte #1(	Oocyte #2(	Oocyte #3(	Zygote #1(	Zygote #2(	Zygote #3(	2-cell emb	2-cell emb	2-cell emb
2	C9orf152	0	0	0	0	0	0	0	0.214	0
3	RPS11	1219.861	1021.069	931.62	875.431	892.406	782.993	1406.73	1382.027	1178.472
4	ELMO2	6.964	12.195	9.299	52.147	70.725	58.647	48.579	47.656	64.419
5	CREB3L1	1	1.424	1.946	1.805	1.435	0.731	1.591	1.264	1.885
6	PNMA1	0.679	1.343	2.125	5.724	4.89	5.648	4.311	4.06	4.363
7	MMP2	0	0	0	0	0	0	0	0	0
8	TMEM216	8.869	12.539	13.851	3.372	1.403	2.003	4.604	4.35	1.996
9	LOC653712	0.125	0.498	0.383	0.553	0.506	0	0.824	0.884	0.774
10	C10orf90	1.451	1.511	2.904	1.234	0.674	0.792	0.887	1.822	1.434
11	ZHX3	15.937	20.835	23.162	30.117	33.387	37.073	23.999	25.435	34.204
12	ERCC5	16.996	21.28	26.289	17.649	16.062	19.834	11.67	10.444	10.272
13	GPR98	0.295	0.546	0.603	0.372	0.582	0.709	0.208	0.273	0.486
14	RXFP3	0	0	0	0	0	0	0	0	0
15	APBB2	6.043	6.564	7.712	7.432	7.699	9.132	5.508	4.688	4.802
16	KLHL13	0	0	0	0	0	0	0	0	0
17	PRSSL1	0	0	0	0	0	0	0	0	0
18	PDCL3	1072.098	670.173	652.315	693.864	607.531	414.215	887.357	759	495.538
19	AEN	3.546	5.633	6.609	14.784	12.572	11.126	8.022	9.014	10.691
20	FRG2	0	0.264	0	0.293	0.268	0.439	0	0	0.137
21	DECR1	0.773	1.279	1.11	3.254	2.997	2.143	1.143	1.397	0.569
22	SALL1	0	0	0	0	0	0	0	0	0
23	GGT3P	0	0	0	0	0.118	0	0.106	0	0.14
24	CADM4	0	0	0	0	0	0	0	0	0
25	RPS18	1116.96	801.516	790.296	955.734	632.221	643.369	1220.316	1034.745	917.353
26	SLC10A7	0.803	0.74	0.623	1.255	1.118	1.305	1.443	0.623	1.041
27	OR2K2	0	0	0	0	0	0	0	0	0
28	BRIX1	157.08	77.659	79.109	134.732	179.965	83.51	165.987	167.796	114.105
29	LMAN1	33.592	31.096	27.733	26.566	20.061	30.579	20.217	23.275	23.503
30	CHD8	21.297	23.99	25.061	16.687	18.12	22.768	16.688	17.446	19.937

Data Filtering & scaling

For genes: expressed in most cells(housekeeping gene)

For cells: total expression is too low.

For scaling:

$$\frac{GeneExpression}{Mean} \rightarrow \log(GeneExpression)$$

Outline

Kmeans

For Kmeans algorithm is not a convex problem, one question is how to find K and choose the initial point.

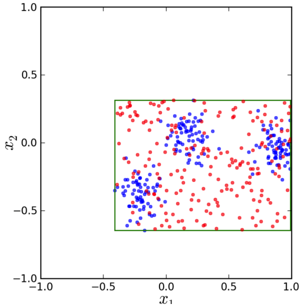
K: Calinski-Harabasz method, gap statistic

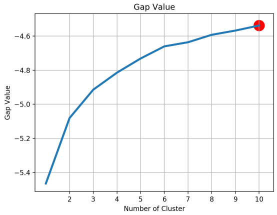
Initial point: kmeans++ method (supported by python package: sklearn.cluster)

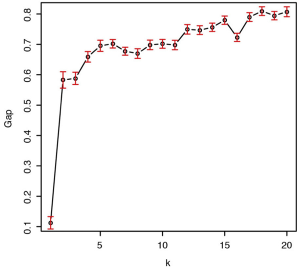
Gap Statistic

Main idea is Monte Carlo method,

$$Gap_n(k) = E(logW^*) - logW$$







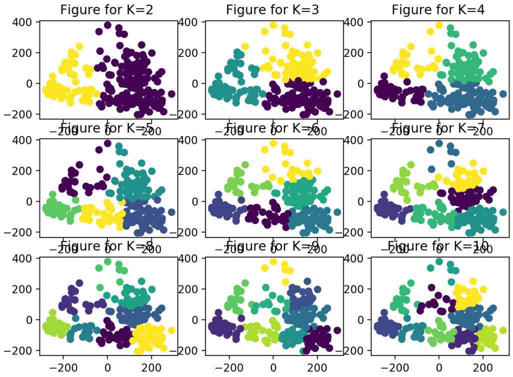
Because K-means is unsupervised learning, gap statistic is just a reference.

Outline

PCA

$$X = U \Sigma V^T$$

PCA works for linear data.



tSNE

$$P_{j|i} = \frac{\exp(-||x_i - x_j||^2/2{\sigma_i}^2)}{\sum_k \exp(-||x_i - x_k||^2/2{\sigma_i}^2)}$$

$$q_{j|i} = \frac{\exp(-||y_i - y_j||^2)}{\sum_k \exp(-||y_i - y_k||^2)}$$

$$C = \sum KL(P_i||Q_i)$$

tSNE usd manifold model and keep the characteristics of neighboring domain.

tSNE

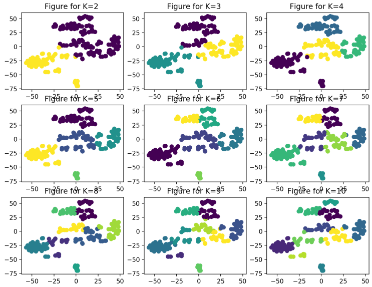
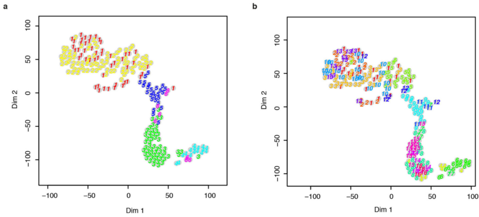


Figure for K=2
Figure for K=3
Figure for K=4
Figure for K=5
Figure for K=6
Figure for K=7
Figure for K=8
Figure for K=9
Figure for K=10


a
b
Dim 1
Dim 2

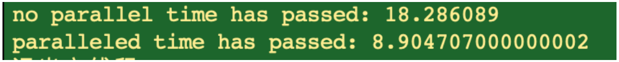
Outline

Multi-thread

Map reduce, Spark, Openmp are popular in parallel programing.
Use 10 threads to accelerating the clustering

$$MeanSpeedUp = 49\%$$

Amdahl's Law



```
no parallel time has passed: 18.286089
paralleled time has passed: 8.904707000000002
```

In R, finish the same work needs 14.2576s

Outline

Reference

[1]Single-cell messenger RNA sequencing reveals rare intestinal cell types, Dominic Gru, Anna Lyubimova, Lennart Kester
[2]Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells, Liying Yan, Mingyu Yang, Hongshan Guo, Lu Yang
[3]Estimating the number of clusters in a data set via gap statistic, Robert Tibshirani, Guenther Weather
[4]Visualizing data using t-SNE, Laurens van der Maaten, Geoffery Hinton

Thank You.