TPK: a single-cell clustering algorithm based on novel feature selection genes

Yaxuan Cui1*, Kunjie Luo1, Zheyu Zhang1 and Saijia Liu2
1 Computer and Information Engineering College, Tianjin Normal University, Tianjin, 300387, China
2 Computer and Information College, Hohai University, Jiangsu Province, Nanjing, 211100, China
*Corresponding author’s e-mail: 1052495127@qq.com

Abstract. With the continuous development of single-cell sequencing technology, through the gene expression data obtained by single-cell sequencing technology, we can have a deeper understanding of the heterogeneity between cells and the underlying mechanisms that exist between cells. However, due to the complexity of the data, single-cell identification and clustering have also brought us huge challenges. We found that many classic clustering algorithms performed poorly in single-cell clustering. Our research found that the key reason was that no mark was found. gene. First remove genes with low expression levels, and then calculate the variance value of genes, select the top 1000 genes with the largest variance, and then perform a T test to remove noise. Finally, the obtained genes are clustered using Cosine similarity algorithm and k-means. Found that it has a good clustering performance.

1. Introduction
The cell can be considered as the basic unit in biology. Biologists know that multicellular organisms are characterized by a large number of different cell types. Although the concept of cell type is intuitively clear, a consistent and strict definition is still elusive. The generation of microfluidics technology made it possible to separate a large number of cells. With the improvement of RNA isolation and amplification methods, the first-generation DNA sequencing technology [1] (Sanger sequencing) born in 1977 directly promoted the completion of the Human Genome Project in 2000. With the development of technology, second-generation sequencing [2] and third-generation sequencing [3] have greatly increased the number of sequencing data and gene expression data. This provides a wealth of data for single-cell clustering. Clustering analysis is to divide the samples according to their similarity to obtain different groups to achieve the idea of clustering by objects. Belongs to a kind of unsupervised learning. Clustering algorithms are very important in many fields. But we found many classic clustering algorithms. The effect on gene expression data is not very good. We found that many problems are not the cause of the clustering algorithm, but a problem in the selection of genes [4]. Feature genes can reduce noise and speed up calculations. Feature selection includes identifying genes with the most abundant information, and selecting the wrong gene leads to poor clustering results.

Therefore, our research first needs to screen genes, build a model to select mark genes from many genes [5], and many current clustering algorithms need to determine the number of clusters by
themselves. This greatly affects the clustering effect. We use PCA to reduce the dimensionality to estimate the number of cells and then cluster by K-means. It is easier to obtain reliable results.

2. Data and experimental environment

2.1. Mouse cell gene expression data
After studying the huge difference between cancer cell genes and normal expression, this is due to cancer cells produced by certain diseased genes [6], so many classic traditional clustering algorithms can easily cluster them and get very good results. This is not enough to prove the reliability of our algorithm. Therefore, we decided to use mouse cells from the study to cluster, which can better prove the stability of our algorithm.

2.2. Experimental environment
All data in the experiment were debugged and run on a CentOS system server with Intel Xeon E2244G 3.80GHz 32 Cores and 128G memory.

3. Methods and steps
The algorithm model of this study is shown in Fig.1.

3.1. Data preprocessing
We first need to remove the noise data, and remove the data with more than 10% of the gene expression level 0 in each row [7]. Then analyze the data. Because the genetic difference is too obvious, we take log2 for the overall data, which reduces the absolute value of the data and facilitates calculation. The difference in different intervals in the entire range of the data has different effects. Prevent noise caused by too large values in the data.

3.2. Selecting characteristic gene
First, we first calculate the variance of each row of genes, and select the top 1000 genes in order from largest to smallest. Because a large number of papers point out that the mark genes that affect the
clustering results will not exceed 1000, use PCA principal components to estimate the number of clusters, but there is still a lot of noise. The t-test is mainly used for small sample sizes. The t-test uses the t-distribution theory to infer the probability of the difference, so as to compare whether the difference between two averages is significant. We use the T test (shown in formula (1)) method such as formula (1) to eliminate the noise gene and get the mark gene.

\[
t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}
\]

(1)

\(\overline{X}_1\) represents the average value of each row of genes, \(\overline{X}_2\) represents the average value of genes predicted by PCA in each cluster, \(S_1^2\) represents the variance of each row of genes, \(S_2^2\) represents the variance of genes predicted by PCA in each cluster.

3.3. Clustering Algorithm

The mark gene and cosine similarity will be obtained. The similarity is evaluated by calculating the cosine value of the angle between two vectors. Two vectors are created and the cosine value of the two vectors is calculated. It is similarity commonly used in statistical methods. It is a very effective method of degree. The cosine measurement algorithm such as formula (2) is used to calculate the similarity between them, and the obtained measurement matrix is clustered using k-means such as formula (3).

\[
\cos(\theta) = \frac{\sum_{i=1}^{n} (x_i \times y_i)}{\sqrt{\sum_{i=1}^{n} (x_i)^2} \times \sqrt{\sum_{i=1}^{n} (y_i)^2}}
\]

(2)

\(x_i\) represents the vector of each row of genes, \(y_i\) represents the vector of each column of genes, and obtains a square matrix of metrics.

\[
E = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2
\]

(3)

\(\mu_i\) is the center \(C_i\) point of the cluster, cluster the data set into \(k\) clusters. The steps of the experimental algorithm are shown in Fig.2.

Fig.2 Schematic diagram of k-means algorithm.
4. Experimental results and analysis

4.1. Time complexity comparison
The data we use are mouse cell data sets yan[8], Biase[9], Goolam[10], pollen[11]. We first analyze the data size and perform statistical algorithm running time (shown in Tab.1).

| Dataset | Number of genes | Number of cells | Operation time(s) |
|---------|-----------------|-----------------|-------------------|
| yan     | 9643            | 90              | 5.456             |
| Biase   | 25737           | 49              | 4.748             |
| Goolam  | 41480           | 123             | 6.051             |
| pollen  | 23730           | 301             | 12.138            |

4.2. Estimating of k
Using PCA to predict the number of cell clusters has a good visualization result. Taking the Biase data as an example, it is obvious that the results are divided into 3 categories (shown in Fig. 3).

![Fig. 3 Schematic diagram of estimated k.](image)

The k value predicted by PCA is shown in Tab.2. It can be seen that it is basically consistent with the original true cell number.

| Dataset | Ref | TCK | SINCERA | SNN-Clip | SEURAT |
|---------|-----|-----|---------|----------|--------|
| yan     | 7   | 6   | 6       | 11       | 13     |
| Biase   | 3   | 3   | 5       | 6        | 10     |
| Goolam  | 5   | 5   | 4       | 21       | 20     |
| pollen  | 11  | 9   | 9       | 14       | 17     |

4.3. Algorithm comparison
We respectively used the SINCERA algorithm [12], the SNN-Clip algorithm [13] and the SEURAT algorithm [14] to compare with our algorithm. The evaluation index adopted is ARI. This index does not consider the clustering method you use. The method is treated as a black box, focusing only on results. It can be said that it is a very "utilitarian" indicator. This kind of evaluation index is used in most papers.
We can see that the TPK algorithm ranks first in ARI in the data of Yan and Pollen (Fig.4, Fig.7), and has the same excellent performance as in the data of Biase (Fig.6), and only ranks second in ARI in the Goolam data (Fig.5). In general, the TPK algorithm has a good performance in the accuracy of clustering and the time complexity and space complexity of the algorithm.

5. Conclusion
Unsupervised clustering may still be the core component of scRNA-seq analysis. Since most downstream analysis is based on clustering, the final conclusion may be strongly influenced by clustering. But in the foreseeable future, it is possible that more attention will be paid to the selection of genes.

This experiment adopted a new type of characteristic gene selection idea. First, the data with the number of gene expression 0 in each row exceeding 10% was eliminated. Calculate the variance of each row of genes, select the first 1000 genes with the largest variance, and take log2 for the genes to get the mark gene. Using cos to measure the similarity of the algorithm, the obtained metric matrix uses PCA to estimate the number of clusters, and uses kmeans to cluster. Finally, the ARI index is used to calculate the final clustering results. The algorithm has a good performance on mouse cell gene expression data, and it remains to be seen whether it has good performance on other species of cells. In addition, it remains to be seen whether this algorithm also performs well in processing other large-scale gene expression data.

References
[1] SANGER F, NICKLEN S, COULSON AR. DNA sequencing with chain-terminating inhibitors [J]. Proc Natl Acad Sci USA. 1977, 74(12): 5463-5467.
[2] SLATKO BE, GARDNER AF, AUSUBEL FM. Overview of Next-Generation Sequencing Technologies[J]. Curr Protoc Mol Biol. 2018, 122(1): e59.
[3] RHOADS A, AU KF. PacBio Sequencing and Its Applications[J]. Genomics Proteomics Bioinformatics. 2015,13(5):278–289.
[4] Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 9, 75 (2017).
[5] Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161, 1202–1214 (2015).
[6] Augsten, M. Cancer-associated fibroblasts as another polarized cell type of the tumor microenvironment. Front. Oncol. 4, 62 (2014).
[7] Wolf, F. A. et al. Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Preprint at bioRxiv https://doi.org/10.1101/208819 (2017).
[8] Yan, L. et al. Nat. Struct. Mol. Biol. 20, 1131–1139 (2013).
[9] Biase, F.H., Cao, X. & Zhong, S. Genome Res. 24, 1787–1796 (2014).
[10] Goolam, M. et al. Cell 165, 61–74 (2016).
[11] Pollen, A.A. et al. Nat. Biotechnol. 32, 1053–1058 (2014).
[12] Guo, M., Wang, H., Potter, S.S., Whitsett, J.A. & Xu, Y. PLoS Comput. Biol.11, e1004575 (2015).
[13] Xu, C. & Su, Z. Bioinformatics https://doi.org/10.1093/bioinformatics/ btv088 (2015).
[14] Macosko, E.Z. et al. Cell 161, 1202–1214 (2015).