# Group project 2 Report

Liucheng Shi, Zhuohui Liang, Ruwen Zhou, Jiying Han

## Introduction

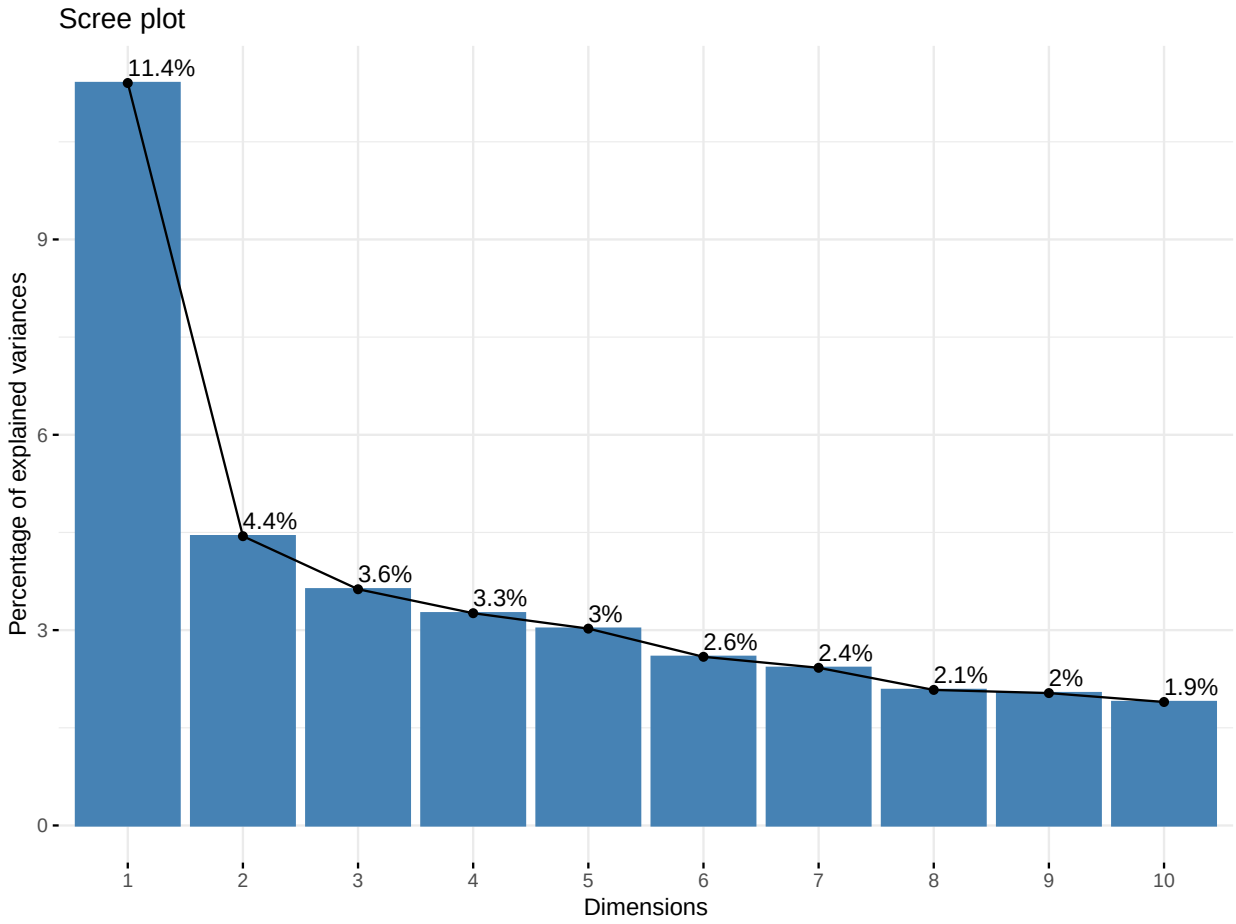
The recent breakthrough in NGS allows us to sequence thousands of RNA simultaneously at individual cell level, which leads to possible insight in heterogeneity in gene expression by dividing cells into subgroups based on their depth of coverage. The objective of this project is to identify the hidden structure in 558 genes using the 716 scRNA sequencing data from breast cancer tumor. We would use clustering method based on GMM model with EM algorithm, in comparison to other methods including hierarchical clustering method.

## Method

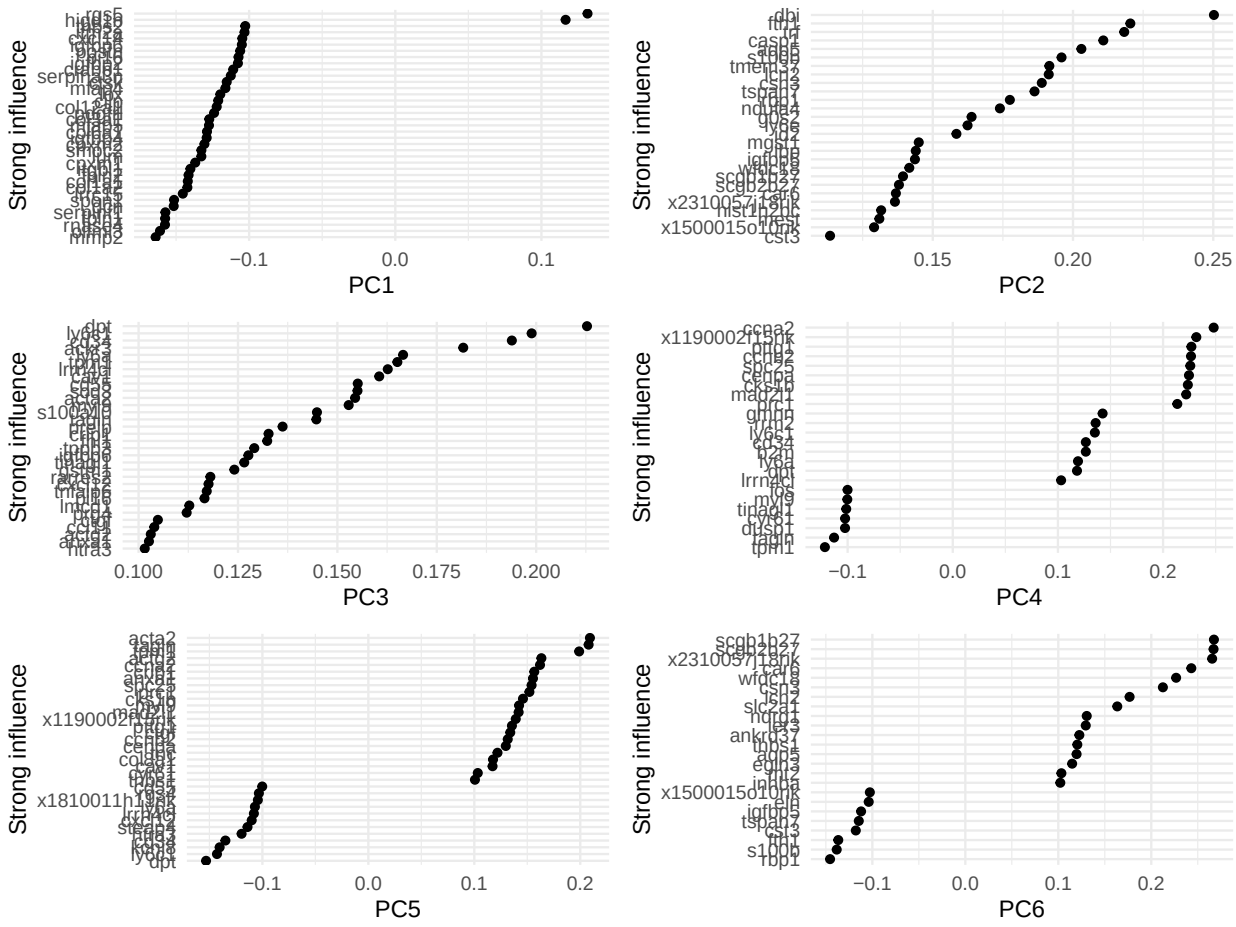
### Data Preparation

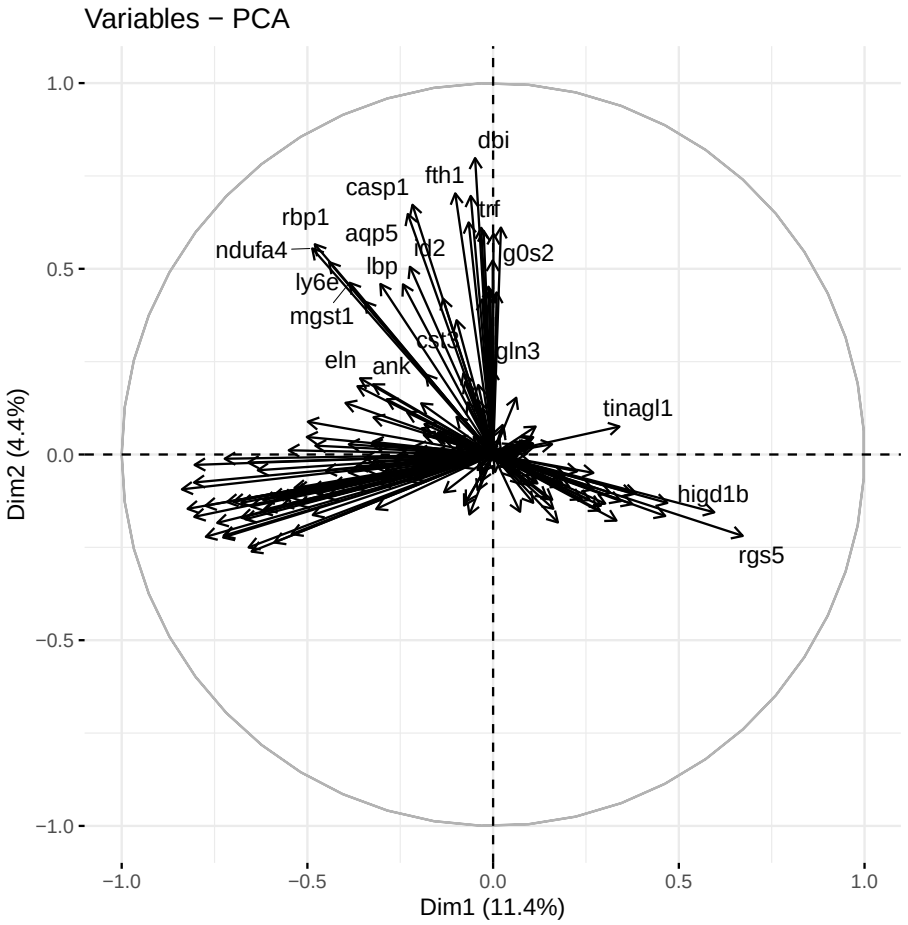
Raw single-cell sequencing data is inflated with 0. According to Pierson (2015), PCA using zero-inflated is a less optimal approach in dimension reduction.Besides, zero-inflation(drop-out) event affecting the converging of EM clustering. Followed the adjustment in the paper, we dropped genes with over 90% of zero inputs. 229 genes is used for further PCA and clustering after filtering.

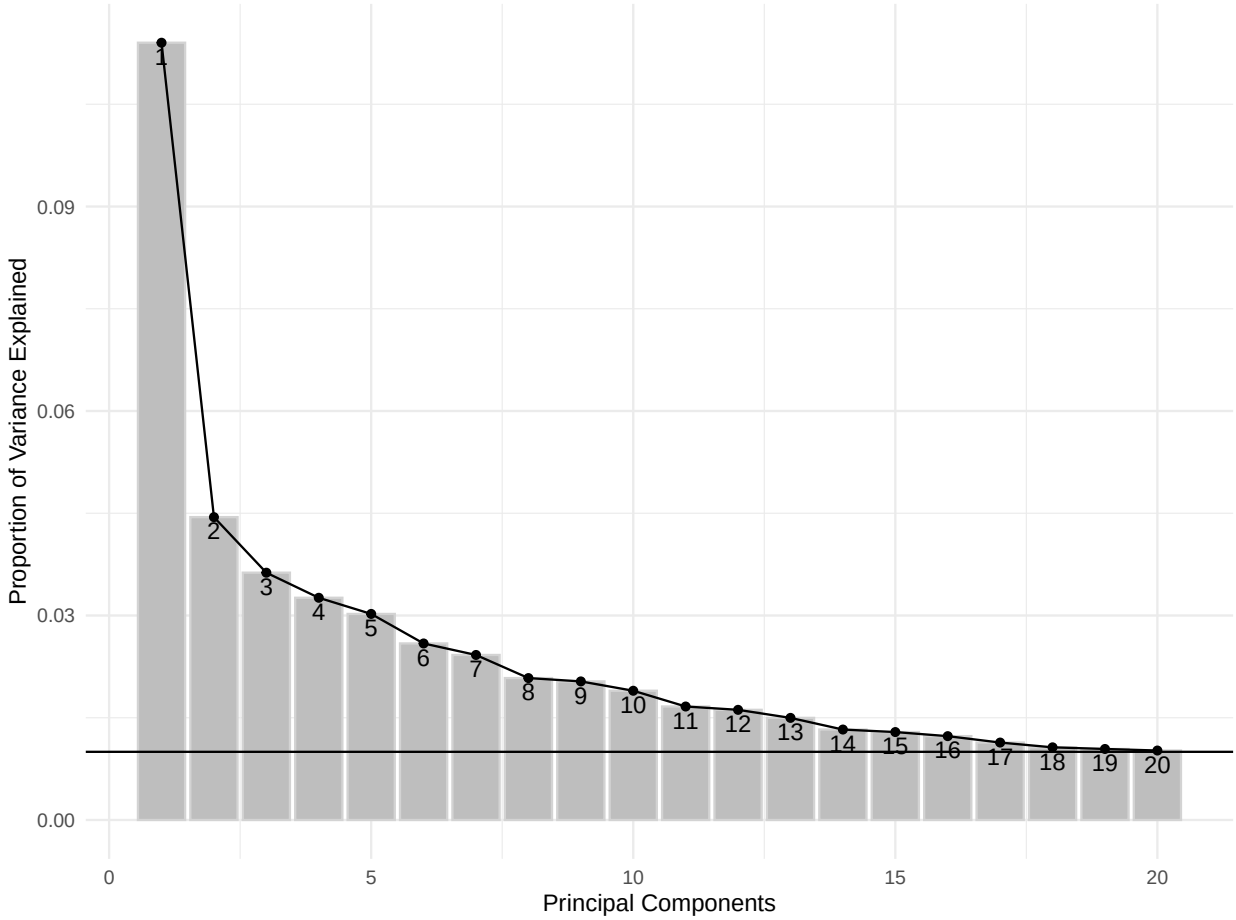
## PCA




Since all depths of coverage and other statistics for gene expression is measured in the same scale, one might have such the intuitive thought that we should not center and scale the single cell expression data. Although both scaled and unscaled methods are implemented in literature, scaling the data would provide us with more details on genome subtypes. Considering gene A required expression level 1000 to be activated and gene B required expression 100, two genes are not comparable and gene A would contribute more to the PCs for bigger variance. Based on the plot, we can see the variance is significantly lower by centering and scaling the expression level.


Strong influence
PC1
PC2
PC3
PC4
PC5
PC6


Variables − PCA
dbi
fth1
casp1
tmf
rbp1
aqp5
rd2
g0s2
ndufa4
lbp
ly6e
mgst1
cst3
gln3
eln
ank
tinagl1
higd1b
rgs5
1.0
0.5
0.0
−0.5
−1.0
Dim2 (4.4%)
Dim1 (11.4%)




The genes contributed the most in the first six PCs are shown below, with absolute value of coefficient larger than 0.1. Based on the proportion of variance explained plot, only the first 20 PCs have the PVE larger than 0.01. All first 20 PCs passed the eigenvalue criterion. Thus, we choose 20 PCs to conduct further clustering.

## EM algorithm

The model is assumed to have gaussian conditional distribution in each cluster, with parameters $\mu_k$ and $\Sigma_k$ for k in 1,2,3…K. Each observation have probability $p_{ik}$ to be any of the K's clusters, observation is assigned to the cluster with the highest probability. The model is presented as followed:

$$\mathbf{x}_i \sim \begin{cases} N(\mu_1, \Sigma_1), \text{with probability } p_1 \\ N(\mu_2, \Sigma_2), \text{with probability } p_2 \\ \quad \vdots \quad , \quad \vdots \\ N(\mu_k, \Sigma_k), \text{with probability } p_k \end{cases}$$

Where i ∈ 1,2,3…N, the completed likelihood function is:

$$L(\theta; \mathbf{x}, \mathbf{r}) = \prod_{i=1}^{n} \prod_{j=1}^{k} [p_j f(\mathbf{x}_i; \mu_j, \Sigma_j)]^{r_{i,j}}$$

The EM algorithm is designed as:

describe

initialize model with random $p_j, \mu_j, \Sigma_j$ given k
while iteration is less than maximum iteration or objective function not converge
    E step: calculate the conditional probability $p(r_j|X, \mu, \Sigma)$
    M step: calculate and update $\mu, \Sigma, p$ and assign clusters
end

After optimizing, calculate observed Likelihood L. EM algorithm is optimized by the observed likelihood, but understanding that assigning each cell to its own cluster would have the model with highest likelihood, but as thus lead to overfilling problem. With above concerns, the AIC loss function($-2(log(L) - n_p)$) is used instead of deviance, where $n_p = G * (K + 1) + K - 1$. As rule of thumb, initial cluster's number is set as 2 to 10[**?**], and final cluster number is determined by model with lowest AIC.

## Hierachical clustering

Hierarchical clustering is an alternative approach to k-means clustering for identifying groups in the dataset. In our project, our purpose is to classify scRNA-seq into different clusters based on their gene expressions and identify potential existence of cell subtypes. We choose agglomerative hierarchical clustering method, which is a bottom-up method. Each object is initially considered as a single-element cluster. At each step of the algorithm, the two clusters that are the most similar are combined into a new bigger cluster. The iteration will not stop until all elements are being classified into one single cluster. This makes the result a tree and can be visualized as a dendrogram.

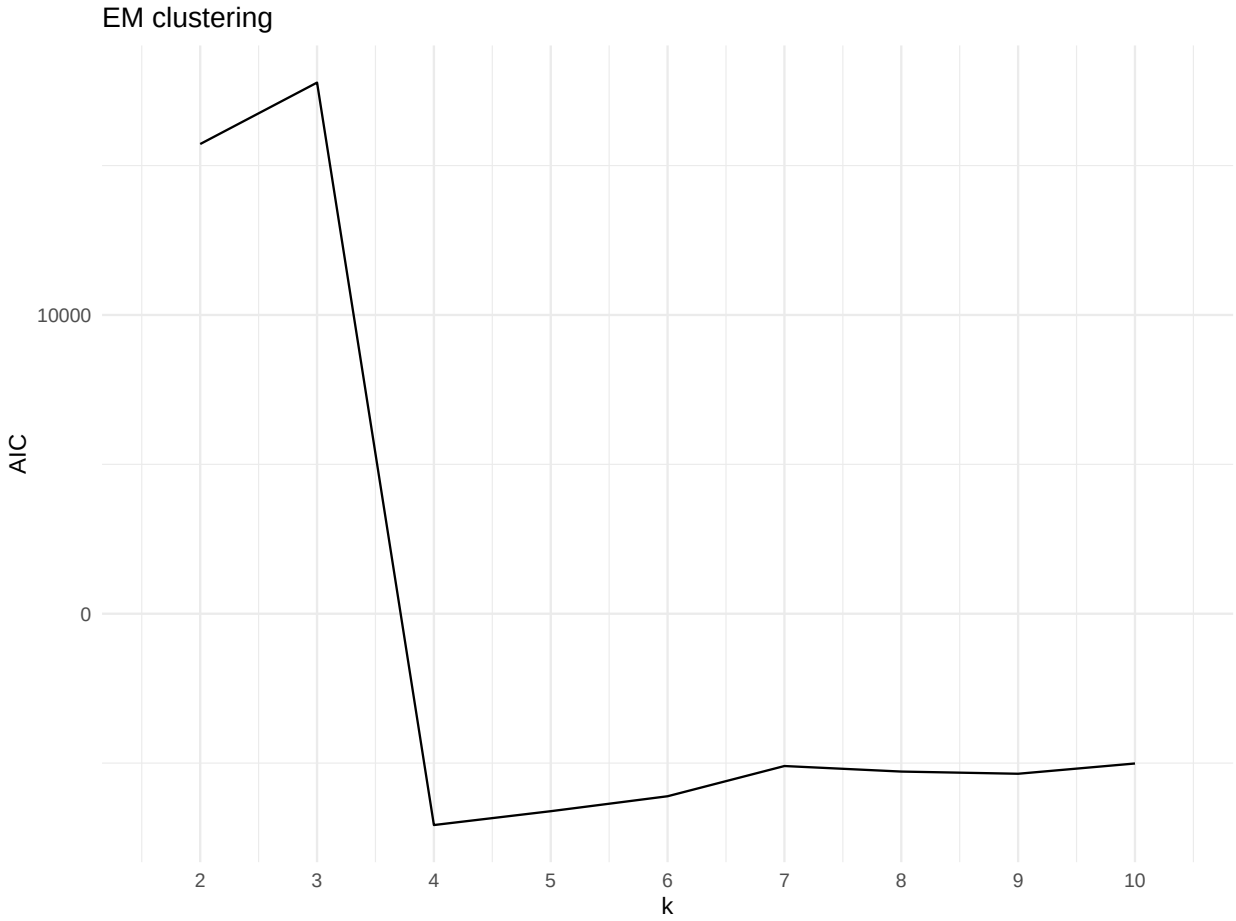
## Density method

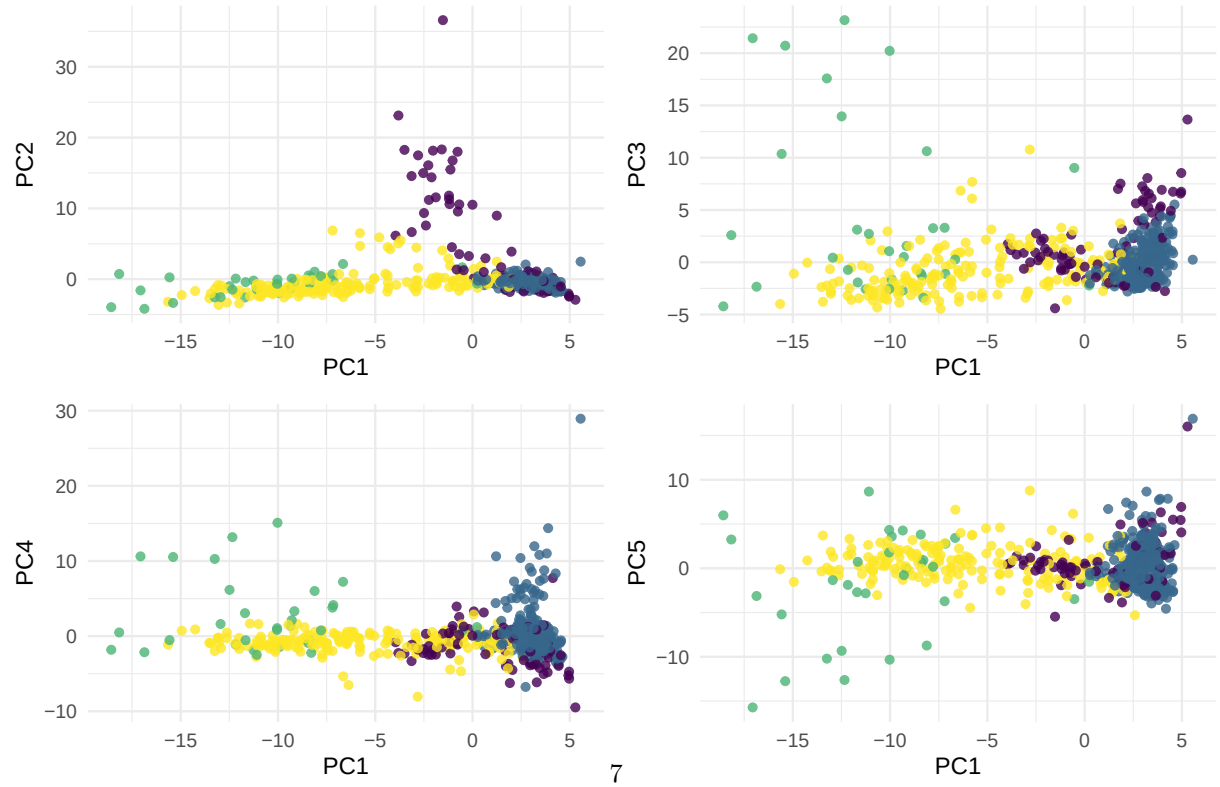
The Density-based clustering method is based on the assumption that points shared similar characteristics would cluster together in a denser format. The method initialize with a core point, and expand the cluster if there are *minPts* number of neighboring points with the radius of *epsilon*.The algorithm would stop when all points are classified as either seeds, borders, or outliers. Two parameters are the minimum number of neighboring points and the searching radius (epsilon).

## Signature selection

Two methods of gene signature selection is performed and compared. Recursively Feature Elimination Support Vector Machine(SVM-RFE) is classification method based on support vectors, and in each iteration, the predictors with smallest importance are removed until the specified subset is reached. Based on cross-validation method, the subset with best loss function, in our case-Accuracy, is selected. Another method is used is Wilcoxson test with Berforroni p-value adjustment, each cluster's differential expressions are evaluated based on rank based method. Not significant genes are removed and re-evaluated with smaller subset.

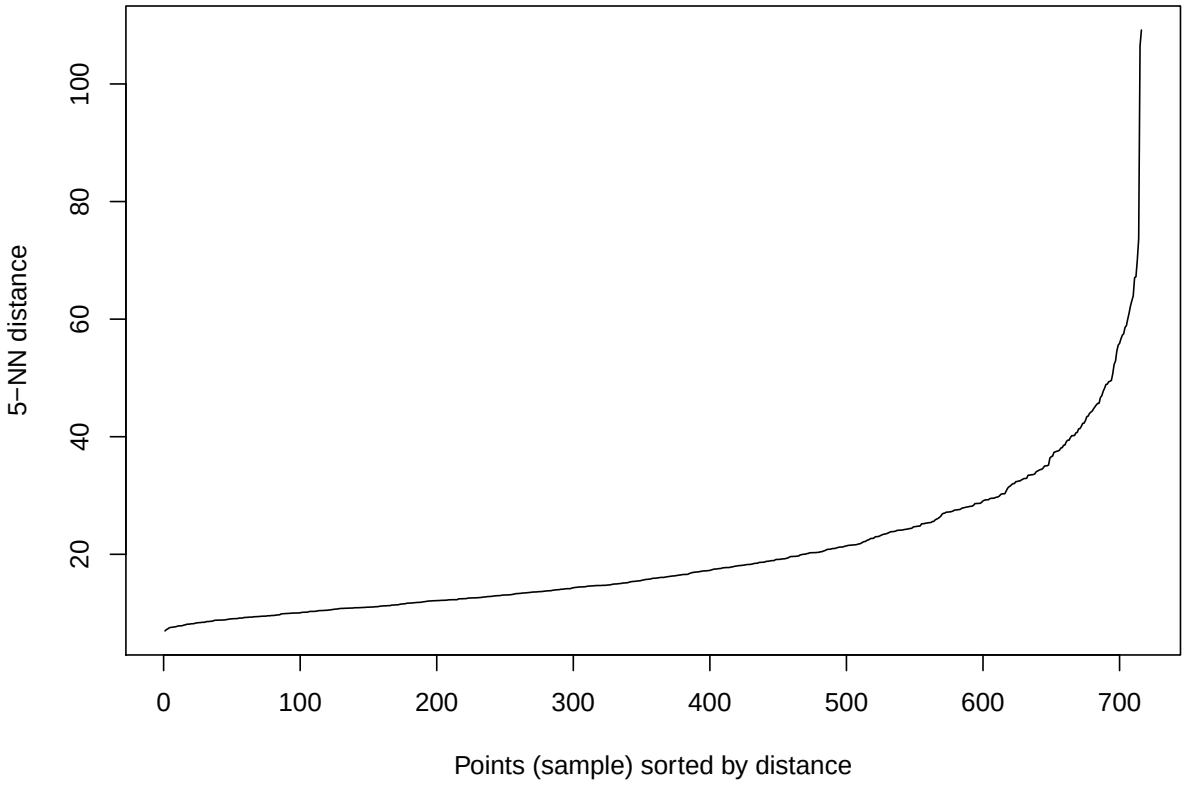
# Result

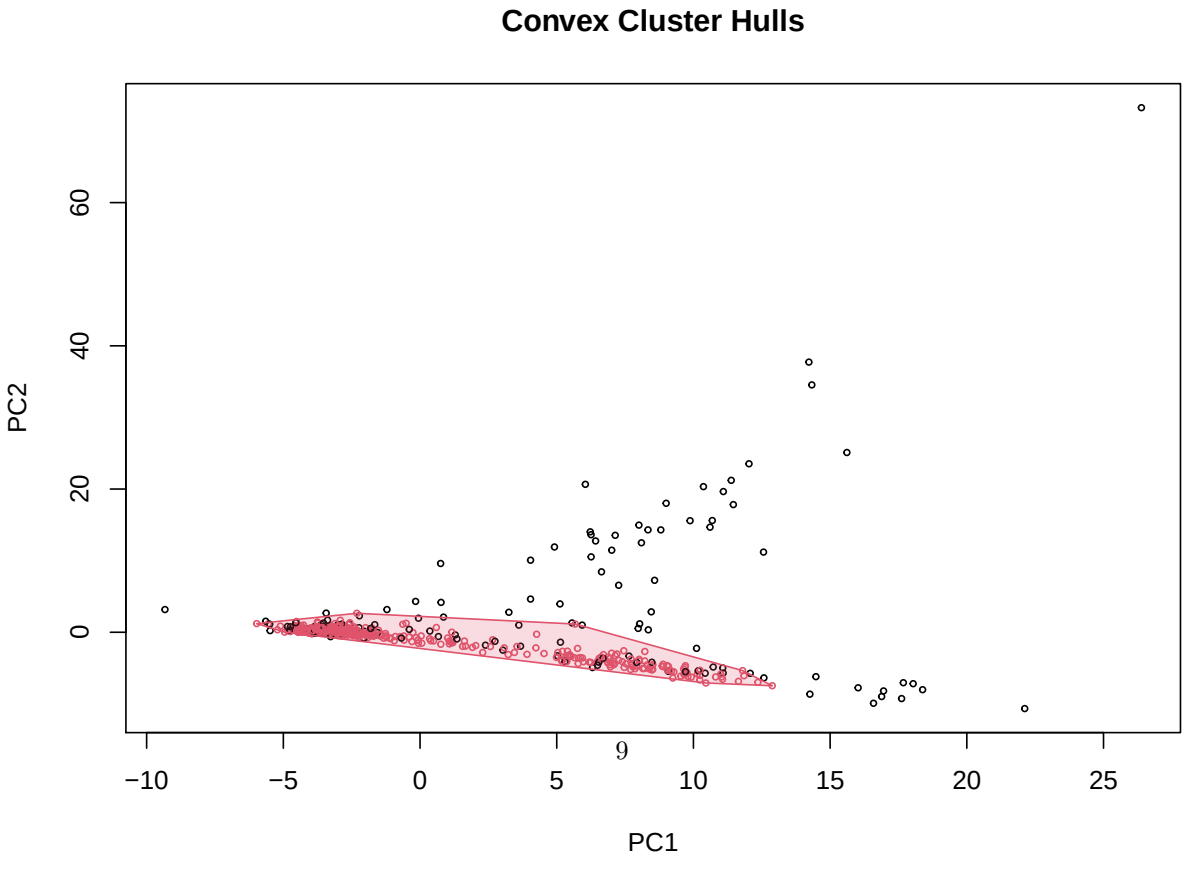

EM clustering
AIC
10000
0
2
3
4
5
6
7
8
9
10
k


PC2
PC1
PC3
PC1
PC4
PC1
PC5
PC1
cluster
1
2
3
4

The number of cluster is selected based on lowest AIC of EM algorithm, and it shows that clustering with 4 clusters has the lowest average AIC. Based on the result of single run, in the case of having 4 clusters, the proportion of cell in each cluster is 0.106, 0.598, 0.045, 0.251. With each cell assigned to cluster, PC component plot displays a well separated cluster pattern.

## Density-based clustering



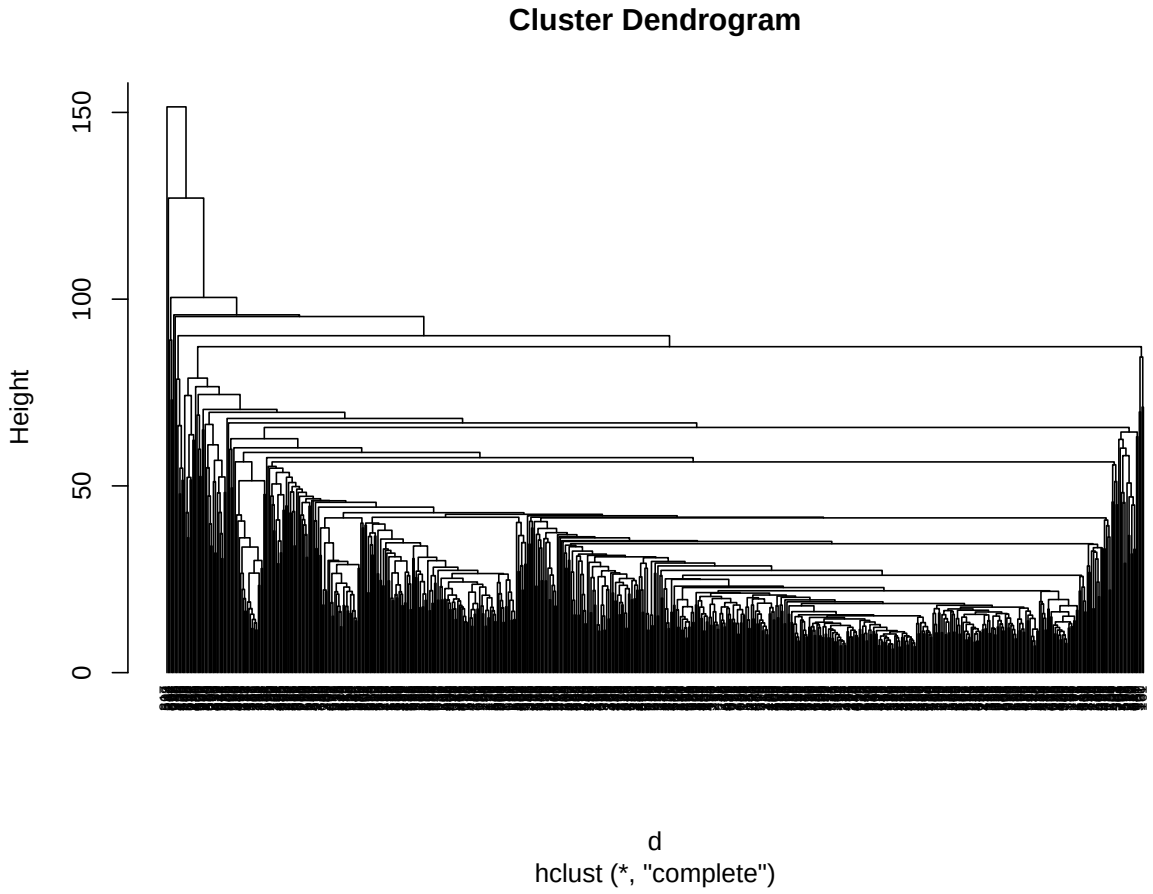



The KNN distance plot have red-flagged the incapability of clustering the . Using minPts 5, and epsilon 25, we get the result of 1 cluster containing 581 points and 135 noises. Density-based clustering fail to capture the underlying genome subtypes possible due to the increasing complication in measuring distance between genes for high-dimensional data. Other clustering methods with greater flexibility should be implemented.

## Hierarchical clustering

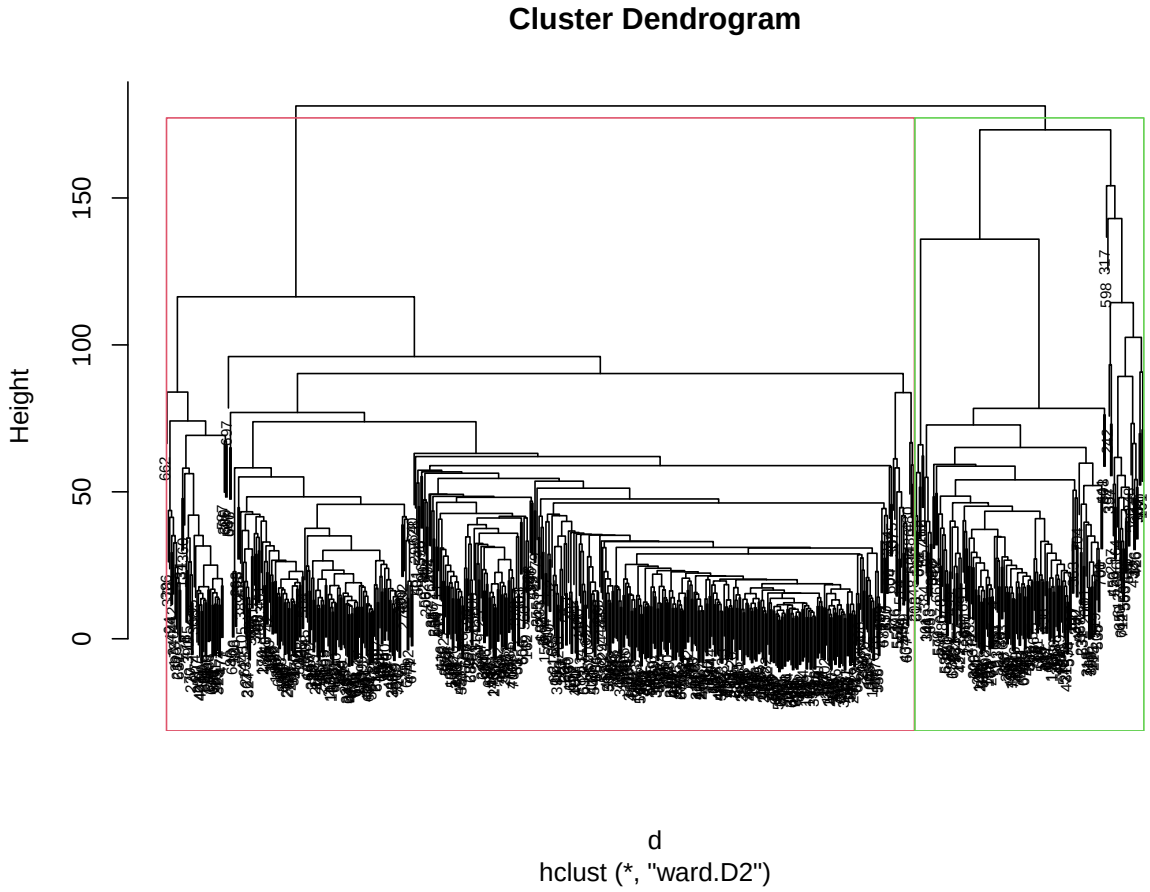
When measuring the dissimilarity between each pair of observations distance, Euclidean distance has been used. However, when measuring the dissimilarity between two clusters of observations, we have applied different methods.

- Minimum or single linkage clustering: It computes all pairwise dissimilarities between the elements in cluster 1 and the elements in cluster 2, and considers the smallest of these dissimilarities as a linkage criterion.

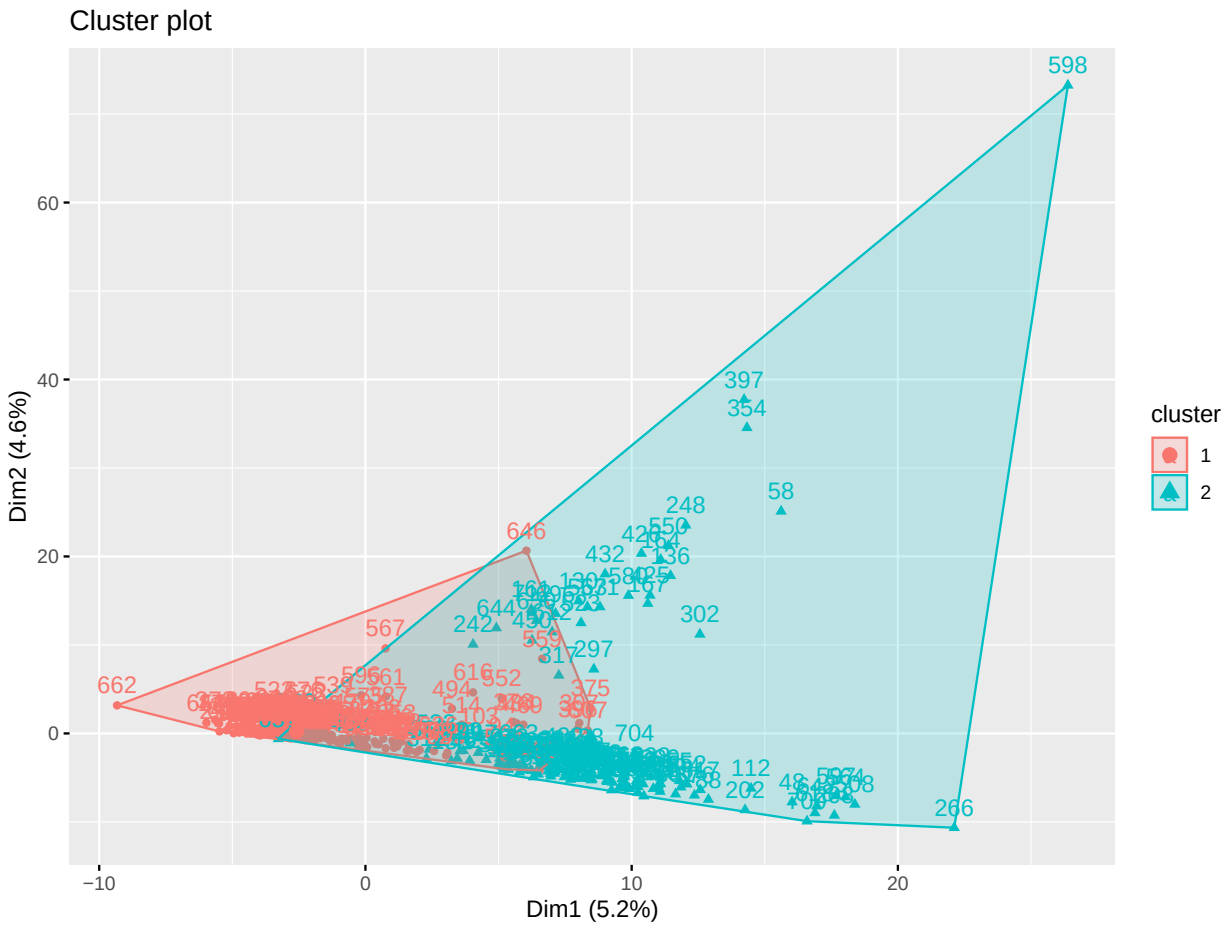



Ward's minimum variance method: It minimizes the total within-cluster variance. At each step the pair of clusters with minimum between-cluster distance are merged.

```
## sub_grp
##   1   2
## 548 168
```

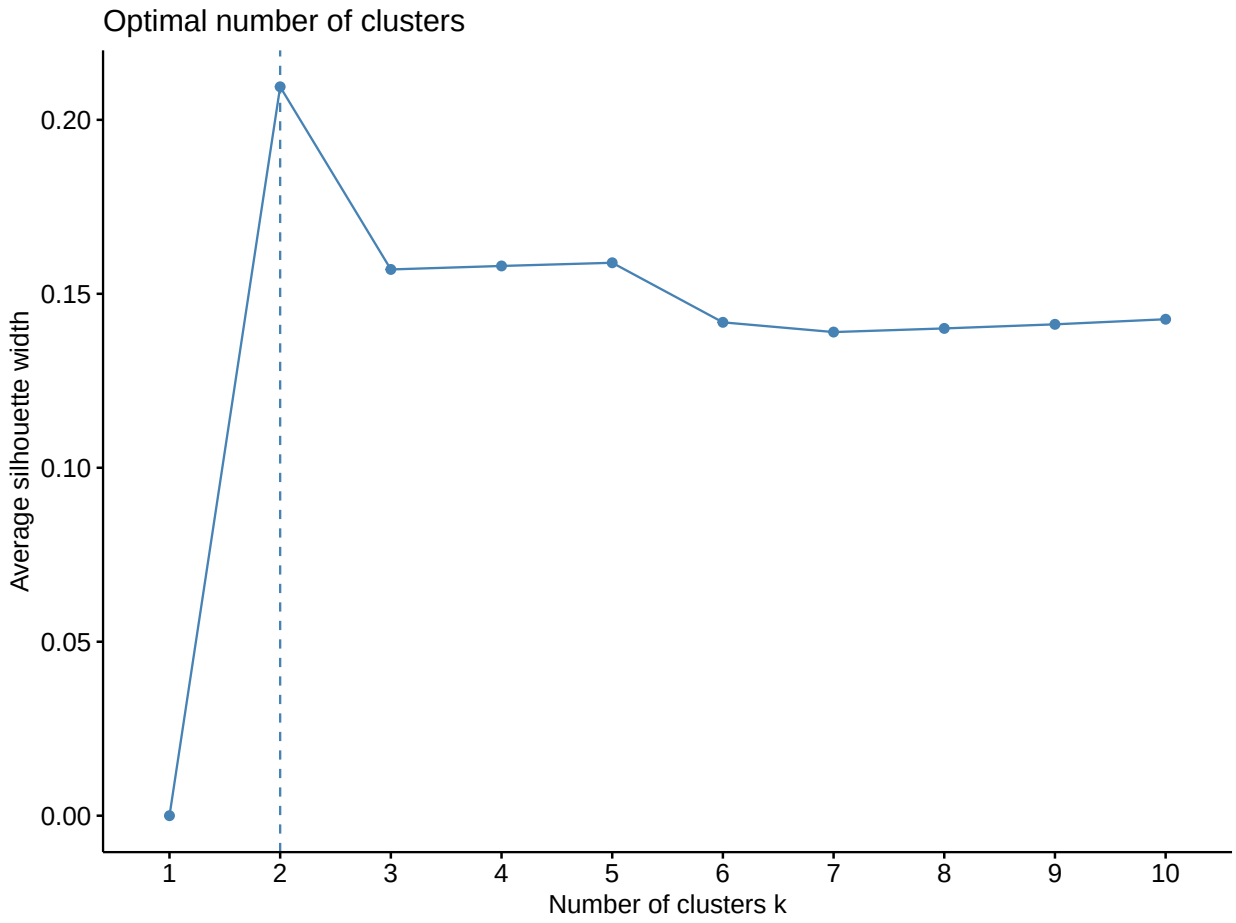



Visualize the result in a scatter plot.

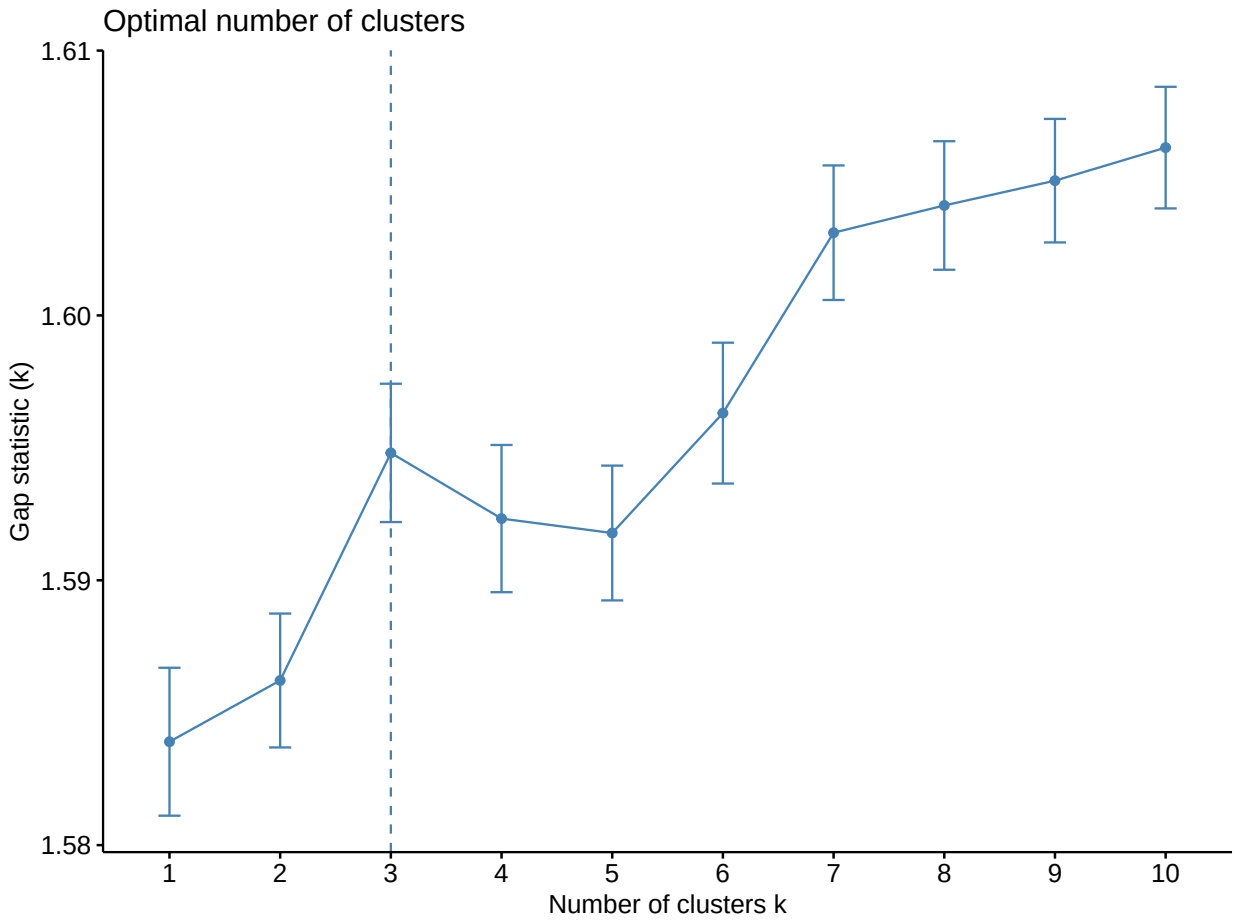



Determining Optimal Clusters

- Average Silhouette Method




- Gap Statistic Method



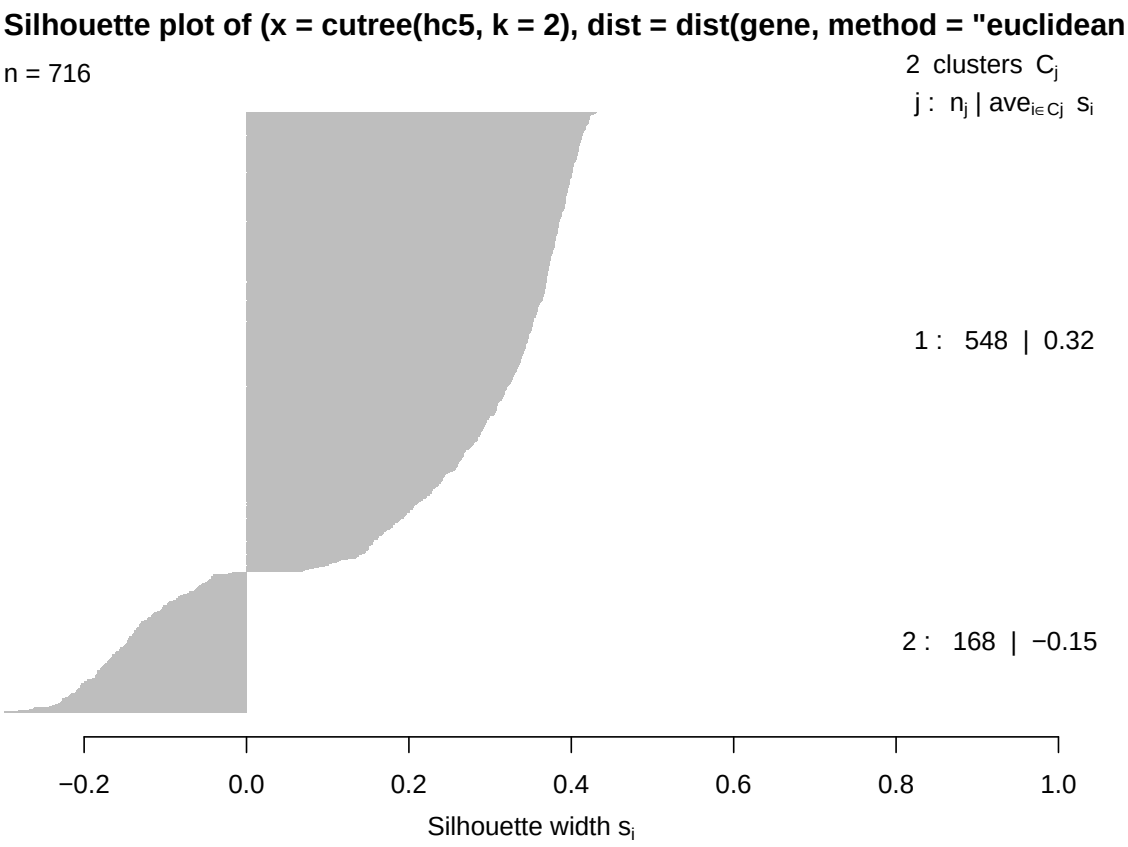

In order to determine the optimal number of clusters, we applied Average Silhouette Method and Gap Statistic Method. Silhouette Method indicates that two-clusters-model fits our data best, while Gap Statistic Method shows that three-clusters is more suitable. After applying different clusters, we found that when we choose cluster = 3, each cluster has 548, 140, 28 items. When cluster = 2, each cluster has 548, 168 items.

## Comparison

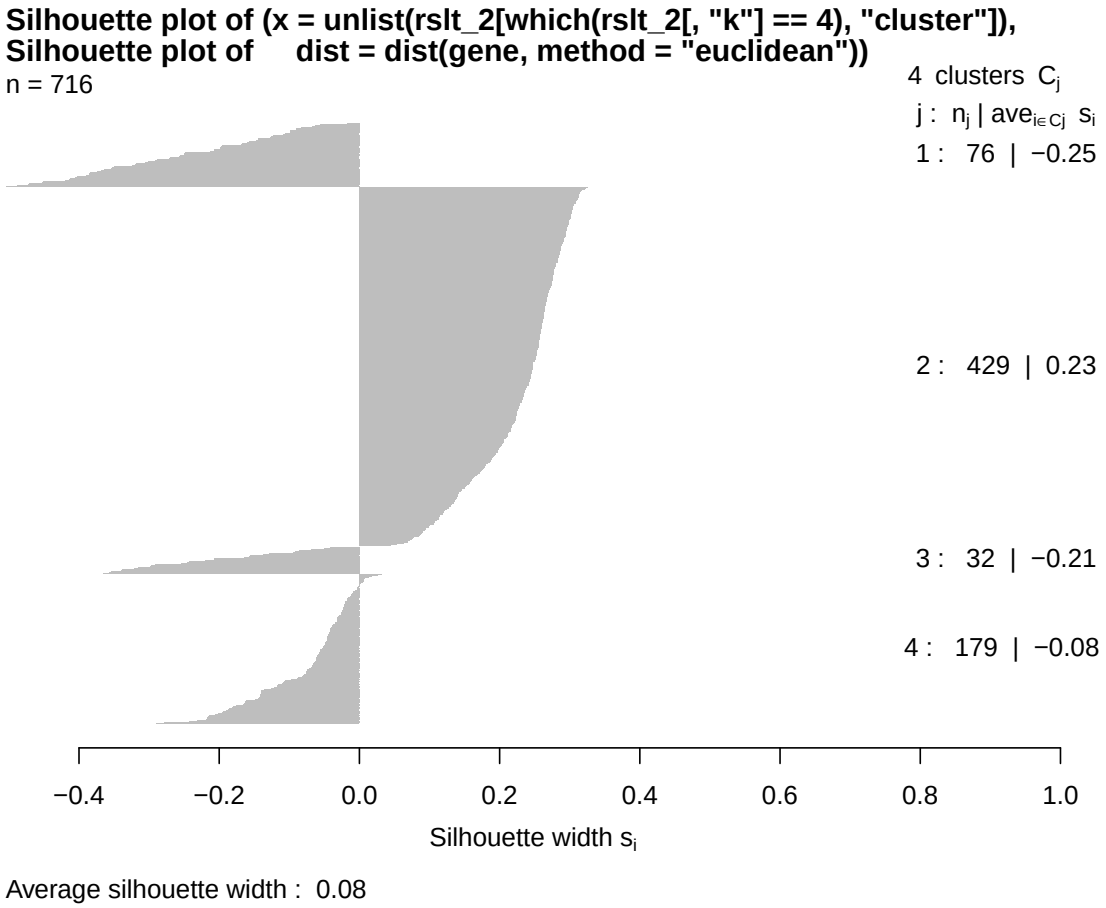
In order to compare the EM algorithm and Hierarchical clustering method, we apply the silhouette coefficient to evaluate the performance of clustering methods. Since the ground truth labels of clusters are unknown, evaluation must be performed using the model itself. The Silhouette Coefficient s for a single sample is then given as:

$$s = \frac{b - a}{max(a, b)}$$

a: The mean distance between a sample and all other points in the same class. b: The mean distance between a sample and all other points in the next nearest cluster.




Average silhouette width : 0.21



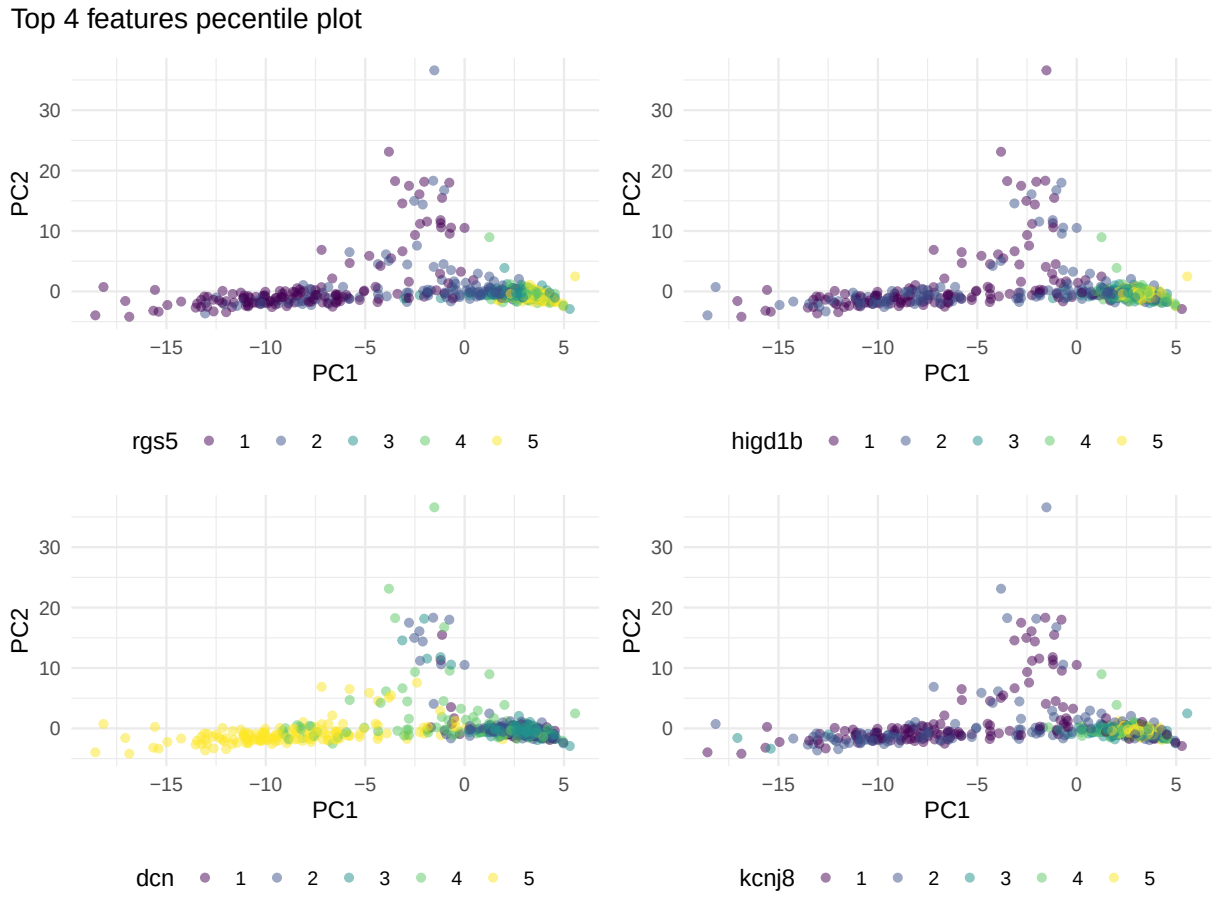

The silhouette coefficient of EM algorithm is 0.12 and silhouette coefficient of hierarchical clustering is 0.21. A higher Silhouette Coefficient score relates to a model with better defined clusters. Therefore, the performance of hierarchical clustering is better than EM algorithm.

## Gene-expression signatures

### SVM-RFE

Table 1: SVM-RFE

| cluster | number_of_signature | top_signature |
|---|---|---|
| 1 | 50 | lcn2\|\|tpm1\|\|wfdc18\|\|dbi\|\|rrad |
| 2 | 80 | rgs5\|\|higd1b\|\|kcnj8\|\|prelp\|\|serpinf1 |
| 3 | 20 | ly6a\|\|ly6c1\|\|isg15\|\|igfbp6\|\|cd34 |
| 4 | 100 | dcn\|\|serpinf1\|\|olfml3\|\|rgs5\|\|higd1b |




SVM-RFE method choosing 20 to full genes subset as signature, indicating incapable to reduce features. For each cluster, SVM selected and rank the importance of each gene, the first five gene signature for cluster 1 is lcn2, tpm1, wfdc18, dbi, rrad, rgs5, higd1b, kcnj8, prelp, serpinf1 for cluster 2, ly6a, ly6c1, isg15, igfbp6, cd34 for cluster 3 and , dcn, serpinf1, olfml3, rgs5, higd1b for cluster 4. The number of gene signatures chosen to classify all clusters is 100, which its top 4 signature are plotted against PC component in quantile scale. According to the plot, these gene display similiar distribution to the the clusters.

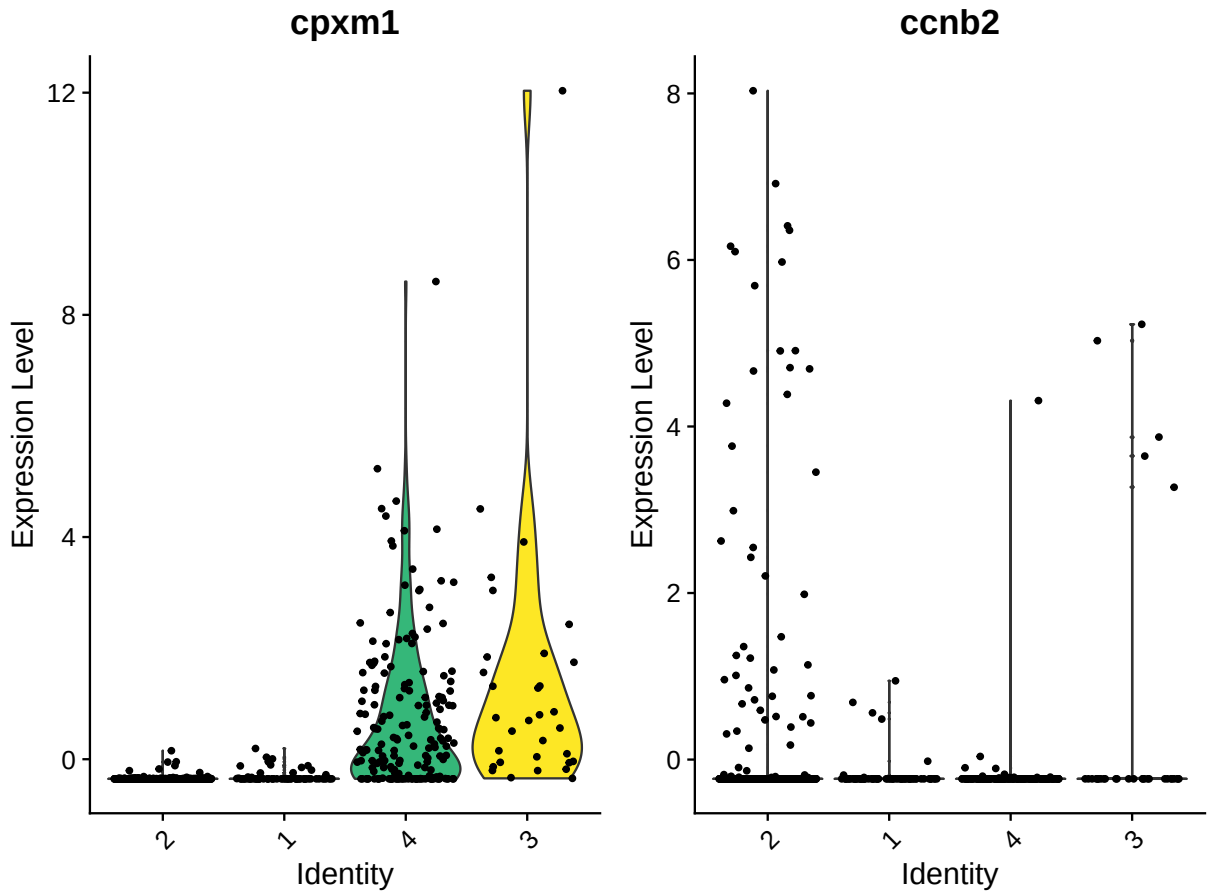
Table 2: Top 6 genes to differentiate the clusters

| cluster_1 | cluster_2 | cluster_3 | cluster_4 |
|---|---|---|---|
| timp1 | higd1b | mfap4 | serpinf1 |
| f2r | kcnj8 | ly6a | olfml3 |
| col6a1 | rgs5 | gpc3 | rgs5 |
| col1a1 | fbln2 | ly6c1 | itgbl1 |
| ctsk | rnase4 | cthrc1 | higd1b |
| rgs2 | serpinf1 | gja1 | rnase4 |

Table 3: Number of signature

| cluster | number_of_signature |
|---|---|
| 1 | 104 |
| 2 | 133 |

| cluster | number_of_signature |
|---|---|
| 3 | 151 |
| 4 | 133 |




**Seurat**

Result from gaussian mixture clustering is used as response variable and conduct classification.

Using the package Seurat, we are able to test the difference in gene expression across cluster, and select the gene signatures which shows significant different patterns among clusters by conducting Wilcoxon rank-sum test. No assumption is assigned using the non-parametric test such that it is validated to find gene signatures. Top 6 gene signatures is detected and shown above, in which each genes is sufficient to differentiate the clusters with significant different distribution for each cluster. Based on the violin plot, the gene CPXM1 in cluster 2 obviously differed with the distribution of expression level in other two clusters. Same for the gene ccnb2 in cluster 3.

Comparing signature of seurat wilcoxon rank test and svm-rfe method, wilcoxon rank test is better in terms of variation of numbers of signature, but the number of signatures are too high. Whereas SVM-RFE successfully selected fewer signature in some cases and maintain high predictabiliy. Top 5 signatures selected are also varies in two methods.

# Conclusion

In this analysis, we use Gaussian-Mixture model with Principal Component Analysis to explore our dataset.Gaussian mixture model performs soft classification, which means that it can gives us the probability that a given data point belongs to each of the possible clusters [1]. Besides, since our data takes on different shape, it's better to use Gaussian. Principal component analysis (PCA) is an essential method for analyzing single-cell RNA-seq (scRNA-seq) datasets, but for large-scale scRNA-seq datasets, since PCA algorithms and implementations load all elements of the data matrix into the memory space, which means that method computation time is long and consumes large amounts of memory. Apart from that, the workflow of PCA is redundant and repeated, which can be replaced by fast PCA algorithms, like "Julia" package[2]. Finally. as for EM algorithm, clustering cell with EM algorithm has several drawbacks. The first of all shared problem with EM algorithm is slow computational time, [**?**] has compared EM based method with hierarchical methods, and EM method is slower. The second problem is that EM algorithm's convergence and convergent time are rely on the initialization, a well separated center for initialization provide fast convergence, on the other hand, bad initialization lead to divergence. But the initialization is randomly assigned uniformly, as a result, convergence is not always guaranteed. The third problem is that gaussian model may not fit RNA sequence expression. [**?**] has pointed out that drop-out event leads to zero-inflated data, and gaussian mixture cannot address this problem. Also, [**?**] provides poisson-mixture as a alternative for gaussian mixture.. Another problem with model-based method is over-estimating. Model-based method has tendency to over-estimate number of cluster, for example, AIC method do not penalize much when K is large, leading to potential of over-estimating.

On the other hand, graphical method, hierarchical and k-means method do not have strong model assumption as EM method. According to [**?**], Graphical method `Seurat` has the fastest run time, followed by biracial method. Model based method is just faster than k-means methods.

## Reference

[1] "Gaussian Mixture Models Clustering Algorithm Explained" from https://towardsdatascience.com/gaussian-mixture-models-d13a5e915c8e [2]"Benchmarking principal component analysis for large-scale single-cell RNA-sequencing " by Koki Tsuyuzaki1etc.