

# Gene Networks of Potato

Jiachen Wang, Yuexu Wang, Jiajun Ma

Supervisor: Dr. Hong Gu

Data and Background courtesy of: Dr. Helen Tai

## 1. Introduction

### 1.1 Background introduction (written by Yuexu, checked by Jiachen)

Potato is the 4th most important crop in the world, our objective is to improve potato so that it can be easier to grow, have increased disease resistance and is good to eat. There are nine traits of potato: CHIP, MATURITY, GLUCOSE, RED\_SKIN, EMERGENCE, FLOWER\_COLOR, FLOWER\_TIME, SCAB\_LS and Sprout\_LS. The strategy to improve traits is to use breeding, which is done by cross-pollinating with plants that have good traits. The progeny of the plants have different combinations of traits from the parents. We are interested in finding out what genes control these traits using patterns of inheritance in the progeny. When we find the genes, we can use the DNA markers to assist in breeding. Also, finding genes can also enable use of genome editing methods to modify genes to study them and improve traits. The performance of biological traits is controlled by genes, and a trait is not controlled by a single gene, in general, one trait is the product of many genes. There are 12 chromosomes in each cell, and also thousands of genes on each chromosome. It is very meaningful to group genes and to find the control of each cluster on traits. Firstly, the genes were encoded again according to their LOD scores. Secondly, chi-square test and normal test based in cross table were used to find not only the relationship between genes but also the distance between each pairs of genes, and after comparison, the normal distribution method was finally selected. Finally, DBSCAN based on density was used to cluster genes, and non-negative matrix factorization was used to cluster genes, which are related to traits and Cytoscape was used to show final result.

### 1.2 Data introduction (written by Yuexu, checked by Jiachen)

The locations in a chromosome that linked to traits are called quantitative trait loci (QTL), and LOD score was used to present a QTL in the interval is determined.

$$LOD = \log_{10} \left( \frac{\text{probability of QTL presence}}{\text{probability of QTL absence}} \right)$$

The LOD scores of 6290 genes and 9 traits will be used for preliminary analysis in the thesis. Here is a list of 9 traits.

<b>Tuber traits:</b> Tubers were harvested from the field stored at 7C, and 5 tubers from each clone and each block were bagged separately, finally, tubers from all 5 plants were bagged together.	
<b>Sprout_LS</b>	Tuber sprouting in storage was scored at three months after storage for the group of 5 tubers from a scale of 0-5, where 0 is no sprout, 1 is 0-0.5 cm, 2 is 0.5-1cm, 3 is 1-2cm, 4 is 2-3 cm and 5 is over 3cm.
<b>CHIP</b>	A chip slice was taken from each of 5 tuber and fried. The darkness of the chip was scored from 10 (dark) to 100 (light).
<b>GLUCOSE</b>	Glucose was measured using a strip assay. The average measurement for 5 tubers is used for each clone from each block. High glucose leads to darkened chips.
<b>RED_SKIN</b>	Tubers were scored for presence and absence of pigmentation
<b>Field traits</b>	
<b>EMERGENCE</b>	The average number of days after planting for plants to emerge from the ground to reach 10 cm, the average was taken over the 5 plant in each block
<b>FLOWER_COLOR</b>	Flowers were for presence and absence of pigmentation
<b>FLOWER_TIME</b>	The average number of days after planting for plants to flower, the average was taken over the 5 plant in each block. Some plants did not flower.
<b>MATURITY</b>	Visual scoring done at 100 days after planting on a scale of 1-9, where 1 is a dead plant and 9 is a healthy plant.
<b>Common scab disease rating</b>	
<b>SCAB_LS</b>	Common scab is a disease that causes the formation of scabby lesions on the surface of the potato tuber. The disease severity was visually rated on a numerical scale.

Table1. Introduction of 9 traits.

## 2. Methods

### 2.1 Preliminary preparations

#### 2.11 Data recode and Contingency Table (written by Yuexu, checked by Jiachen)

For every gene, the LOD score of 3 is a standard threshold indicative of a region with

a significant QTL, and 3 as a standard value could separate the LOD scores of every gene into two situations: significant and non-significant, so we recoded 1 if the LOD score is larger than 3 which means it's significant, otherwise recoded to 0 which means it's non-significant. This classification method can be regarded as a binomial distribution and it suggests to delete the columns of gene only include 0 which means this gene has no significant LOD scores, then the final result includes 6098 genes.

Contingency table is a table in which frequency counts in a contingency correspond to variables. After re-encoding the LOD to 0 and 1, each gene has only two variable values 0 or 1, on this basis, a  $2 * 2$  contingency table can be built for a pairs of genes.

## 2.12 Using Chi-square Test of Independence and Normal distribution to explore the relationship or distance between genes and between genes and traits

(written by Yuexu, checked by Jiachen & Jiajun)

The chi-square test of independence is used to test for a relationship between two categorical variables by using contingency table. We could use chi-square statistics and do a transformation to find the distance. The function of chi-square test of independence is:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

O: observed frequency in a cell of contingency table

E: expected frequency in a cell of table, it could with the assumption that the variables in the contingency table are independent.

$$E = \frac{(total_{row}) \times (total_{column})}{total_{grand}}$$

K: degree freedom, r is the number of rows and c is the number of columns.

$$K = (r - 1) * (c - 1)$$

One thing to note when performing the Chi-Square Independence Test is:

- 1) When  $N \geq 40$  and all theoretical frequencies  $T \geq 5$ , use Pearson's chi squared test. If the calculated p value is close to the specified significance level (such as 0.05), use Fisher's exact test.
- 2) When  $N \geq 40$  but there is a theoretical frequency  $T < 5$  of a certain grid, use Yates's correction for continuity, or use Fisher's exact test.
- 3) When  $N < 40$ , or the theoretical frequency  $T < 5$  of a certain grid, use Fisher's exact

test.

Actually we only judge two genes as highly associated if they are both significantly related to the same SNPs. So, in addition to the chi-square test of independence, we could use the p-value under the normal distribution to explore the correlation between genes, and the principle is as follows:

- 1) For any two genes A and B, count how many (1,1) outcome in Contingency table, and denoted as X. Suppose the data have N rows (total no. of SNPs in the data matrix). Then we have other outcomes total of N-X.
- 2) If A gene's proportion of 1's is  $p_1$ , B gene's proportion of 1's is  $p_2$ . Then if A and B are independent, then X follows Binomial ( $N, p_1 * p_2$ ).
- 3) Using CLT, X approximately follow Normal( $N * p_1 * p_2, N * p_1 * p_2 * (1 - p_1 * p_2)$ ).
- 4) Calculating the right tail p-value using the above distribution could let us pick up the most significant pairs of genes which have unusually larger overlap of 1's on the same SNPs.
- 5) Use the above p-value to do a transformation to find distance and perform cluster analysis.

## **2.2 Cluster analysis between genes by DBSCAN** (written by Yuexu, checked by Yuexu)

DBSCAN is a clustering method based on density is different. It only assume that clusters are continuous "dense" areas in the data space, separated by low density areas (Kriegel et al. 2011). This method based on the density could find the high-density regions with arbitrary shape and it also could identify noise points in regions with low density (Hahsler, Piekenbrock and Doran, 2019).

Hahsler et al. (2019) positively identified that DBSCAN clustering starts with a dataset D containing a set of points which belongs to D. Density-based algorithms need to obtain a density estimate over the data space, and DBSCAN estimates the density around a point using the concept of  $\epsilon$  neighborhood. It will choose a point randomly and circle with  $\epsilon$  as radius, and include some neighboring points in the circle, the minimal number of points which will be included in the circle is called minpts.

For example, if DBSCAN starts at point x randomly and there are more than minpts neighbors with distance smaller or equal to  $\epsilon$  ( $\epsilon$ ), then x is a core point and the neighbors will be included in the circle which means they have a high-density and clustered with x as a group, and it will start again by choosing a point which inside the circle as a new start point and do the same steps by assigning all points in its neighborhood to the cluster, finally if there is no more core points are found in the expanded neighborhood, then the cluster is complete and the remaining points are

choose to see if another core point can be found to start a new cluster. In figure1, point A and all the neighbors in the same circle is called direct density reachable, and the red points which are neighbors of A again as new core points, then their neighborhoods are called density reachable. Moreover, in the set, if a point is not a core point like B and C, it will be regarded as a border point, and all points in the same set are called density connected, however, if a point is not density connected it will be regard as a noise point like N. The implementation of this example of DBSCAN used here developed by Schubert et al. (2017).

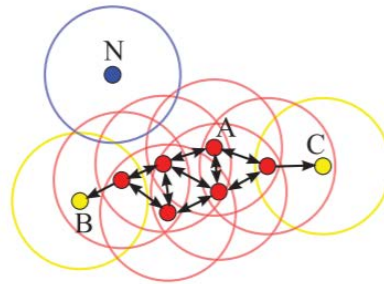


Figure 1: Illustration of the DBSCAN model.

Retrieved from: <https://dl.acm.org/doi/pdf/10.1145/3068335>

## 2.3 Cluster analysis of genes which are related to traits by NMF

(written by Jiachen, checked by Jiachen)

NMF can be an efficient method to identify distinct molecular patterns and a powerful tool for class discovery (Chunhong et al. 2009). The basic idea of Non-negative matrix factorization (NMF) can be simply described as: for any given non-negative matrix  $V$ , the NMF algorithm can find a non-negative matrix  $W$  and a non-negative matrix  $H$ . Thereby decomposing a non-negative matrix into the product of the left and right non-negative matrices, that is,  $V=W*H$ . It can be understood that the column vector of the original matrix  $V$  is the weighted sum of all the column vectors in  $W$ , and the corresponding weight coefficient is the element of the column vector of  $H$ , so  $W$  is called basis matrix,  $H$  is called coefficient matrix.

NMF is applied to cluster analysis, and the purpose of data reduction is to estimate the essential structure of the original data, that is, to obtain the basis matrix  $W$ . In fact, the basis matrix  $W$  can also be understood as the "pattern" in the pattern recognition theory, and the basis vector space forms the "class" in the K-means clustering. Therefore, NMF is to dig out the essential structure of the original features, find the "classes" in it, gather similar local features into one category, and then use the extracted "classes" to represent the original features.

Before using the NMF method for cluster analysis, it is extremely important to set the number of clusters to be classified, we called this as rank. For any rank  $k$ , the NMF algorithm groups the samples into clusters. The key issue is to tell whether a given

rank  $k$  decomposes the samples into “meaningful” clusters. The criteria for selecting the rank value is not mandatory nor unique. It can be judged whether a rank value is reasonable in many ways according to the needs of the data and the results of the clustering. Generally speaking, the selection criteria for rank are as follows:

- Select the minimum Rank value at which the cophenetic correlation coefficient begins to decrease.
- The point before the maximum change of the cophenetic value with the rank change.( Brunet et al. , 2004 )
- The first value where the RSS curve presents an inflection point.

This paper needs to perform nine clusters analysis on the genes related to the nine traits. Most of them choose the second criteria above as the standard, and refer to the specific clustering situation to determine the final rank value of each cluster analysis.

One question about the NMF decomposition is its non-uniqueness. Thus usually there is a normalization constraint to make the solution unique. In this regard, the method used in this article is adding a diagonal matrix  $D$  on the basis of  $V=W*H$ , then

$$V_{m*n} = W_{m*k} * D_{k*k} * D^{-1} * H_{k*n} \quad (1)$$

$$V_{m*n} = NW_{m*k} * NH_{k*n} \quad (2)$$

Where  $NW_{m*k} = W_{m*k} * D_{k*k}$ ,  $NH_{k*n} = D^{-1} * H_{k*n}$ , and specific steps are as follows:

1. Calculate the L1 norm of the  $W$  matrix, that is, calculate the sum of each column of the  $W$  matrix as  $W1norm, W2norm, \dots, Wknorm$ .
2. Calculate the mean of L1 norm of  $V$ , as  $MV = \text{mean}(V^T * 1)$
3. Make a  $k*k$  diagonal matrix  $D$ . The values on the diagonal of the  $D$  matrix are  $MV/W1norm, MV/W2norm, \dots, MV/Wknorm$ .
4. We will get a new  $W$  matrix,  $NW = W * D$
5. We will get a new  $H$  matrix,  $NH = D^{-1} * H$

Because NMF is a soft clustering method, its determination of gene category is not all-or-nothing. When Rstudio performs cluster analysis, the  $W$  matrix and  $H$  matrix obtained each time are not the same, so we need to increase the number of calculations to ensure accuracy. It should be noted that in order to save memory, only the optimal result will be stored in the memory during calculation, and the default output will also be the best result.

### 3. Analysis

### 3.1 Analysis of Preliminary preparations

#### 3.11 Chi-square Test of Independence and Yate's Correction for Continuity

(written by Yuexu, checked by Jiachen)

First of all, chi-square test of independence is used to find the relationship for pairs of genes, and a large statistic of chi-square represents a strong relationship between the pairs of genes. Meanwhile, yate's correction for continuity is used to correct the error caused by the approximate continuous distribution of discrete probabilities of frequencies in the contingency table be approximately continuous distribution, however, it's necessary to ensure that in this case, yate's correction is not overused. Thus, we used yate's correction for the pairs of genes who has the at least one cell with expected value smaller than 5, and plotted the chi-squares before and after correction.

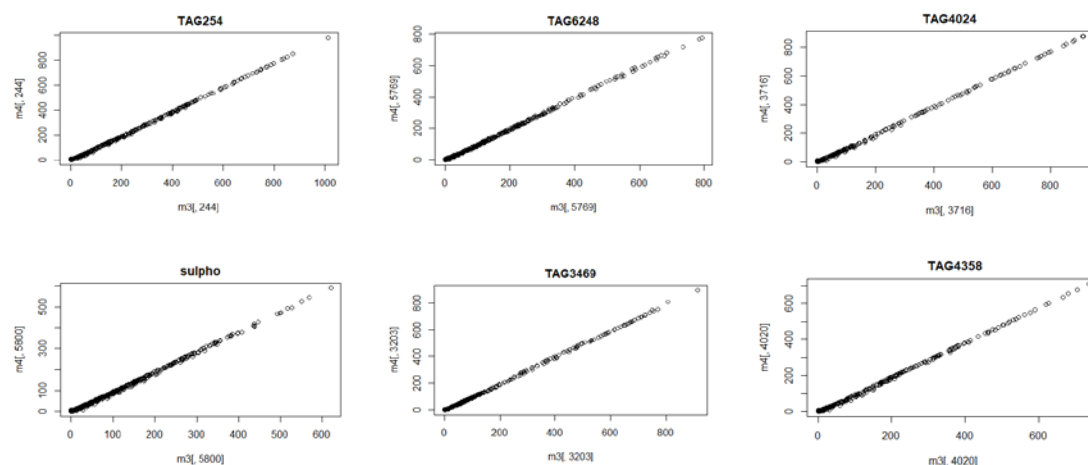


Figure 2: There is no obvious difference between before and after yate's correction.

Figure 2 shows chi-square before and after yate 'correction for combination of 6 genes and other remaining genes whose cross table has at least one cell with expected value smaller than 5. The abscissa shows value of before the correction, and the ordinate shows value of after the correction. These plots show that there are many overlapping points in the lower right corner, so most genes combinations still have the same chi-square value. In contrast, only a small number of chi-square values have obvious changes, so we chose the chi-square values before correction for subsequent calculation.

#### 3.12 Compare Chi-square Test of Independence and Normal Distribution

(written by Yuexu, checked by Jiachen)

Though both chi-square test of independence and normal test are based on

contingency table, there are some differences for them. The Pearson chi-square test uses all four factors in the contingency table, and normal distribution only use the value of [1,1] cell. As we mentioned before, if two genes both show 1 in the same row, it can be considered that both of them are significantly related to the same SNPS. For Pearson chi-square test, every value in contingency table will participate the calculate progress, so all cells in the table can influence the final result which is not as targeted and accurate as the method of normal that only use the common 1s to calculate.

### 3.2 Analysis of relationship and distance between genes and between genes and traits

#### 3.21 Distance selection of genes (written by Yuexu, checked by Yuexu)

P-value of normal test will be used for calculation of distance, small p-value means two genes are highly connected and there should be a small distance between them, so distance is supposed to proportional to the value of p-value. If the distance is very small, a large number of genes will be clustered together, and the distance is not clearly divided, the number of clusters will be very small and the clustering results will not be accurate, meanwhile, if the distance is very large and the genes are too scattered, then there will be many clusters, so the clustering results will also be not accurate enough. After summary, the minimum normal p-value was  $5.747885 \cdot 10^{(-305)}$ , the maximum normal p-value was 0.9999982. The lower bound of normal p-value was too small which leded a wide range covered by the upper and lower limit, so “log10” was used to reduce the scope of the distance. Meanwhile, we also need to ensure that p-value is proportional to distance, so the type of distance

$Dist = \frac{1}{-\log_{10}(p-value)}$  will be used.

$Dist = \frac{1}{-\log_{10}(p-value)}$					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	9.3	16.3	67.0	36.9	2603278.5

Table 2: min, mean, max value of norm p-value and distance.

Table 2 shows that after the transformation, the range of the distance is reduced, but the maximum value is still as high as 2603278.5. It is possible that only a very small part of the distances has a large value, so it’s better to find the distance range of most points by plotting 99% of whole distances.



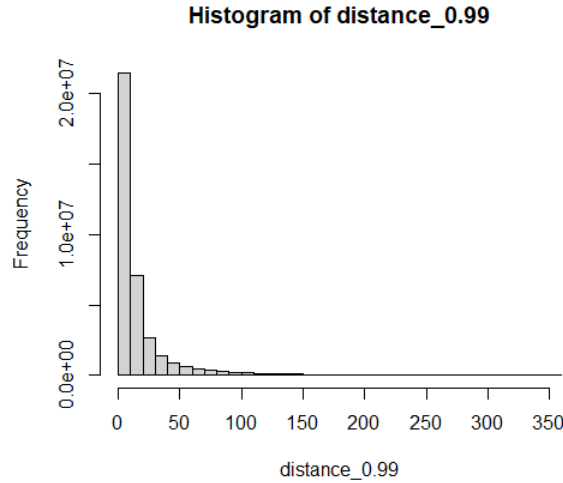


Figure 3: Histograms for 99% of all distances. The distance produced by  $1/(-\log_{10}(p\text{-value}))$  performs good, and 99% of distances between 0 to 150.

### 3.22 Selection of genes related to traits (written by Jiachen, checked by Jiachen)

When selecting genes related to traits, we also use two methods, chi-square independence test and normal distribution. By sorting the chi-square and p-value obtained by the chi-square independence test, and choosing a confidence level of 0.01 ( $p < 0.01$ ), we screened out the genes that are respectively related to the nine traits. The result obtained by the chi-square independence test is:

Traits	Number of genes related to the trait
CHIP	259
EMERGENCE	561
FLOWER_COLOR	433
FLOWER_TIME	1282
GLUCOSE	334
MATURITY	1515
RED_Skin	886
SACB_LS	629
Sprout_LS	418

Table 3: number of genes of chi-square independence test, the number of genes with strong association with trait MATURITY is the most which is 1515, and the second trait is FLOWER\_TIME with 1282 genes, and for other traits with 300 to 890 genes, however, trait CHIP has the smallest number of genes which is 259.

By sorting P-values obtained by the normal distribution at the level of 0.01 ( $p < 0.01$ ), the result of genes that are respectively related to the nine traits is:

Traits	Number of genes related to the trait
--------	--------------------------------------

CHIP	266
EMERGENCE	503
FLOWER_COLOR	415
FLOWER_TIME	1312
GLUCOSE	314
MATURITY	1297
RED_Skin	618
SACB_LS	653
Sprout_LS	536

Table 4: number of genes of normal, the number of genes with strong association with trait FLOWER\_TIME is the most which is 1312, and the second trait is MATURITY with 1297 genes, and for other traits with 300 to 618 genes, and same as the result of chi-square independence test, trait CHIP still has the smallest number of genes which is 266.

From table 3 and table 4, there is a slight change in the number of genes corresponding to each trait, however, on the whole, the results obtained by the two methods are largely the same. Those genes that are strongly related to traits are selected under both methods, and genes with weaker correlations are either added or deleted. Just like comparing the two methods in the previous article, the normal distribution performs more sharp and accurate than Pearson chi-square test, so in the next cluster analysis we will use the results obtained from the normal distribution.

### 3.3 Clustering analysis of all genes by DBSCAN (analyzed and written by Yuexu, checked by Yuexu)

There are two important parameters in DBSCAN which are eps and minpts. At present, there is no clear way to choose the most suitable eps and minpts. For minpts, Sander et al. (1998) noted that twice dimension of dataset could be used as minpts in general, however, Schubert et al. (2017) found that minpts has a smaller impact on the clustering result and there is “no significant differ” between different value of minpts.

For same eps with different minpts, the larger the minpts is, the more noise points will be, and the result with more noise points could be able to better concentrate on pairs of genes by comparison, but a smaller minpts could help to get less noise points and smaller number of clusters. After comparison, finally decided to use minpts=7.

For eps, Schubert et al. (2017, p.11) noted that “the value of  $\epsilon$  also depends on the distance function. In an ideal scenario, there exists domain knowledge to choose this parameter based on the application domain.” In gene network, the purpose is to find cluster of genes that highly correlated which means p-value of normal test for pairs of genes is small, generally speaking, when p-value is less than 0.01, there is a strong association between the two genes, so the distance calculated by p-value= 0.01 is selected as eps=0.5 which can ensure that almost all the associated points are included,

but this result is not very satisfactory, because DBSCAN shows that there is only one cluster for all genes, this means that the eps value is too large and needs to be further reduced. By using `KNNdisplot()` which can show k-nearest neighbor distances in a matrix of point in R, Figure 5 shows most of the distances are between 0 and 0.02, so the eps will be chosen from this range.

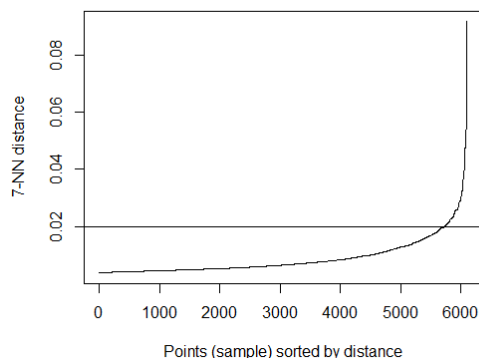


Figure 4: `KNNdisplot` shows that distance for most of genes are in the range from 0 to 0.02.

Four values of eps were selected for comparison. As Schubert et al. (2017) noted that if there is a cluster contains more than 20% to 50% points, the cluster method should be use again with a smaller eps. Based on this rule, some smaller values are selected from range 0 to 0.02 as eps, and table 5 shows the four DBSCAN results.

Cluster	Eps=0.006	Eps=0.007	Eps=0.008	Eps=0.01
Number	31	34	24	9
Noise points	2565	1867	1383	776
Top 5 clusters	cluster1:687 cluster2:433 cluster3:420 cluster4:385 cluster5:266	cluster1:925 cluster2:725 cluster3:513 cluster4:237 cluster5:194	cluster1:3673 cluster2:519 cluster3:188 cluster4:86 cluster5:30	cluster1:5239 cluster2:19 cluster3:16 cluster4:10 cluster5:9
Corresponding p-value	$10^{-\frac{502}{3}}$	$10^{-\frac{1002}{7}}$	$10^{-125}$	$10^{-100}$
20% of total number	1219.6			
50% of total number	3049			

Table 5: DBSCAN results for 4 different eps with `minpts=7`. We can see that the larger eps is, the smaller number of clusters be. The number of genes included in cluster1 with eps=0.008 and cluster1 with eps=0.01 exceeds 20% of the total number of genes, so these two cases do not perform well. Meanwhile, with the same parameter conditions, it is more advantageous to choose clusters with fewer noise points, so eps=0.007 will be used for gene network.

From table 5, when  $\text{eps}=0.008$ , cluster 1 has 3676 genes and when  $\text{eps}=0.01$ , cluster 1 has 5239, both of these two clusters contain more than 50% of whole genes, based on the rule from Schubert et al. (2017), these two cases are not perform well. In general, it's better to choose clusters with less noise points, so  $\text{eps}=0.007$  will be selected for plotting gene network.

### 3.4 Cluster analysis of genes related to traits by NMF (analyzed and written by Jiachen & Jiajun, checked by Jiachen & Jiajun)

#### 3.41 The final result of clustering

In order to facilitate the calculation, we used the `nmf` package in Rstudio for cluster analysis. There are 11 built-in NMF algorithms ("brunet", "KL", "lee", "Frobenius", "offset", "nsNMF", "ls-nmf", "pe-nmf", "siNMF", "snmf/r", "snmf/l") and 4 initialization methods ("ica", "nndsvd", "none", "random") in the NMF package. We chose the "brunet" algorithm and the "random" or "ica" method as the initialization method.

Since the rank value is not unique, we set the number of clusters in advance by referring to the cophenetic coefficient, so we need to use the consensus matrix to see whether the rank value we set can reasonably cluster genes. For example, the information of CHIP is as follow:

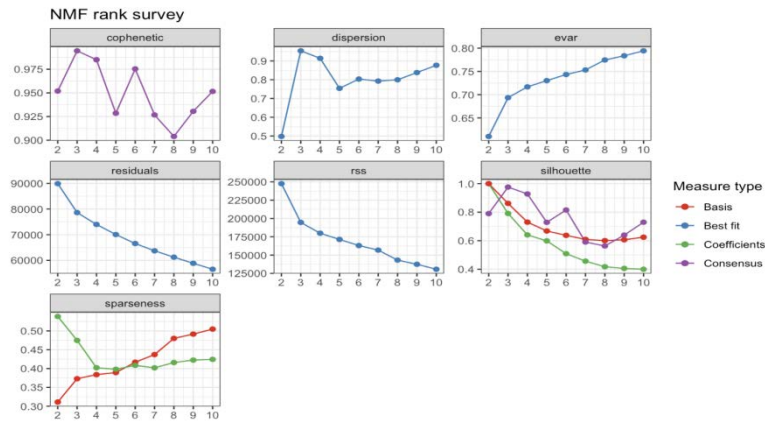


Figure 5: Various coefficients related to CHIP.

According to the point before the maximum change of the cophenetic value with the rank change, we choose rank=4, and the consensus matrices of rank=2:10 are as follows:

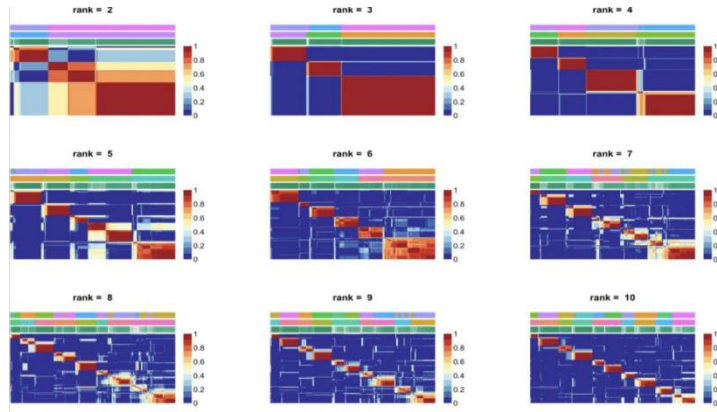


Figure 6: consensus matrices of CHIP.

Figure 6 shows the consensus matrices generated for ranks  $K=2:10$ . The clear red square on the diagonal indicates that the clustering result is acceptable when rank=3 and 4. As the rank value increases, the red squares on the diagonal are scattered, which means that the clustering results under this rank value are not ideal.

The clustering results of genes related to the nine traits are as follows:

Traits	Number of clusters being clustered (the value of rank)
CHIP	4
EMERGENCE	3
FLOWER_COLOR	3
FLOWER_TIME	8
GLUCOSE	3
MATURITY	7
RED_Skin	4
SACB_LS	4
Sprout_LS	4

Table 6: Number of clusters being clustered (the value of rank).

After clustering analysis of genes related to traits, 1312 genes related to FLOWER\_TIME were clustered into 8 clusters, and 1297 genes related to MATURITY were clustered into 7 clusters. The genes related to the other seven traits are all clustered into 3 clusters or 4 clusters.

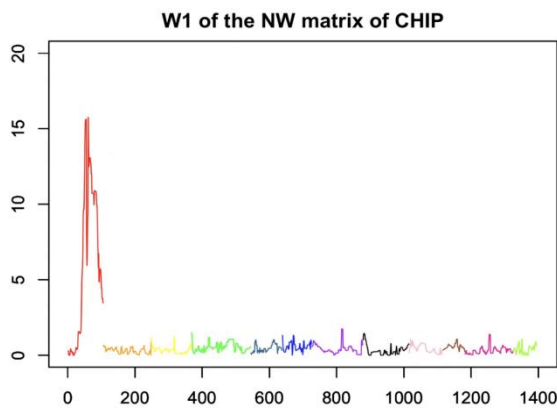
### 3.42 Details about the W and H matrices (analyzed and written by Jiachen & Jiajun, checked by Jiachen & Jiajun)

The core idea of NMF is to decompose the original matrix  $V$  into the product of two matrices  $W$  and  $H$ , making  $W \cdot H$  infinitely close to  $V$ . In order to standardize the  $W$  matrix, we have improved the NMF so that it becomes:

$$V_{m \times n} = N W_{m \times k} * N H_{k \times n}$$

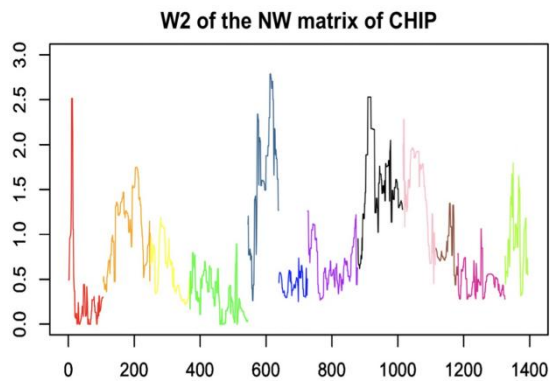
For W or NW matrix, each column of W(NW) matrix stands for a cluster. The highest entry in each row under each column denoted the features emphasized by the respective clusters (Krishan, 2016). For H or NH matrix, the row index of the highest entry under each column indicates the cluster membership of that column. The specific information of NW and NH of the nine traits is as follows:

### 1) CHIP



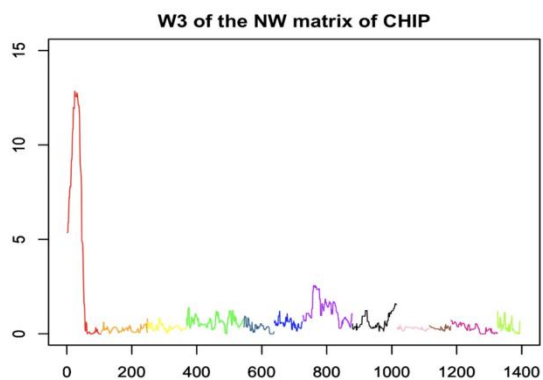
Notes:

- x-axis is 1394 SNPs
- y-axis is the coefficients in the NW matrix
- The 12 colors from left to right represent chromosome 1 to chromosome 12
- The coefficient of the w1 column is between 0 and 20
- The higher coefficient appears on chromosome 1



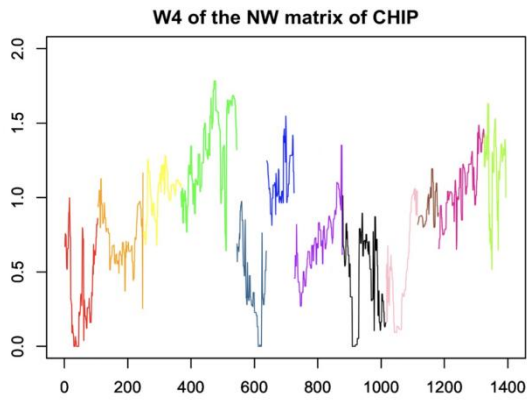
Notes:

- The coefficient of the w2 column is between 0 and 3



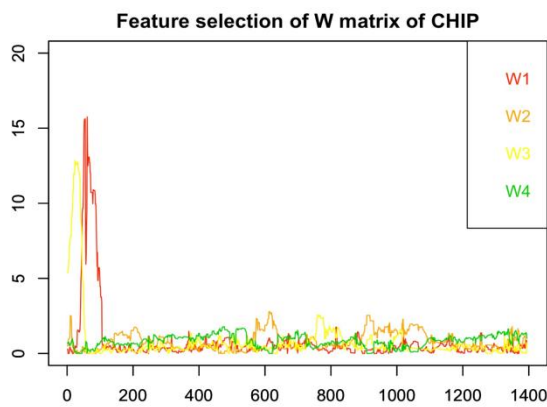
Notes:

- The coefficient of the w3 column is between 0 and 15
- The higher coefficient appears on chromosome 1



Notes:

- The coefficient of the w4 column is between 0 and 2



Notes:

- The four columns of the W matrix are placed in the same graph, and four colors are used to distinguish w1, w2, w3 and w4

Figure 7: details of NW matrix of CHIP.

The heatmap of NH matrix of CHIP is as follow:

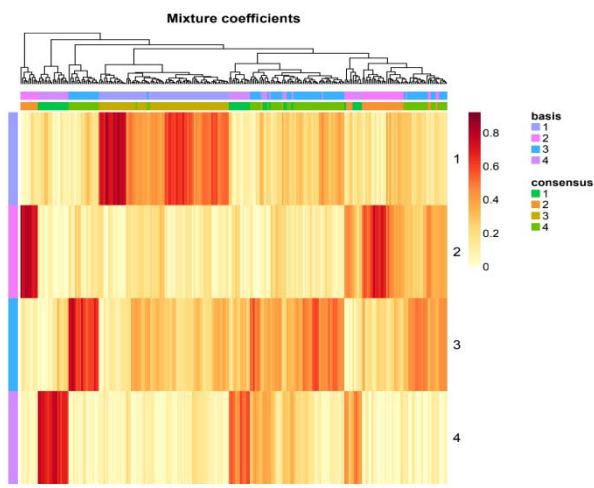
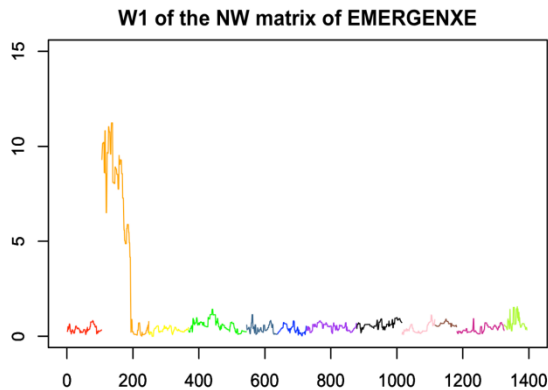


Figure 8: heatmap of NH matrix of CHIP

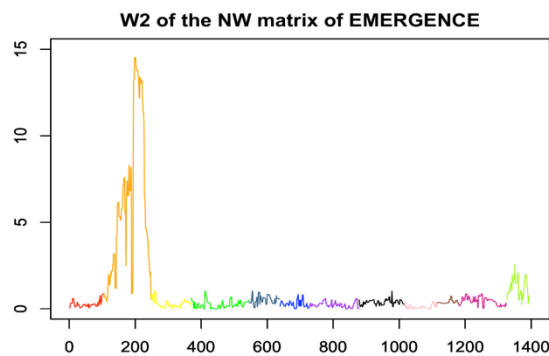
Through the heatmap, we can clearly see that the 266 genes related to CHIP are clustered into 4 clusters, and the number of genes classified as the first cluster is the largest.

## 2) EMERGENCE



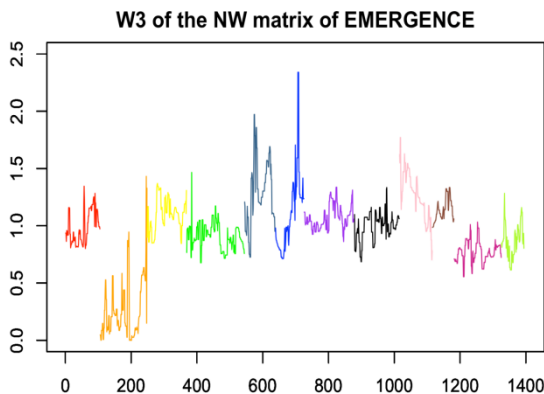
Notes:

- x-axis is 1394 SNPs
- y-axis is the coefficients in the NW matrix
- The 12 colors from left to right represent chromosome 1 to chromosome 12
- The coefficient of the w1 column is between 0 and 13
- The higher coefficient appears on chromosome 2



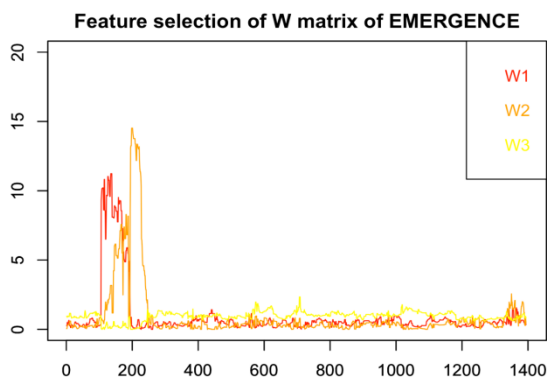
Notes:

- The coefficient of the w2 column is between 0 and 15
- The higher coefficient appears on chromosome 2



Notes:

- The coefficient of the w3 column is between 0 and 2.5
- The higher coefficient appears on chromosome 6



Notes:

- The three columns of the W matrix are placed in the same graph, and three colors are used to distinguish w1, w2, and w3.



Figure 9. details of NW matrix of EMERGENCE.

The heatmap of NH matrix of EMERGENCE is as follow:

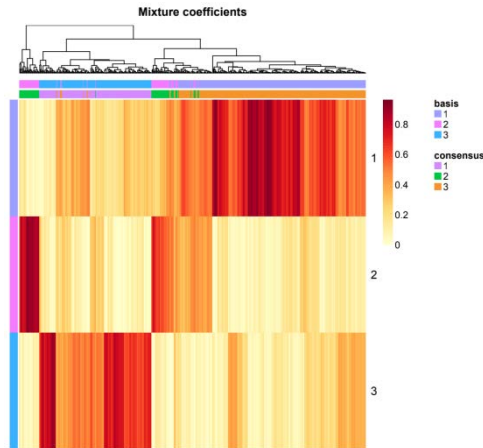
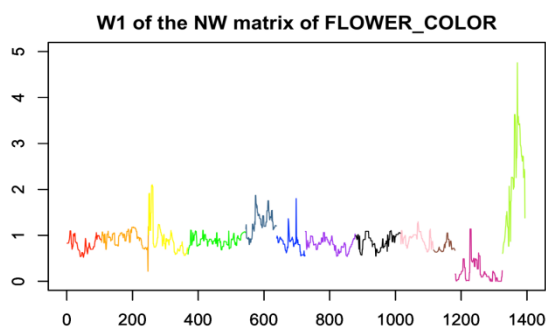


Figure 10: heatmap of NH matrix of EMERGENCE.

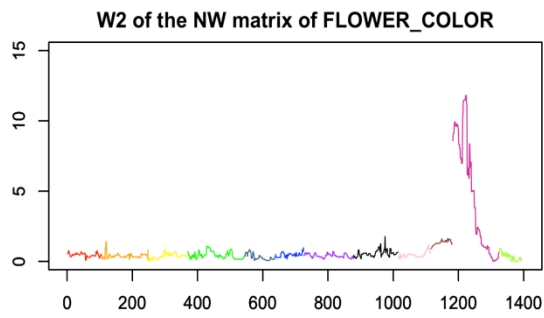
The 503 genes related to EMERGENCE are clustered into 3 clusters, and the number of genes classified as the first cluster is the largest.

### 3) FLOWER\_COLOR



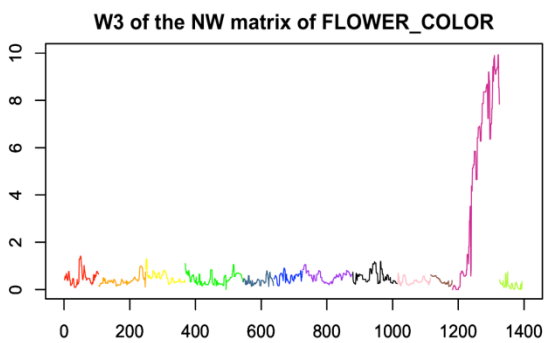
#### Notes:

- The coefficient of the w1 column is between 0 and 5
- The higher coefficient appears on chromosome 12



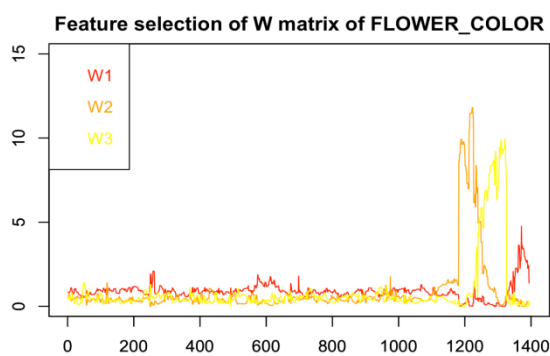
Notes:

- The coefficient of the w2 column is between 0 and 12
- The higher coefficient appears on chromosome 11



Notes:

- The coefficient of the w3 column is between 0 and 10
- The higher coefficient appears on chromosome 11



Notes:

- The three columns of the W matrix are placed in the same graph, and three colors are used to distinguish w1, w2, and w3.

Figure 11: Details of NW matrix of FLOWER\_COLOR.

The heatmap of FLOWER\_COLOR as follow:

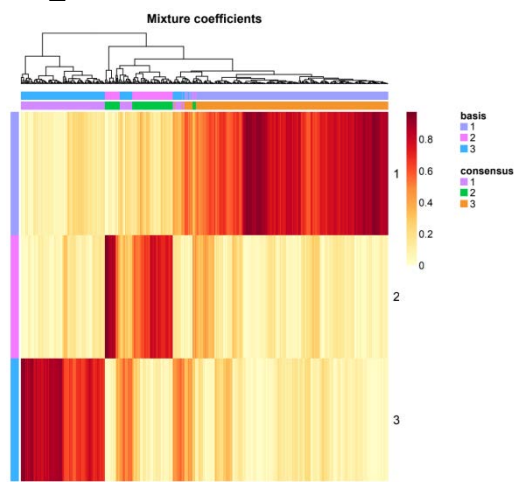
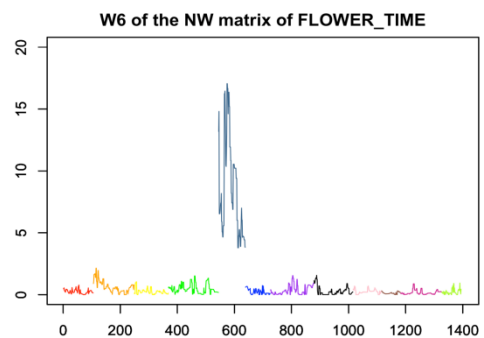
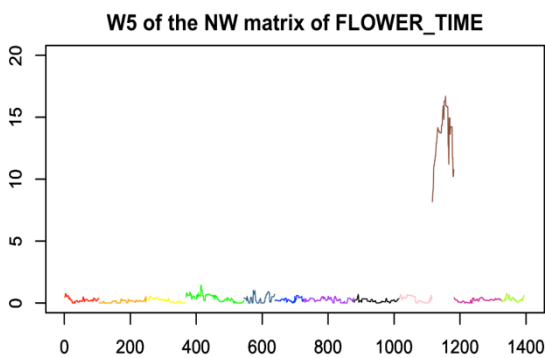
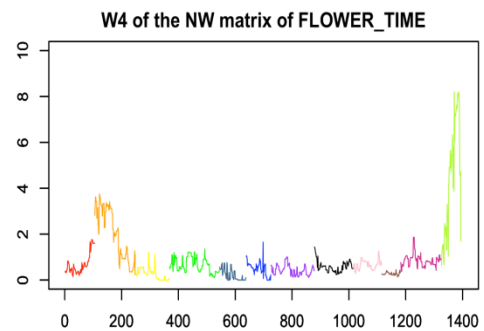
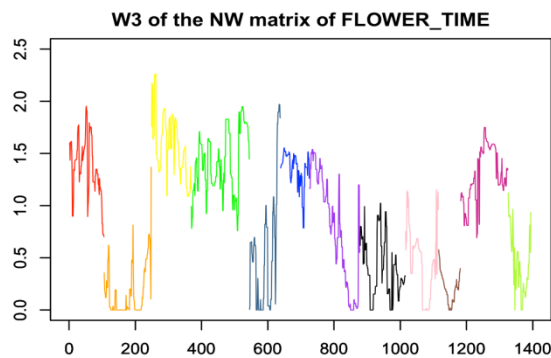
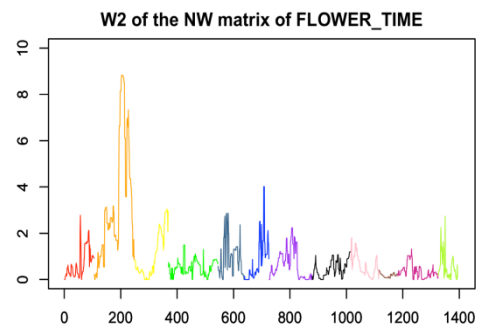
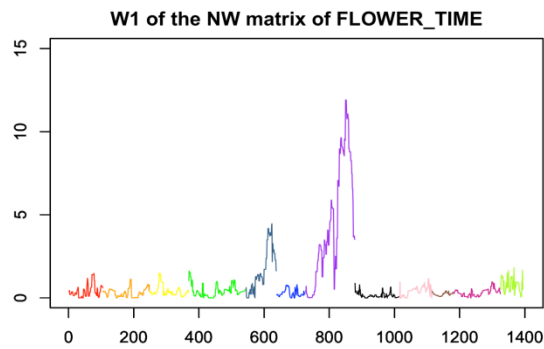
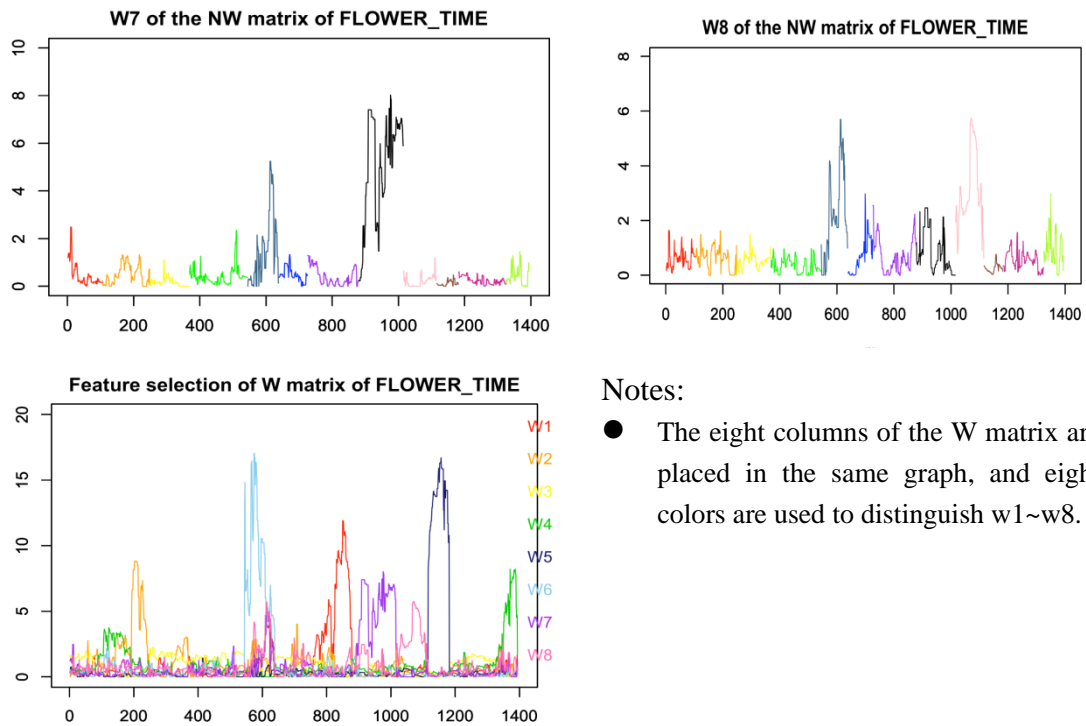


Figure 12: heatmap of NH matrix of FLOWER\_COLOR.

The 415 genes related to FLOWER\_COLOR are clustered into 3 clusters, and the number of genes classified as the first cluster is the largest.

#### 4) FLOWER\_TIME





Notes:

- The eight columns of the W matrix are placed in the same graph, and eight colors are used to distinguish w1~w8.

Figure 13: Details of NW matrix of FLOWER\_TIME.

The heatmap of NH matrix of FLOWER\_TIME as follow:

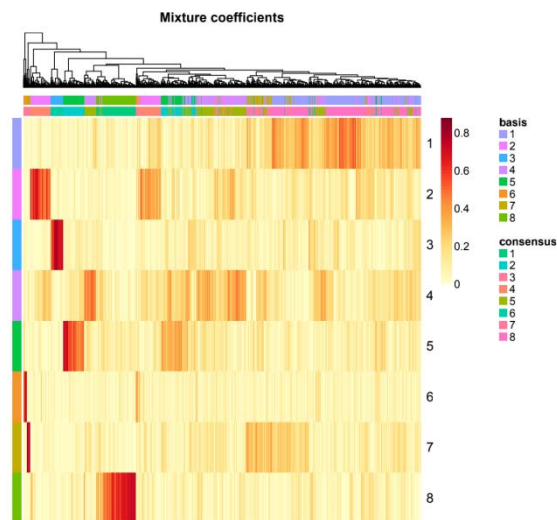
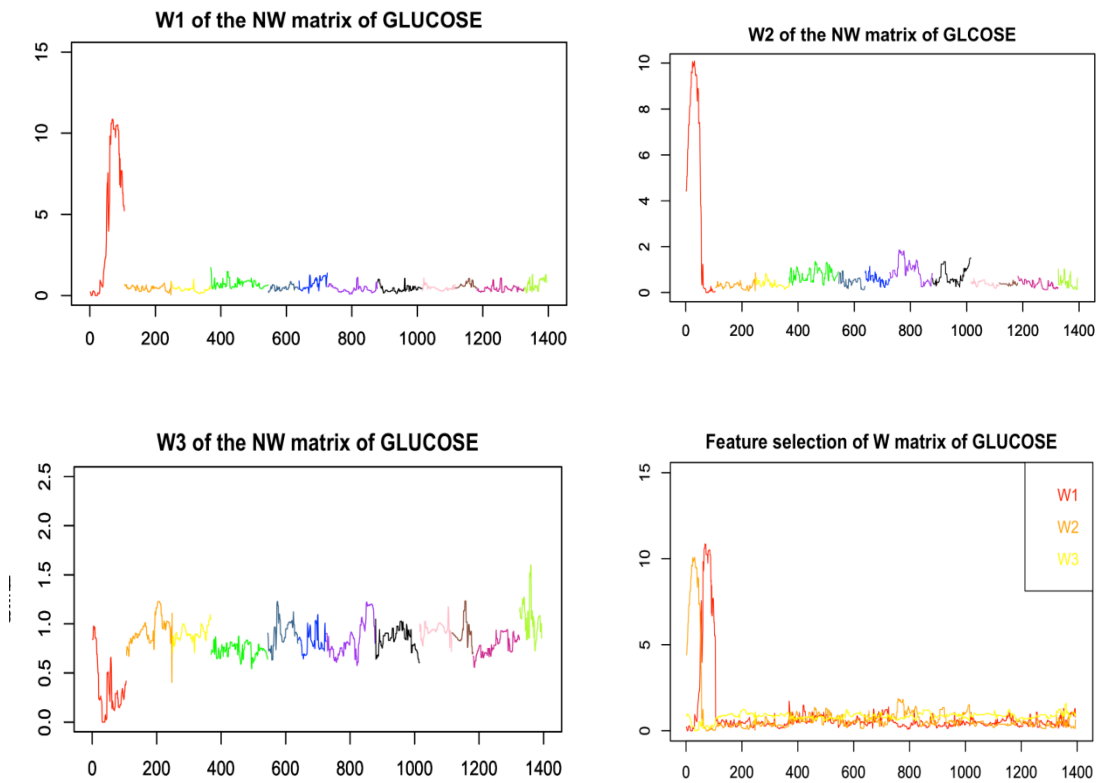


Figure 14: heatmap of NH matrix of FLOWER\_TIME.

The 1312 genes related to FLOWER\_TIME are clustered into 8 clusters, and the number of genes classified as the first cluster is the largest.

## 5) GLUCOSE

Figure 15: Details of NW matrix of GLUCOSE.



The heatmap of NH matrix of GLUCOSE as follow:

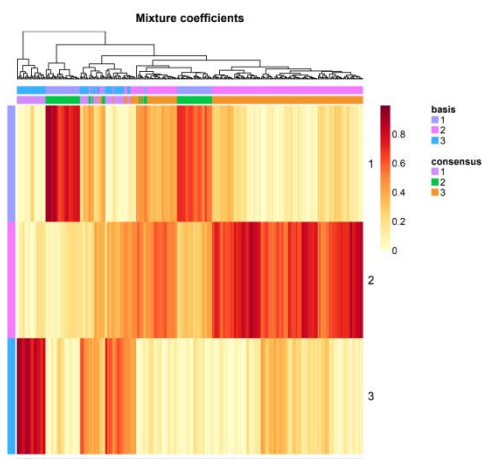
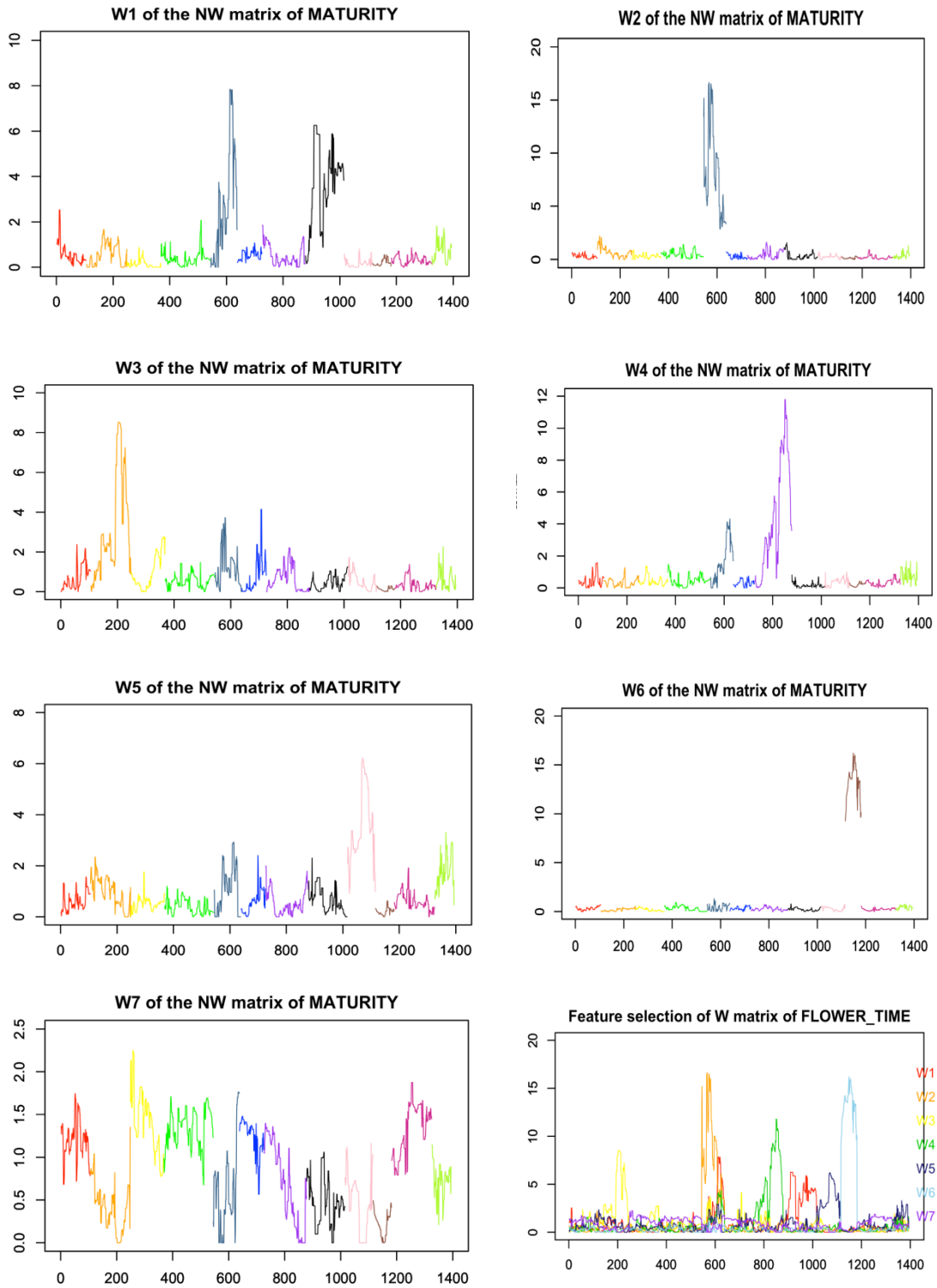


Figure 16: heatmap of NH matrix of GLUCOSE.

The 314 genes related to GLUCOSE are clustered into 3 clusters, and the number of genes classified as the second cluster is the largest.

## 6) MATURITY

Figure 17: Details of NW matrix of MATURITY.



The heatmap of NH matrix of MATURITY as follow:

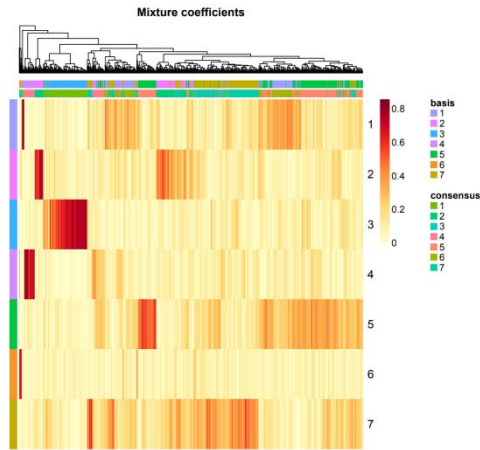
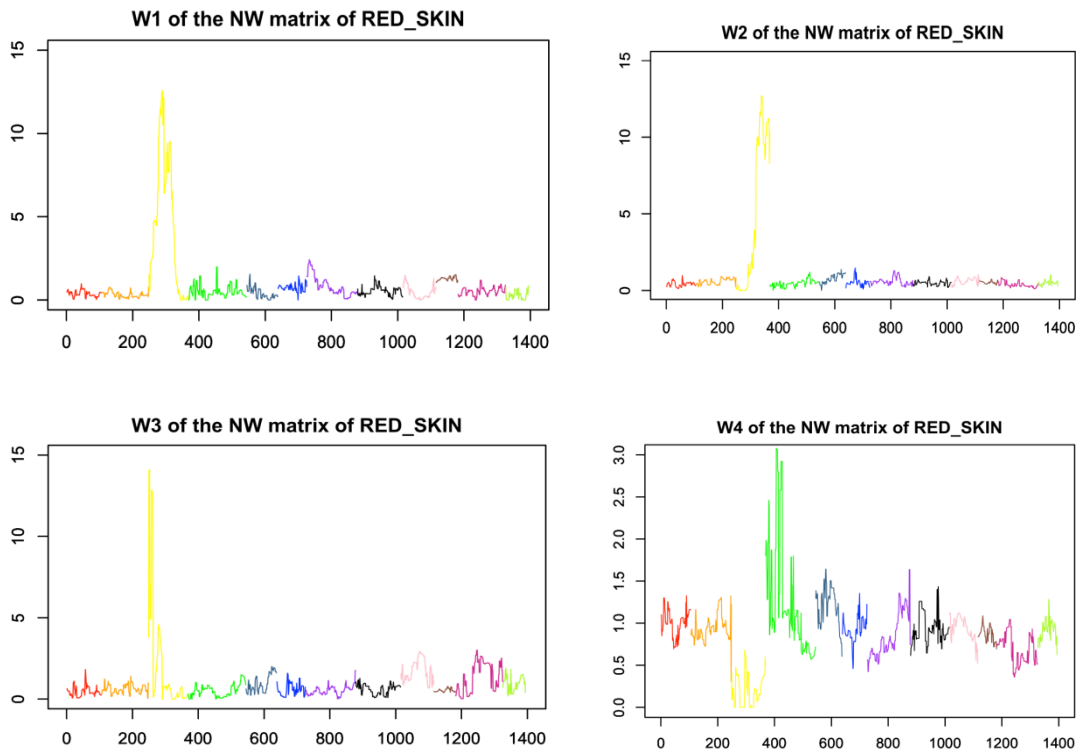


Figure 18: heatmap of NH matrix of MATURITY.

The 1297 genes related to MATURITY are clustered into 7 clusters.

7) RED\_SKIN



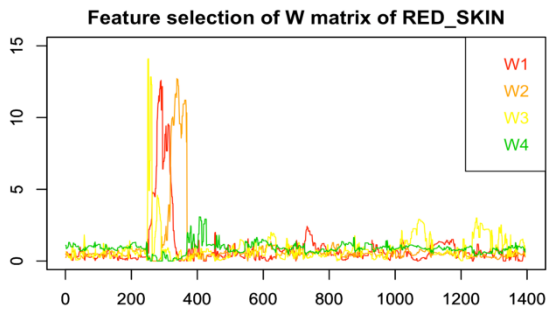


Figure 19: Details of NW matrix of RED\_SKIN.

The heatmap of NH matrix of RED\_SKIN as follow:

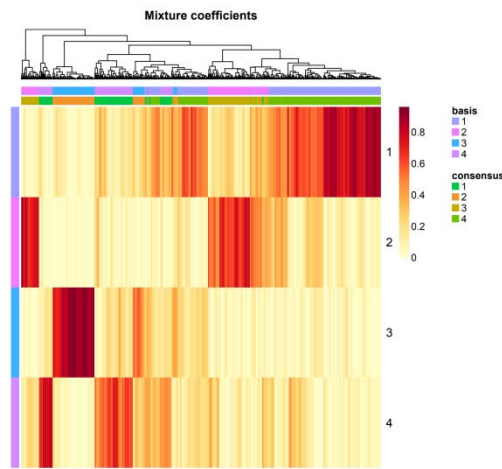
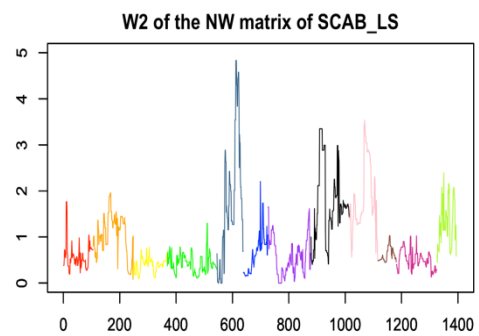
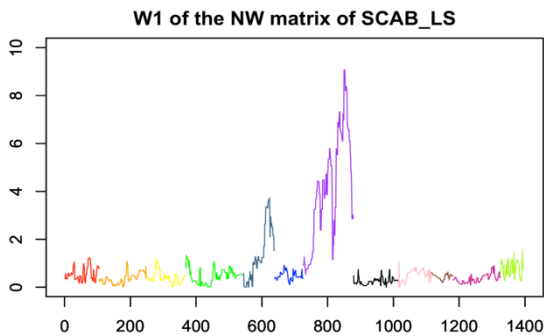


Figure 20: heatmap of NH matrix of RED\_SKIN.

The 618 genes related to RED\_SKIN are clustered into 4 clusters, and the number of genes classified as the first cluster is the largest.

### 8) SCAB\_LS





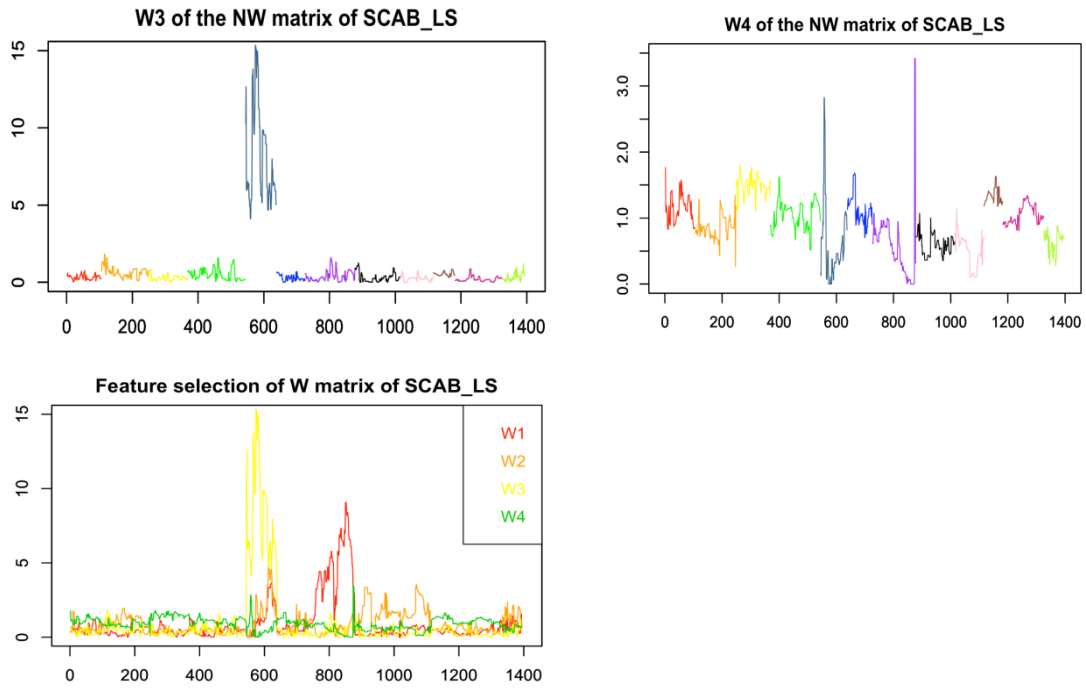


Figure 21: Details of NW matrix of SCAB\_LS.

The heatmap of NH matrix of SCAB\_LS as follow:

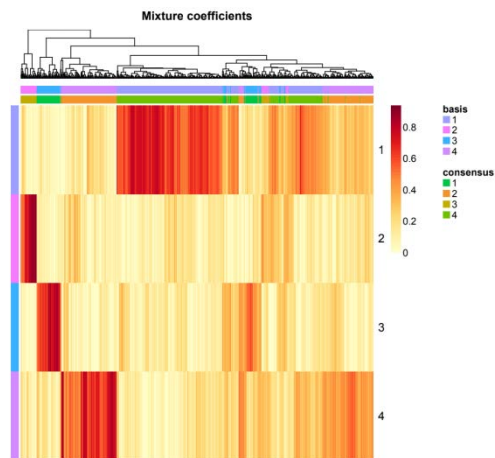


Figure 22: heatmap of NH matrix of SCAB\_LS.

The 653 genes related to SCAB\_LS are clustered into 4 clusters, and the number of genes classified as the first cluster is the largest.

## 9) SPROUT\_LS

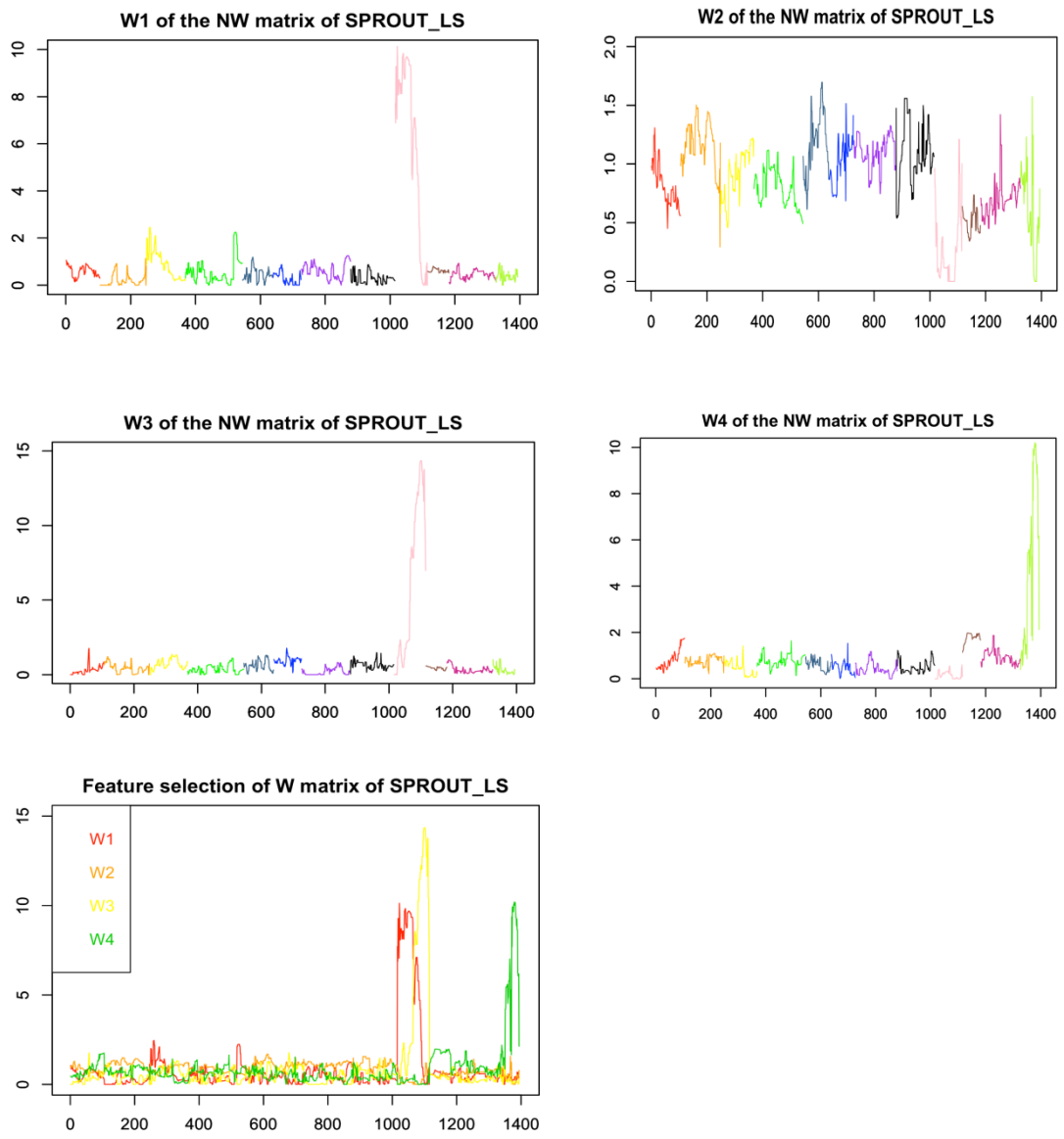


Figure 23: Details of NW matrix of SPROUT\_LS.

The heatmap of NH matrix of SPROUT\_LS as follow:

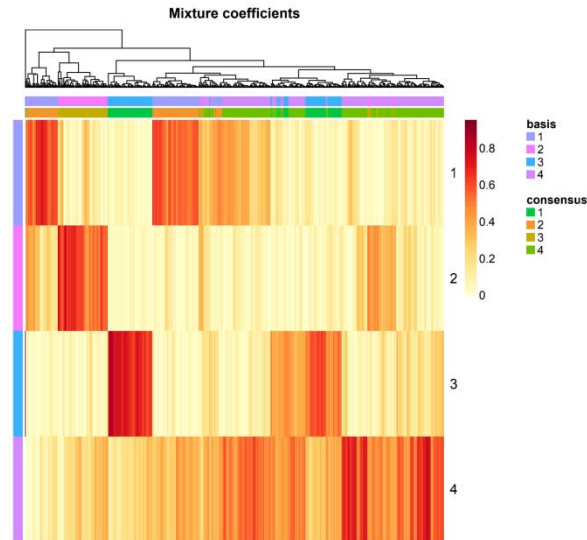


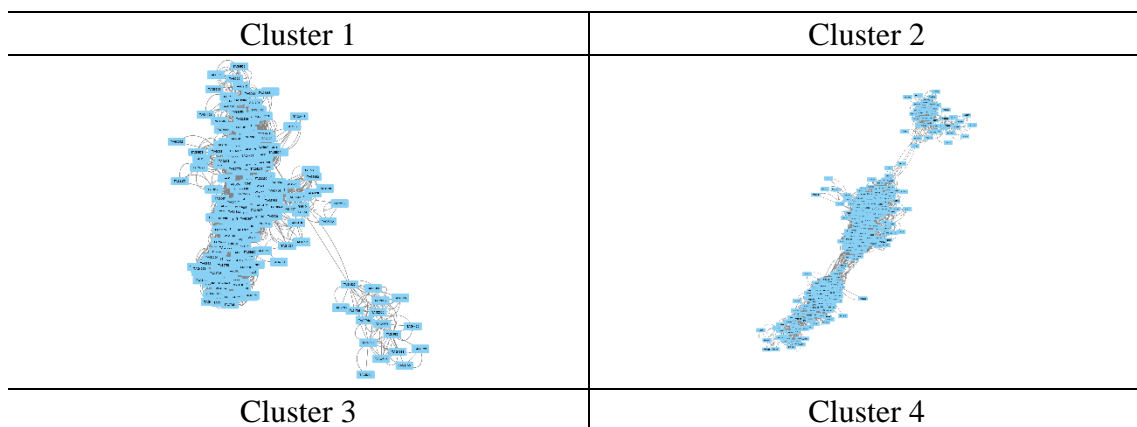
Figure 24: heatmap of NH matrix of SPROUT\_LS.

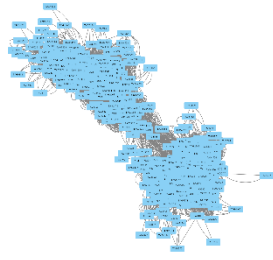
The 536 genes related to SPROUT\_LS are clustered into 4 clusters, and the number of genes classified as the fourth cluster is the largest.

### 3.5 Gene Network form Cytoscape

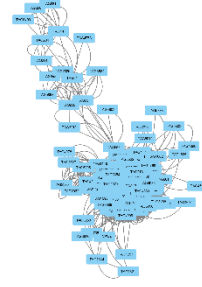
#### 3.51 Genes Network (written by Yuexu, checked by Yuexu)

Using  $\text{eps} = 0.007$  as an example. After defining the genes in each cluster, for each gene, found the other gene with the distance within the  $\text{eps}=0.007$  from it in the cluster, then took the names of each pair of gene as the starting point and the destination point, the distance between them used as the edge contribution. The Cytoscape gave the final result.

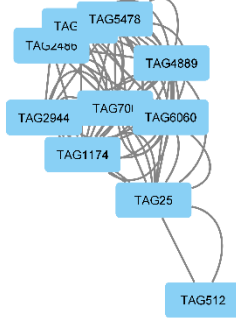




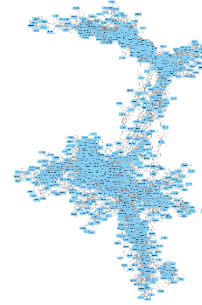
Cluster 5



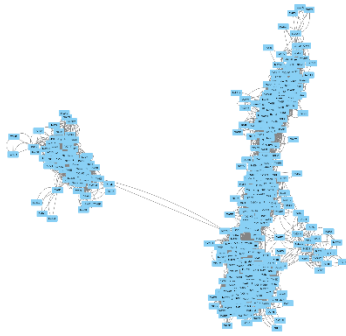
Cluster 6



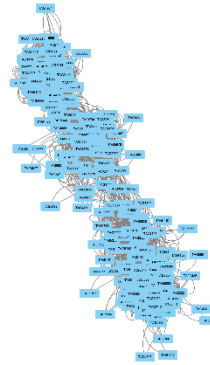
Cluster 7



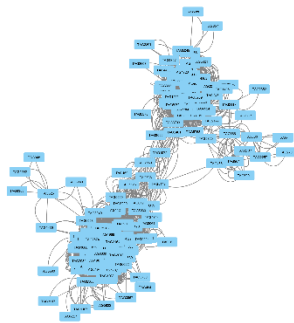
Cluster 8



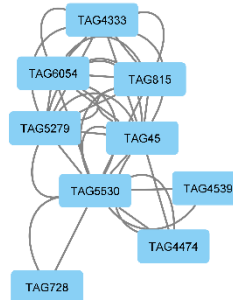
Cluster 9



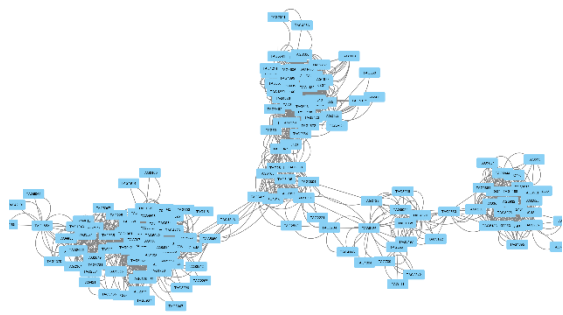
Cluster 10



Cluster 11



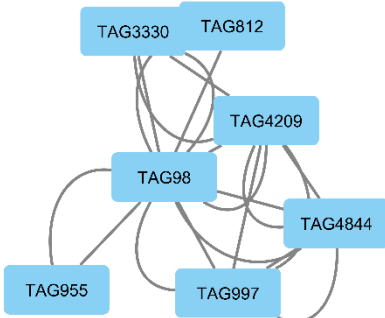
Cluster 12



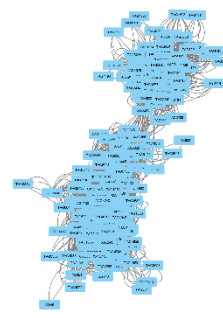
Cluster 13



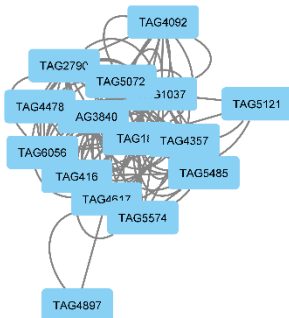
Cluster 14



Cluster 15



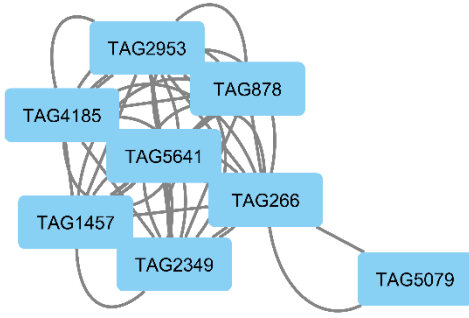
Cluster 16



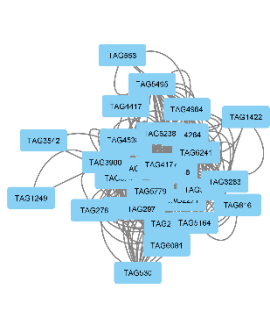
Cluster 17



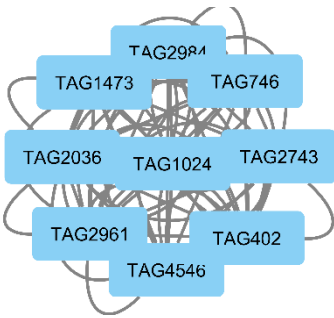
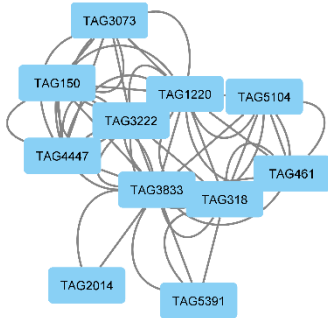
Cluster 18



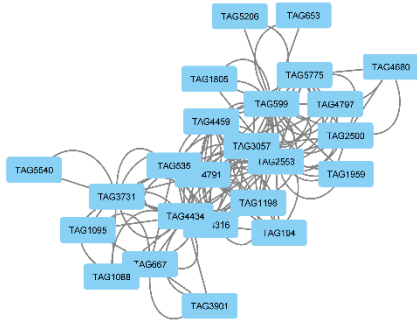
Cluster 19



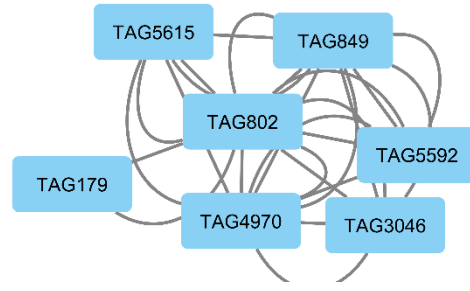
Cluster 20



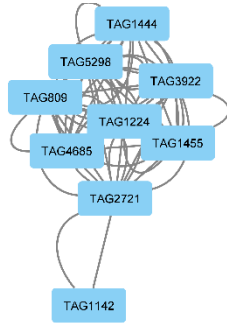
Cluster 21



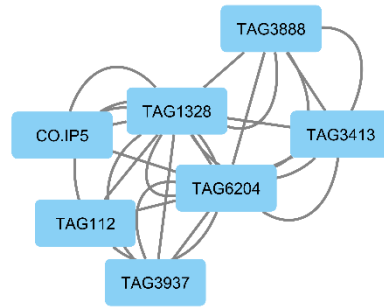
Cluster 22



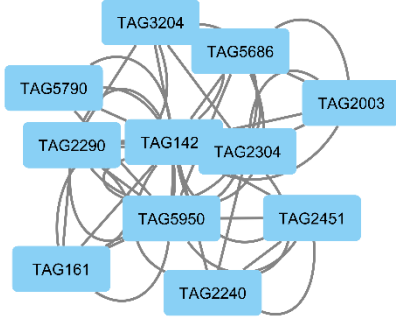
Cluster 23



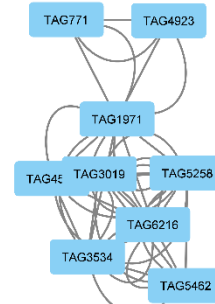
Cluster 24



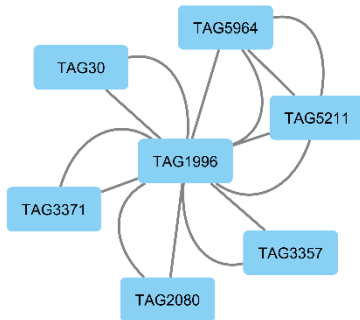
Cluster 25



Cluster 26



Cluster 27



Cluster 28



Cluster 29



Cluster 30



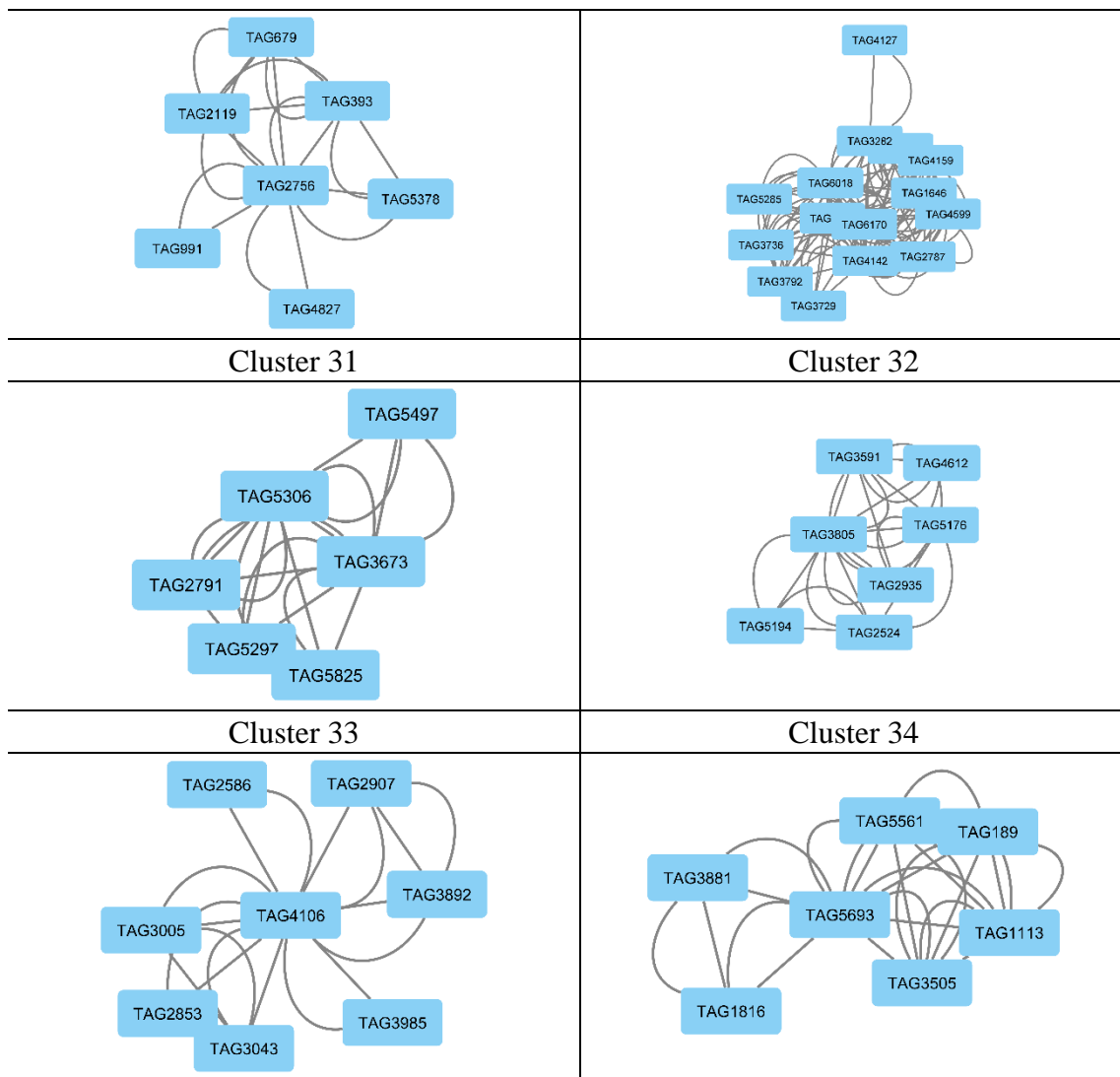


Figure 25: Gene networks with 34 clusters.

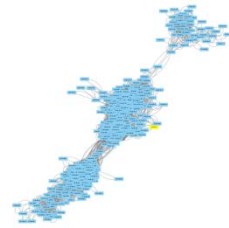
### 3.52 Marking the genes related to CHIP in genes network (written by Jiachen, checked by Jiajun)

The 266 genes related to CHIP were clustered into 4 clusters. We distinguished these four clusters of genes related to CHIP with red, orange, yellow, and green, and labeled these genes into the gene network which is clustered into 34 clusters. The results show that most of the genes related to CHIP are concentrated in cluster 9, and a small part of the genes are scattered in the other 15 clusters.

Cluster 1	Cluster 2
-----------	-----------



Cluster 3



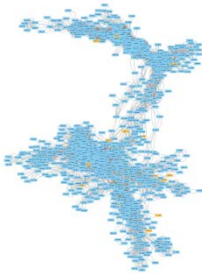
Cluster 4



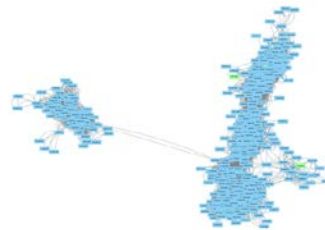
Cluster 6



Cluster 7



Cluster 8



Cluster 9



Cluster 11



Cluster 12



Cluster 14



Cluster 16



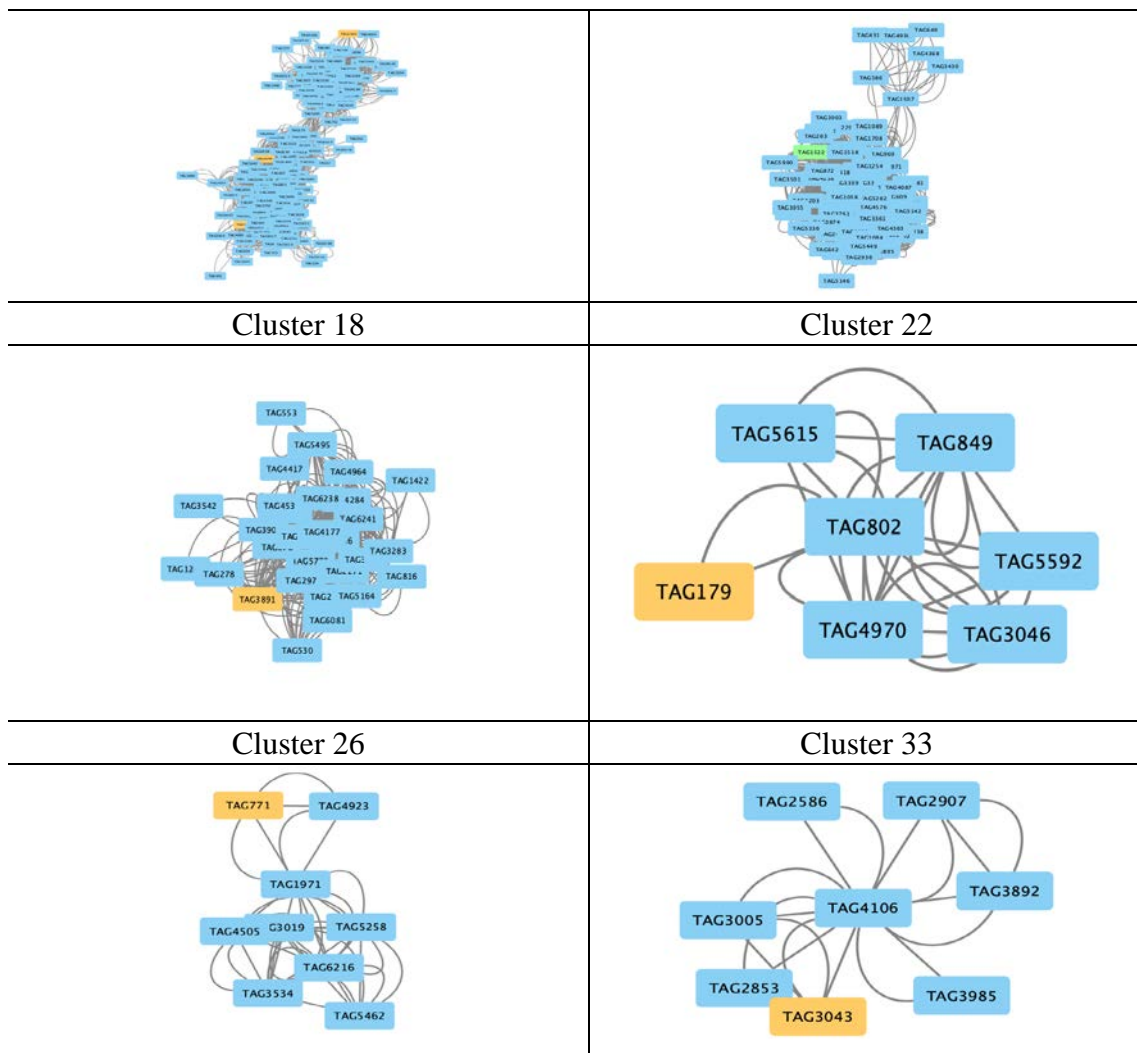
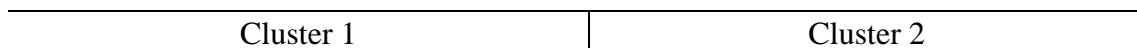
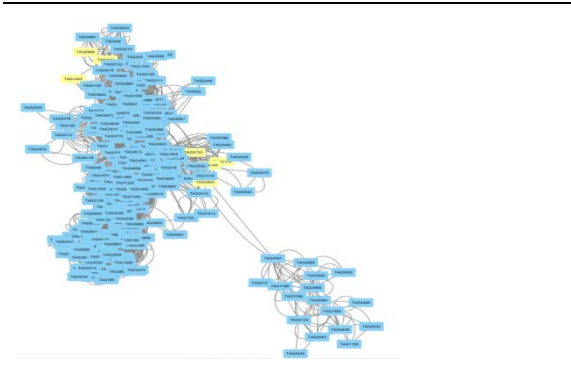


Figure 26: Distribution of genes related to CHIP.

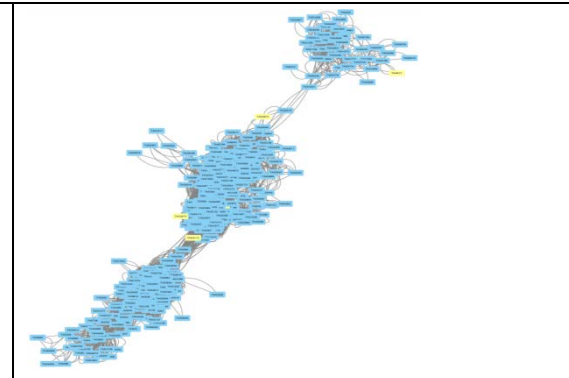
### 3.53 Marking the genes related to EMERGENCE in genes network (written by Jiachen, checked by Jiajun)

The 503 genes related to EMERGENCE were clustered into 3 clusters. We distinguished these three clusters of genes related to EMERGENCE with red, orange, and yellow. The results show that most of the genes related to EMERGENCE are concentrated in cluster 3, and a small part of the genes are scattered in the other 14 clusters.

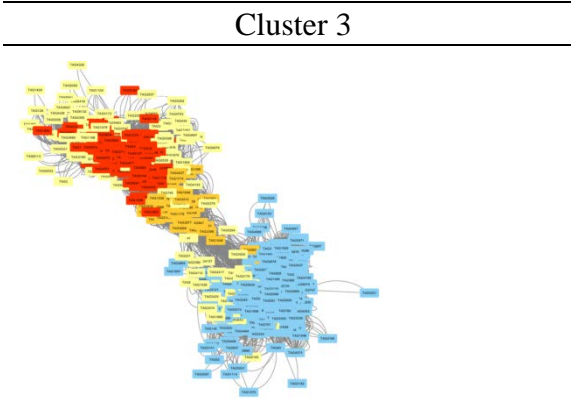




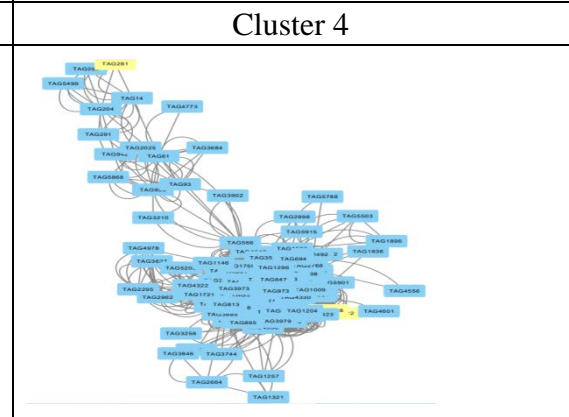
Cluster 3



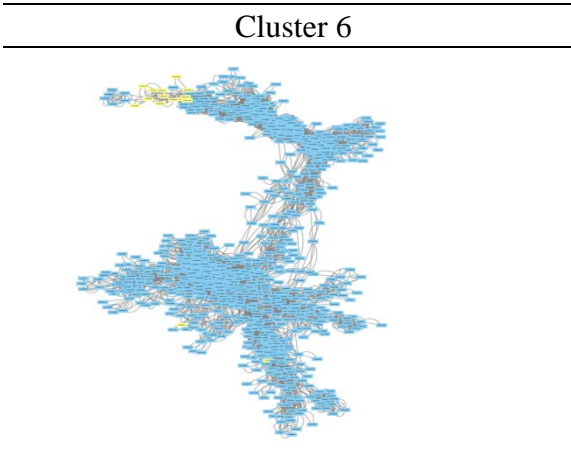
Cluster 4



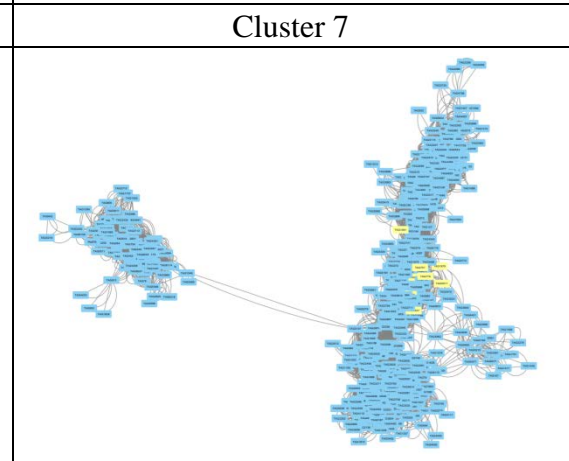
Cluster 6



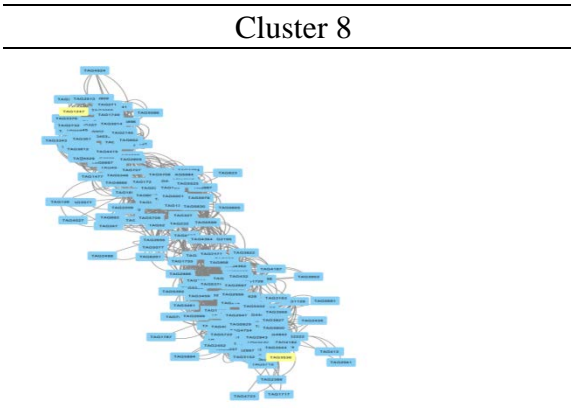
Cluster 7



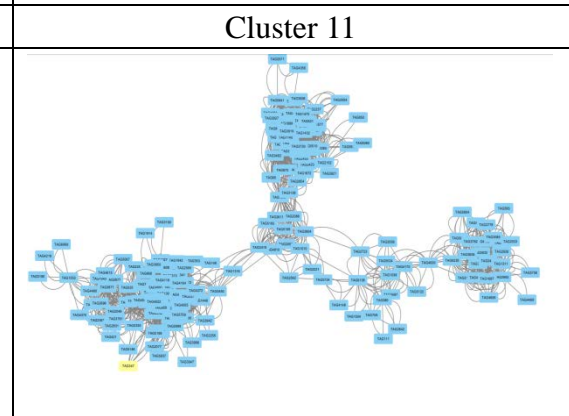
Cluster 8



Cluster 11



Cluster 13



Cluster 18

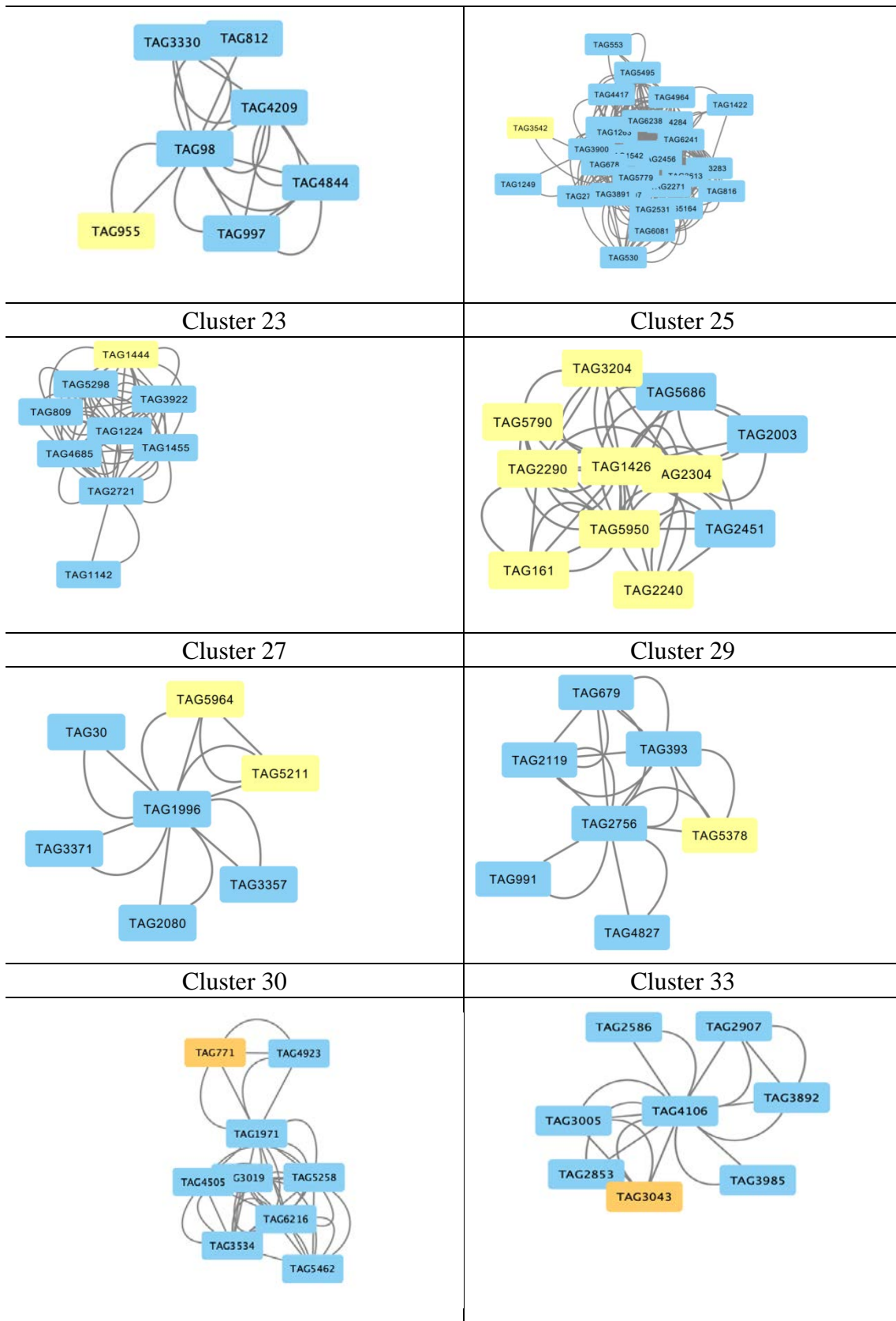
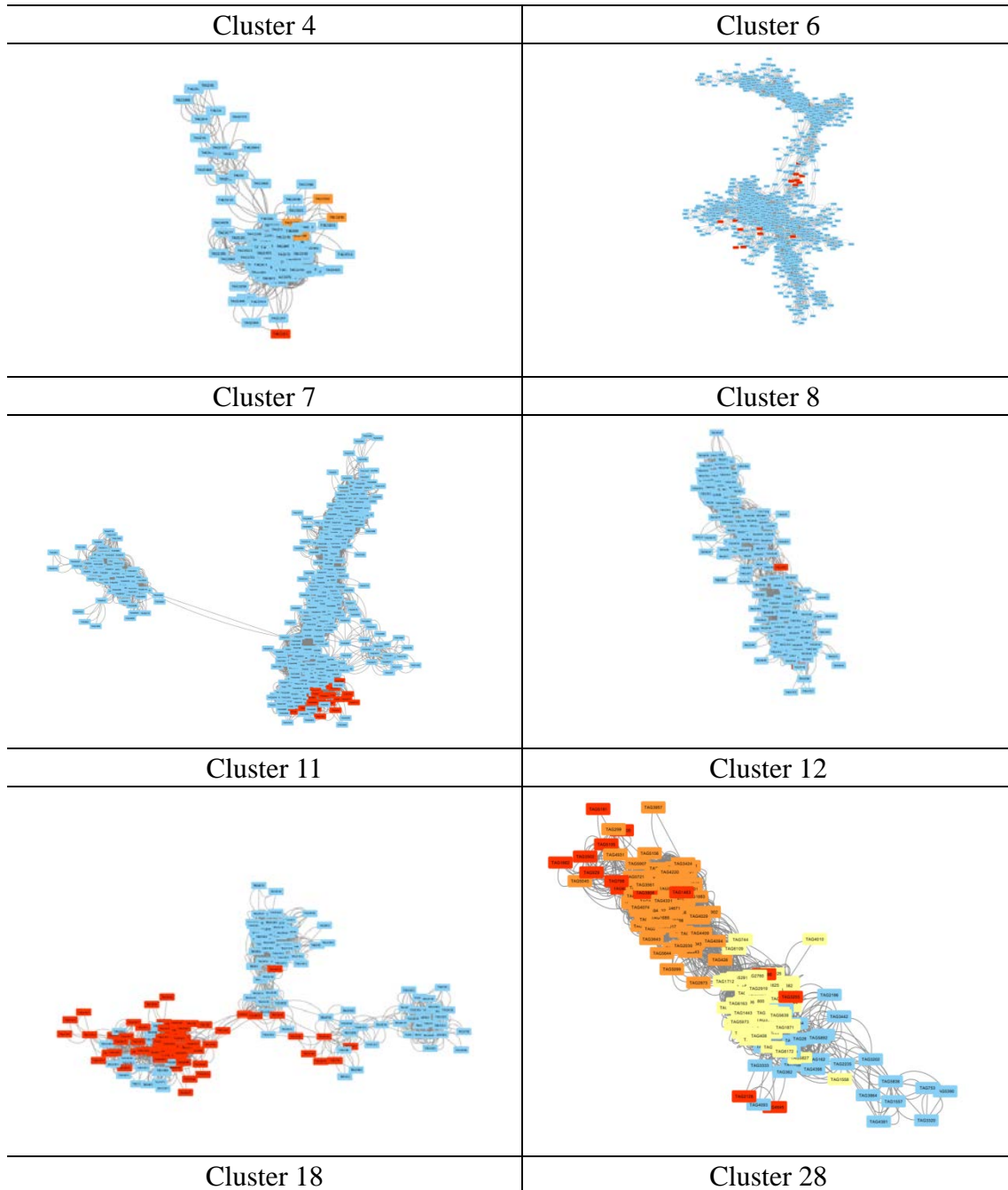


Figure 27: Distribution of genes related to EMERGENCE.

### 3.54 Marking the genes related to FLOWER\_COLOR in genes network (written

by Jiachen, checked by Jiajun)

The 415 genes related to FLOWER\_COLOR were clustered into 3 clusters. We distinguished these three clusters of genes related to FLOWER\_COLOR with red, orange, and yellow. The results show that most of the genes related to FLOWER\_COLOR are concentrated in cluster 12, and a small part of the genes are scattered in the other 7 clusters.



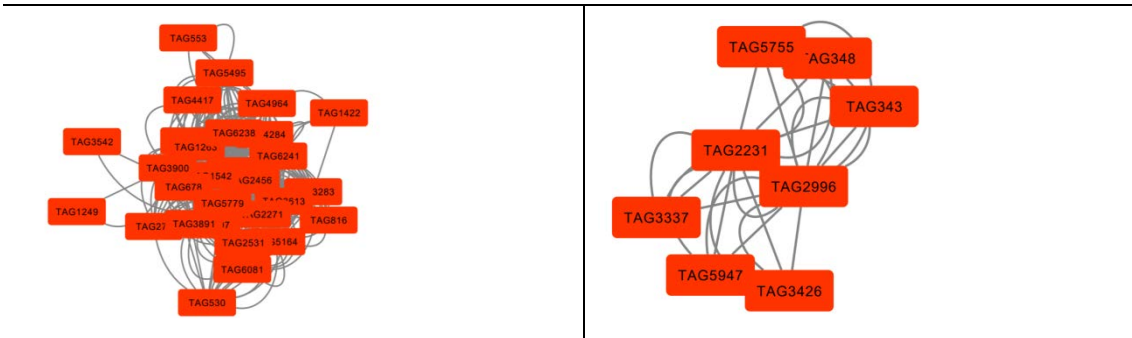
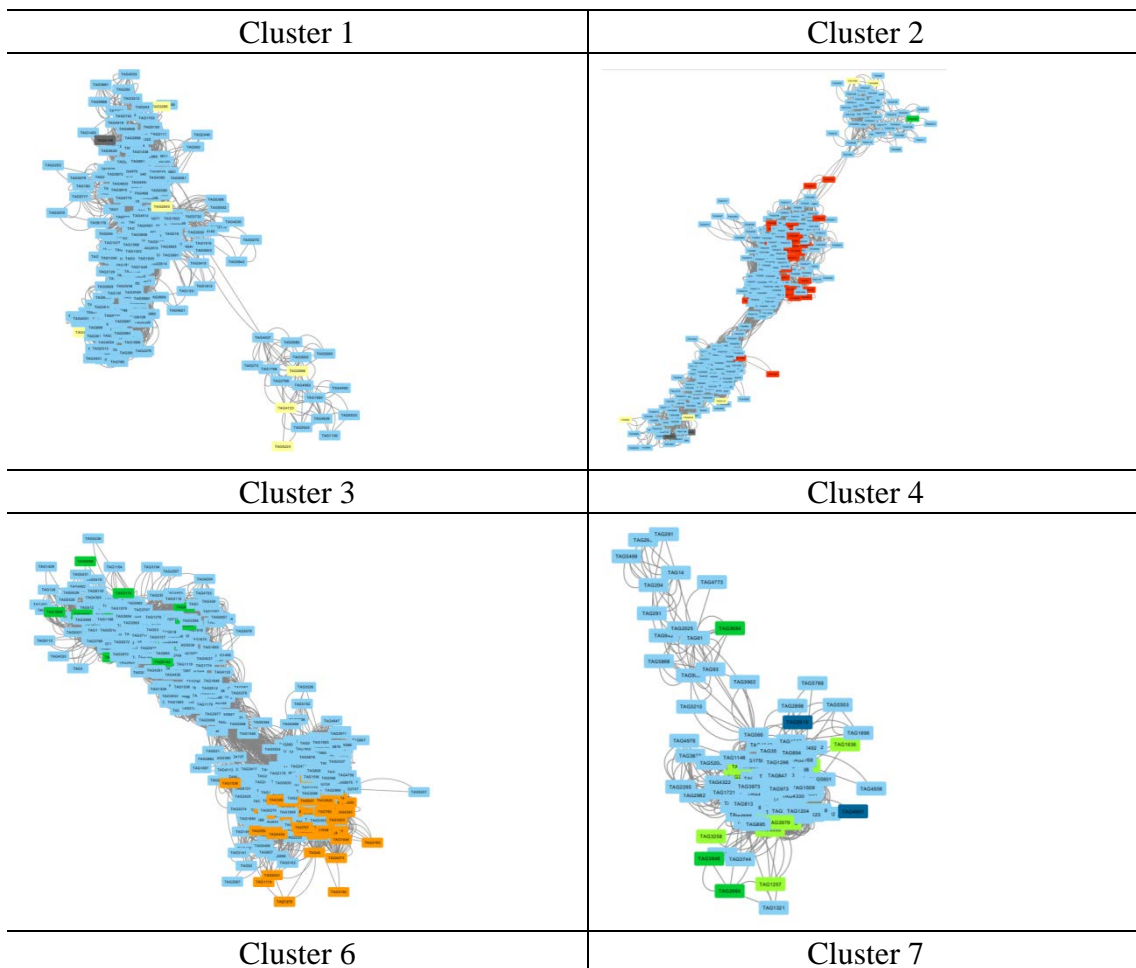
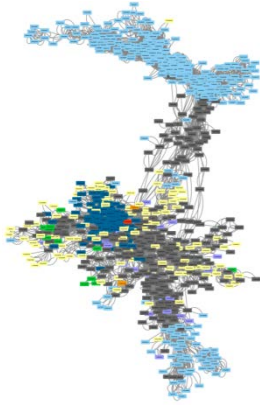


Figure 28: Distribution of genes related to FLOWER\_COLOR.

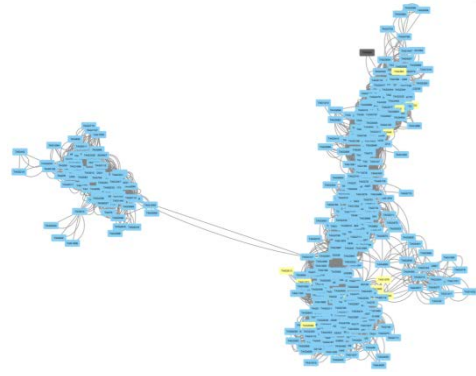
### 3.55 Marking the genes related to FLOWER\_TIME in genes network (written by Jiachen, checked by Jiajun)

The 1312 genes related to FLOWER\_TIME were clustered into 8 clusters. We distinguished these eight clusters of genes related to FLOWER\_TIME with red, orange, yellow, green, greenyellow, midnightblue, purple and grey. The results show that most of the genes related to FLOWER\_TIME are concentrated in cluster 6, and a small part of the genes are scattered in the other 18 clusters.





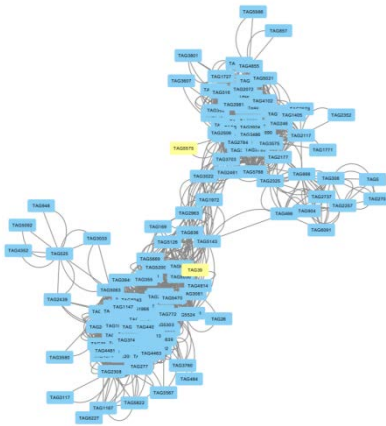
Cluster 8



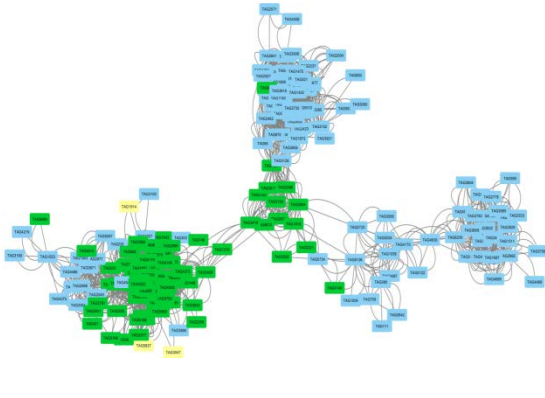
Cluster 9



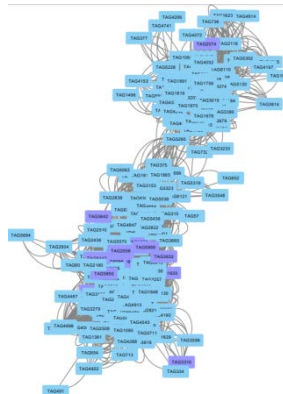
Cluster 11



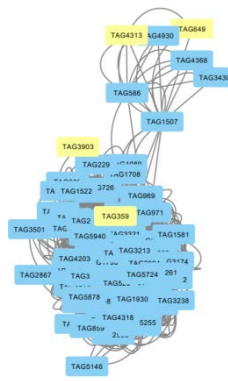
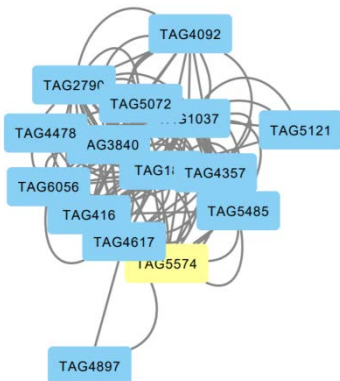
Cluster 14



Cluster 15



Cluster 16



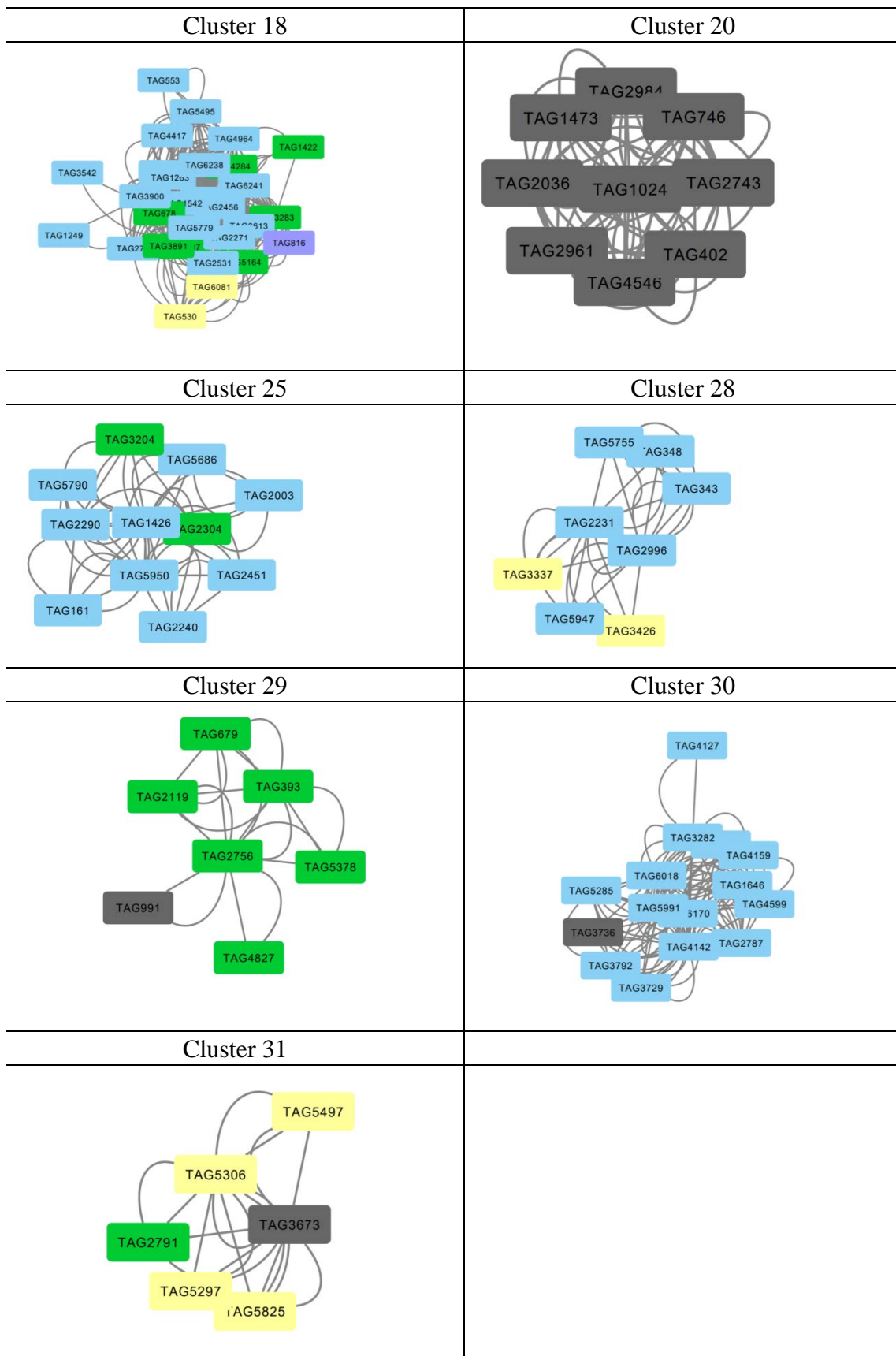

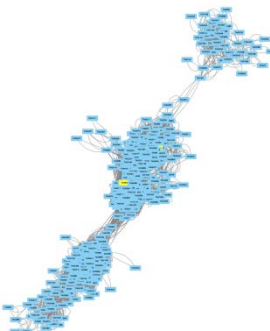

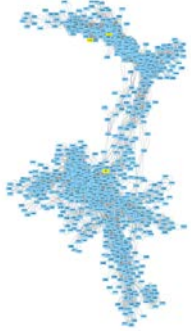




Figure 29: Distribution of genes related to FLOWER\_TIME.

**3.56 Marking the genes related to GLUCOSE in genes network** (written by Jiachen, checked by Jiajun)

The 314 genes related to GLUCOSE were clustered into 3 clusters. We distinguished these three clusters of genes related to GLUCOSE with red, orange, yellow, green. The results show that most of the genes related to GLUCOSE are concentrated in cluster 9, and a small part of the genes are scattered in the other 12 clusters.

Cluster 1	Cluster 2
	
Cluster 3	Cluster 6
	
Cluster 7	Cluster 8
	
Cluster 9	Cluster 12



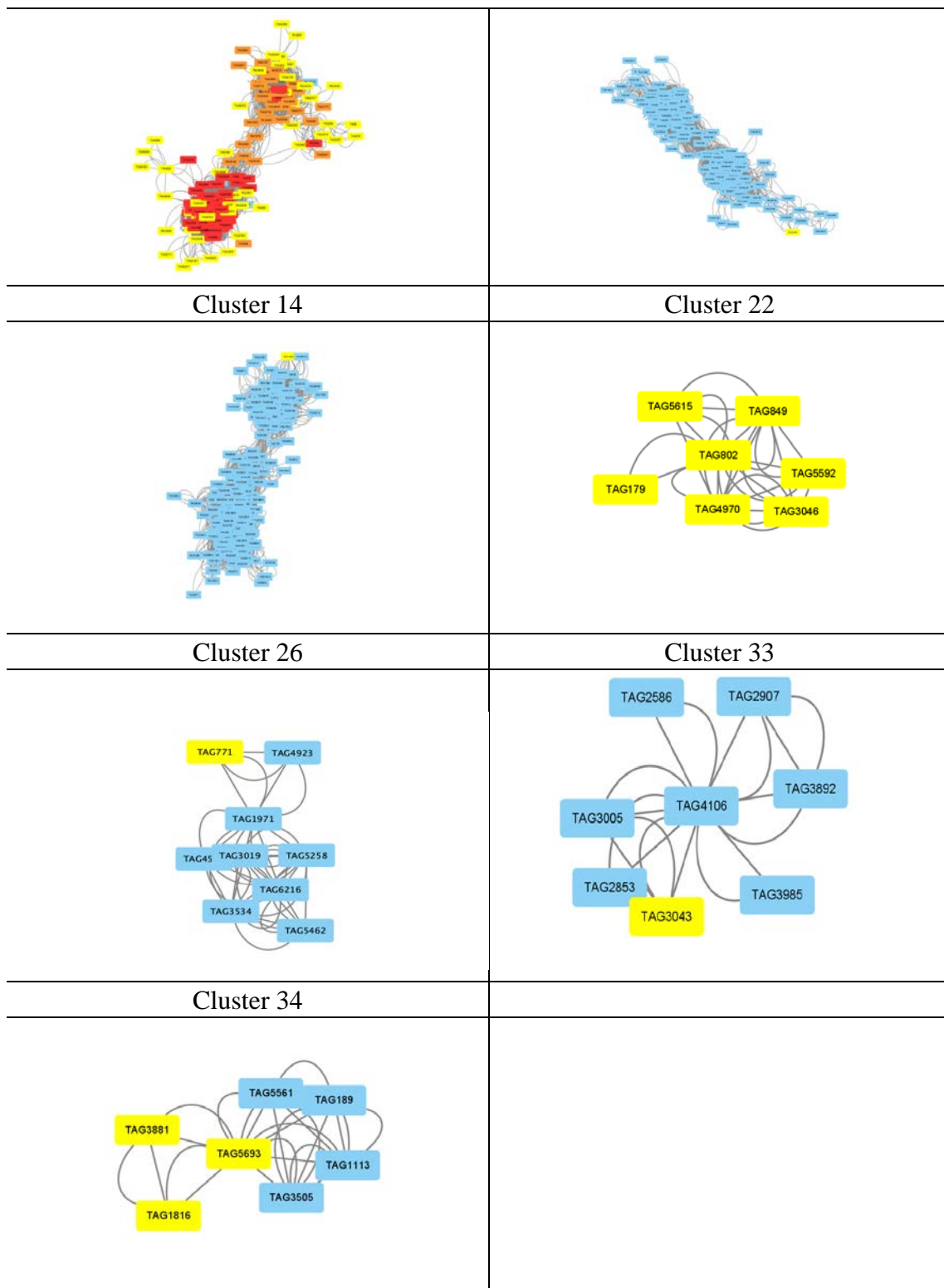
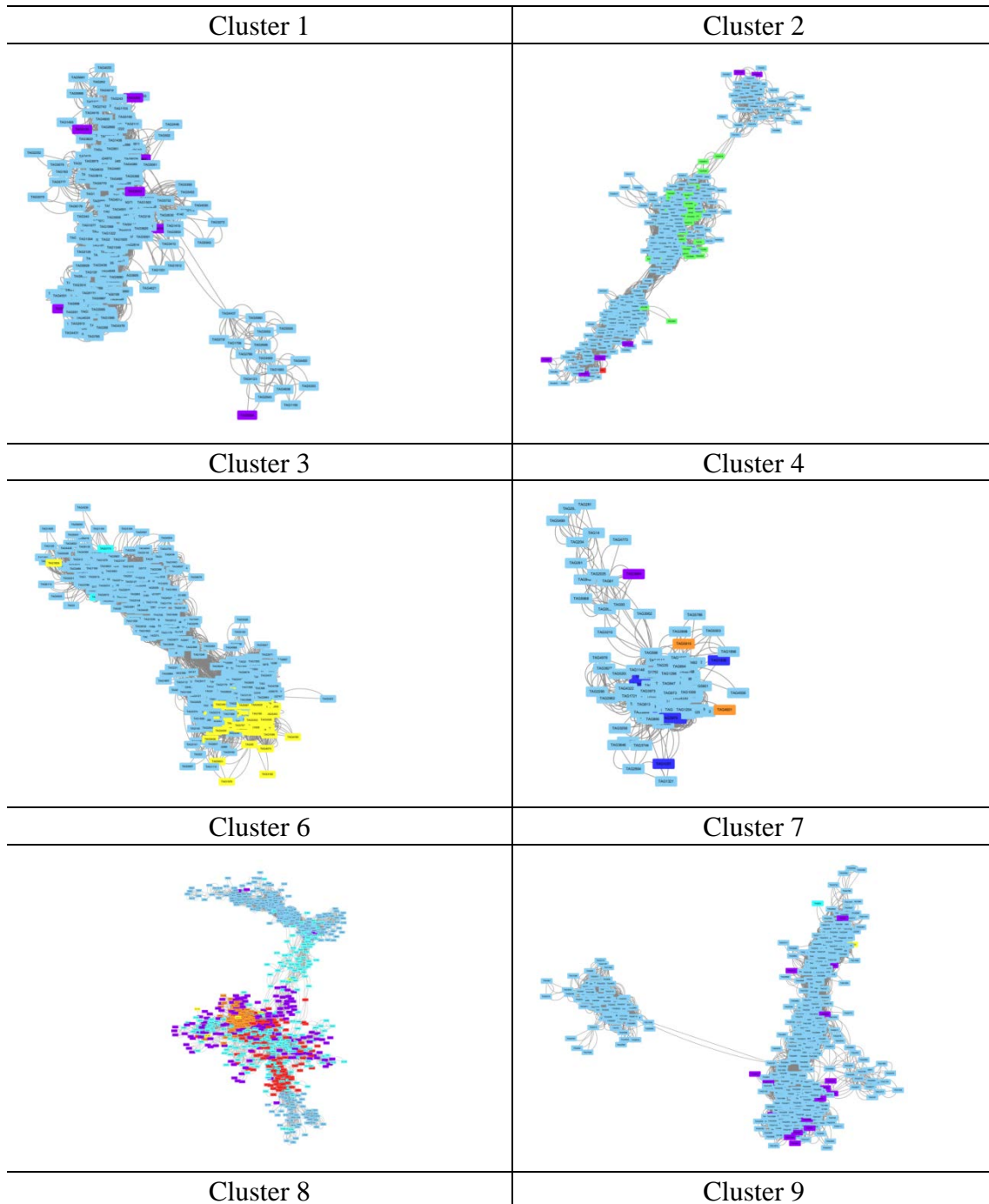


Figure 30: Distribution of genes related to *GLUCOSE*.

### 3.57 Marking the genes related to MATURITY in genes network (written by Jiachen, checked by Jiajun)

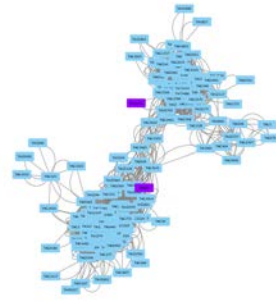
The 1297 genes related to MATURIT were clustered into 7 clusters. We distinguished

these seven clusters of genes related to MATURITY with red, orange, yellow, green, skyblue, midnightblue and purple. The results show that most of the genes related to MATURITY are concentrated in cluster 6, and a small part of the genes are scattered in the other 17 clusters.

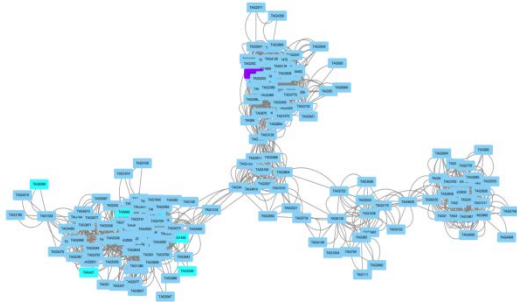




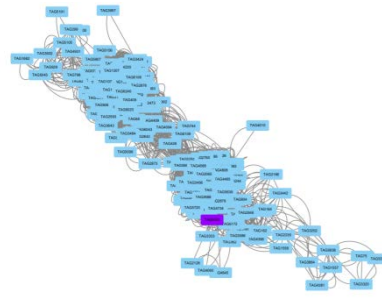
Cluster 11



Cluster 12



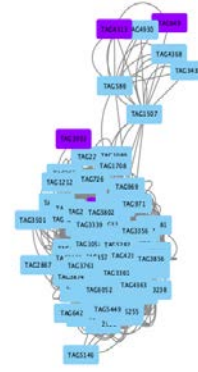
Cluster 14



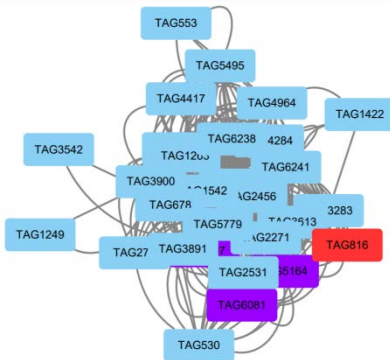
Cluster 16



Cluster 18



Cluster 20



Cluster 28



Cluster 29

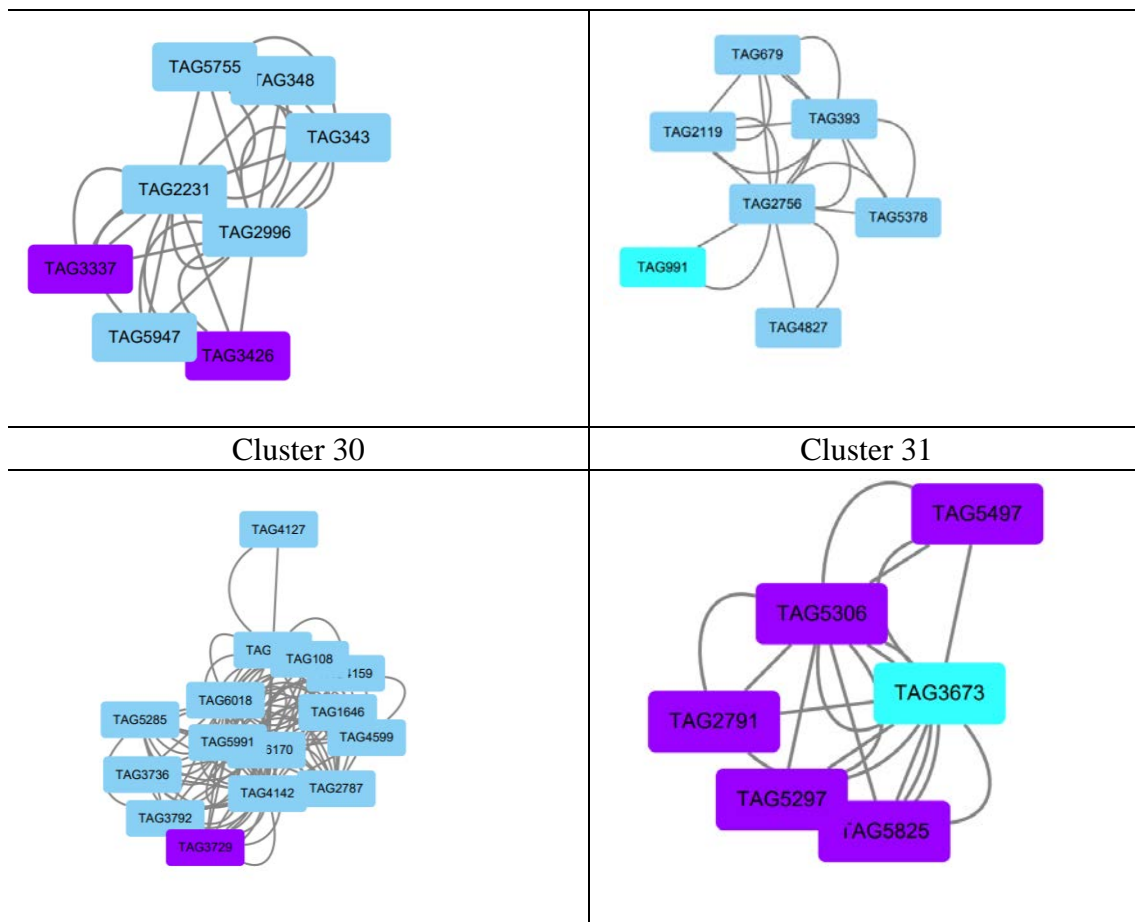


Figure 31: Distribution of genes related to MATURITY.

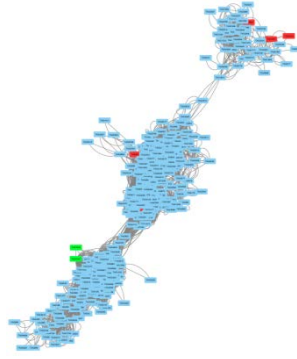
**3.58 Marking the genes related to RED\_SKIN in genes network** (written by Jiachen, checked by Jiajun)

The 618 genes related to RED\_SKIN were clustered into 4 clusters. We distinguished these four clusters of genes related to RED\_SKIN with red, orange, yellow, green. The results show that most of the genes related to RED\_SKIN are concentrated in cluster 7, and a small part of the genes are scattered in the other 9 clusters.

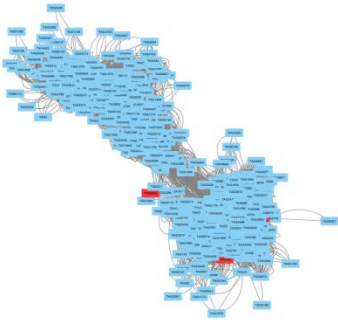
Cluster 1	Cluster 2
-----------	-----------



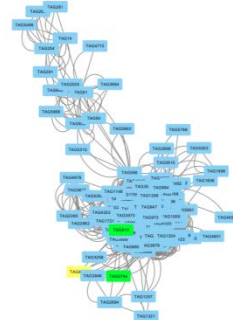
Cluster 3



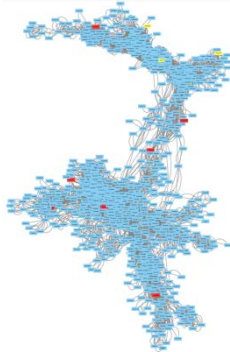
Cluster 4



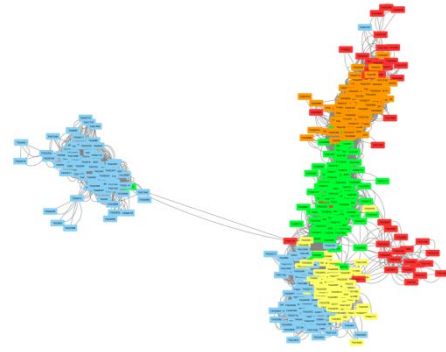
Cluster 6



Cluster 7



Cluster 8



Cluster 11



Cluster 16



Cluster 30

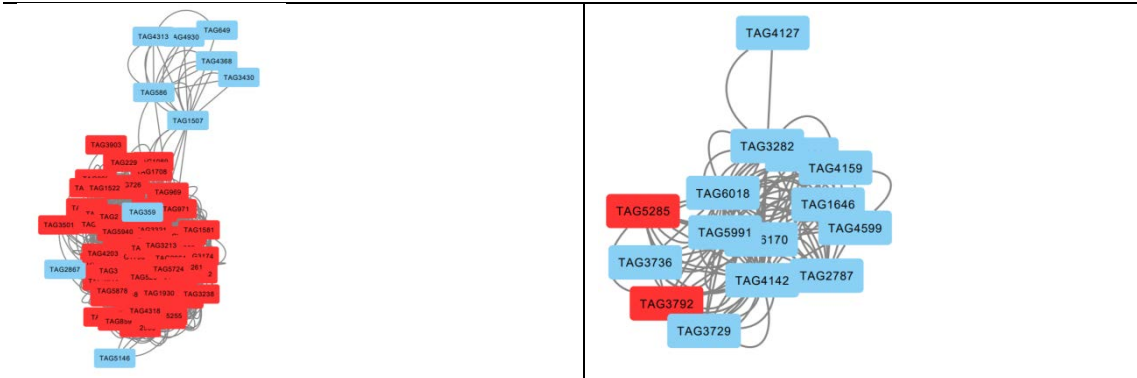
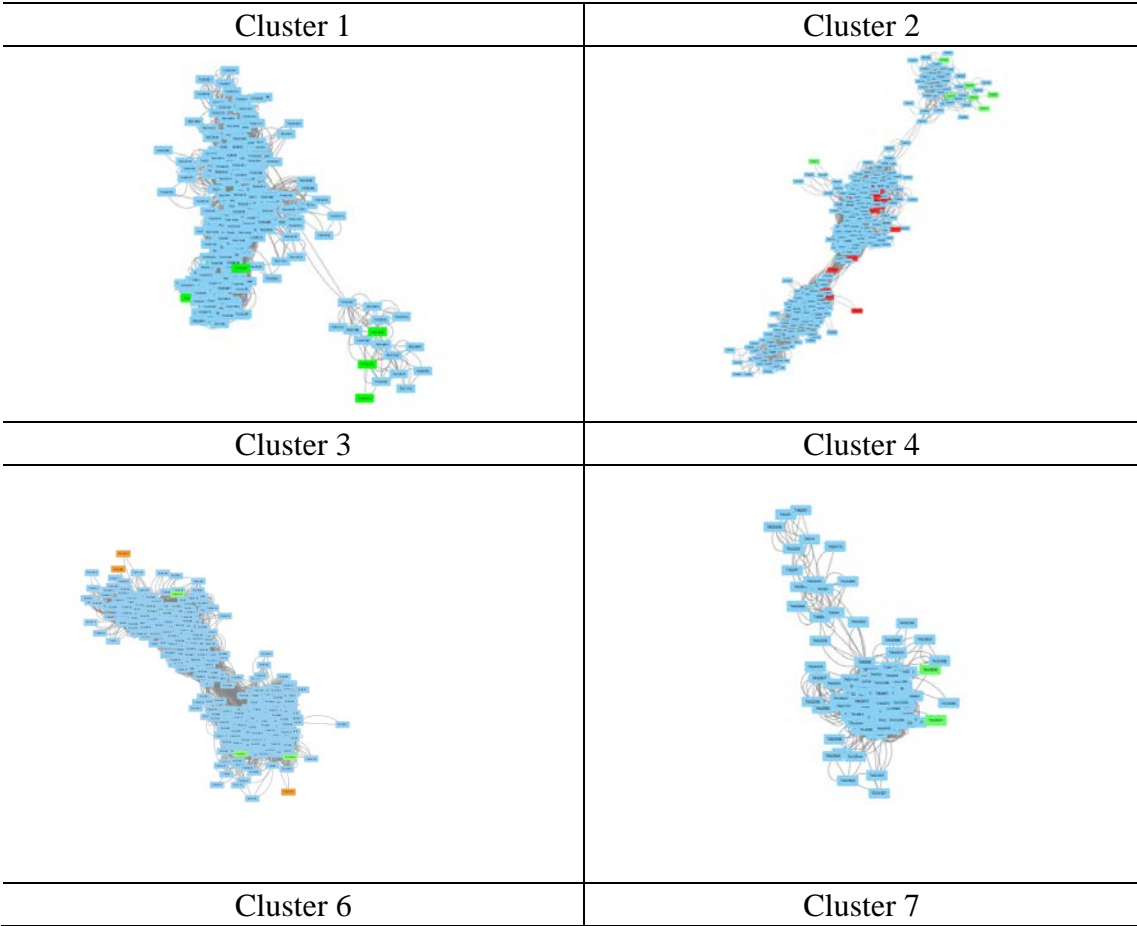


Figure 31: Distribution of genes related to RED\_SKIN.

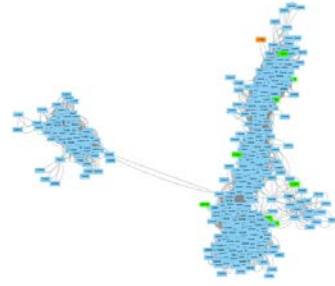
**3.59 Marking the genes related to SCAB\_LS in genes network** (written by Jiachen, checked by Jiajun)

The 653 genes related to SCAB\_LS were clustered into 4 clusters. We distinguished these four clusters of genes related to SCAB\_LS with red, orange, yellow, green. The results show that most of the genes related to SCAB\_LS are concentrated in cluster 6, and a small part of the genes are scattered in the other 13 clusters.





Cluster 8



Cluster 9



Cluster 12



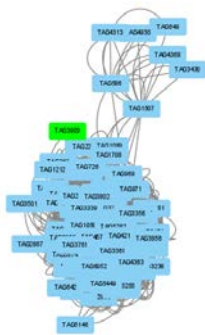
Cluster 14



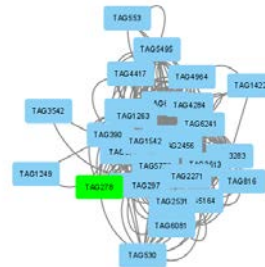
Cluster 16



Cluster 18



Cluster 20



Cluster 33

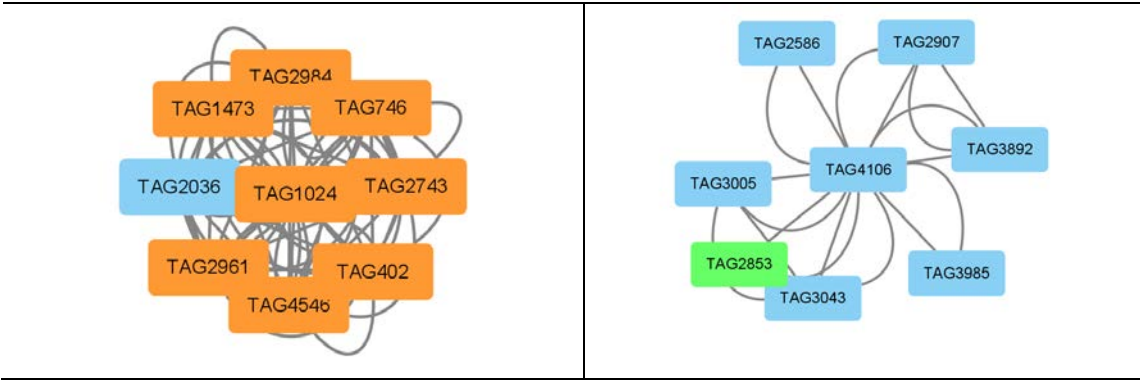
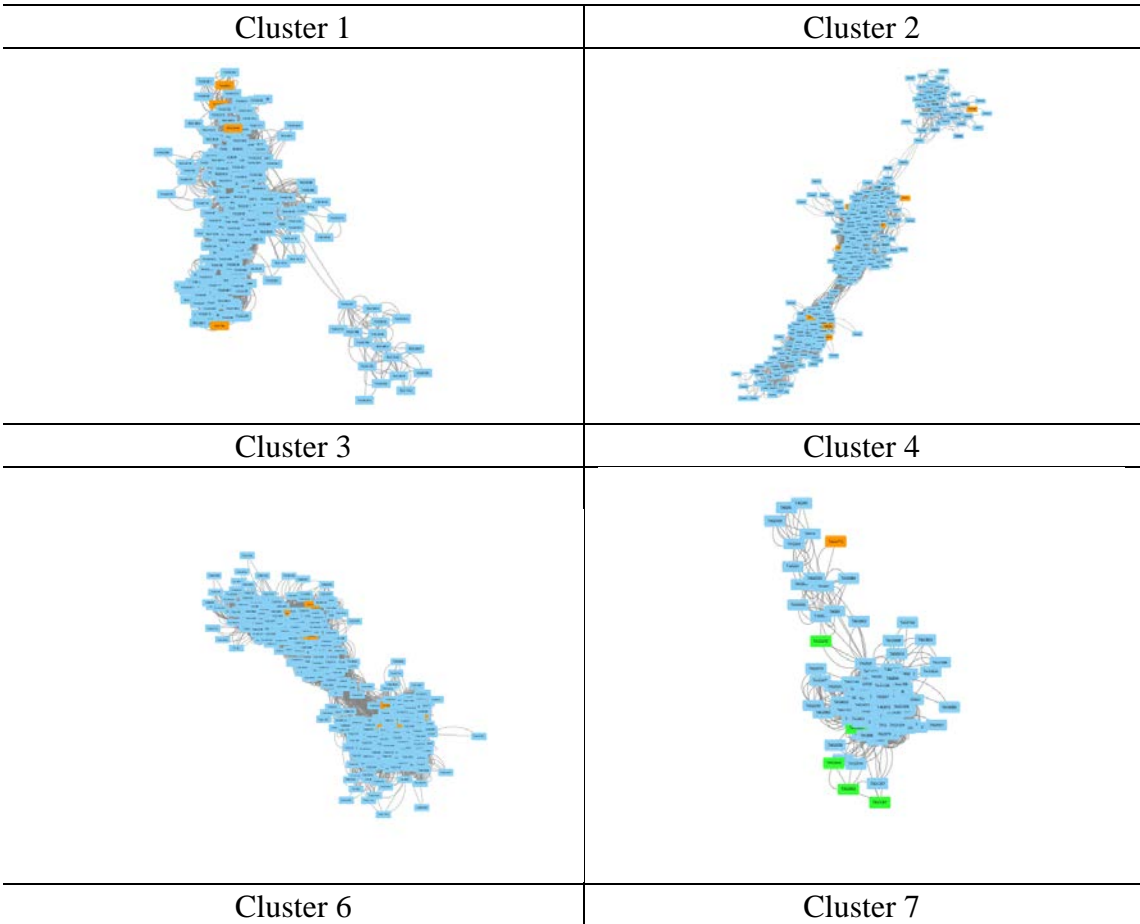


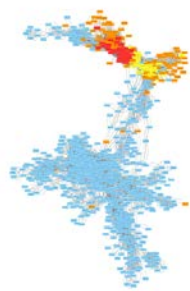
Figure 32: Distribution of genes related to SCAB\_LS .

**3.59\* Marking the genes related to SPROUT\_LS in genes network** (written by Jiachen, checked by Jiajun)

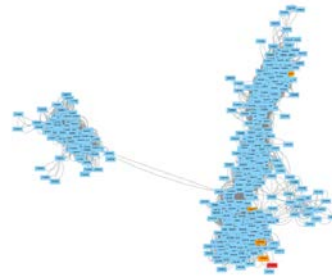
The 536 genes related to SPROUT\_LS were clustered into 4 clusters. We distinguished these four clusters of genes related to SPROUT\_LS with red, orange, yellow, green. The results show that most of the genes related to SPROUT\_LS are concentrated in cluster 6 and cluster 11, and a small part of the genes are scattered in the other 13 clusters.







Cluster 9



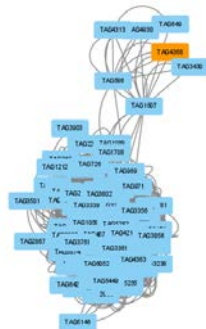
Cluster 11



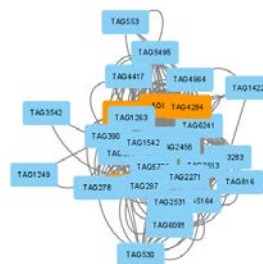
Cluster 16



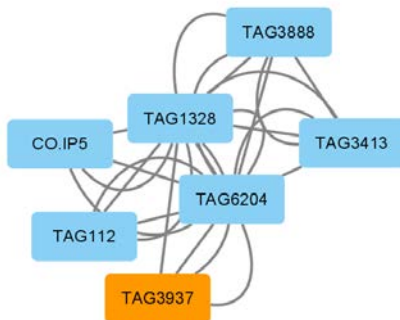
Cluster 18



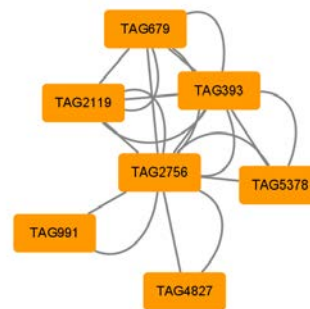
Cluster 24



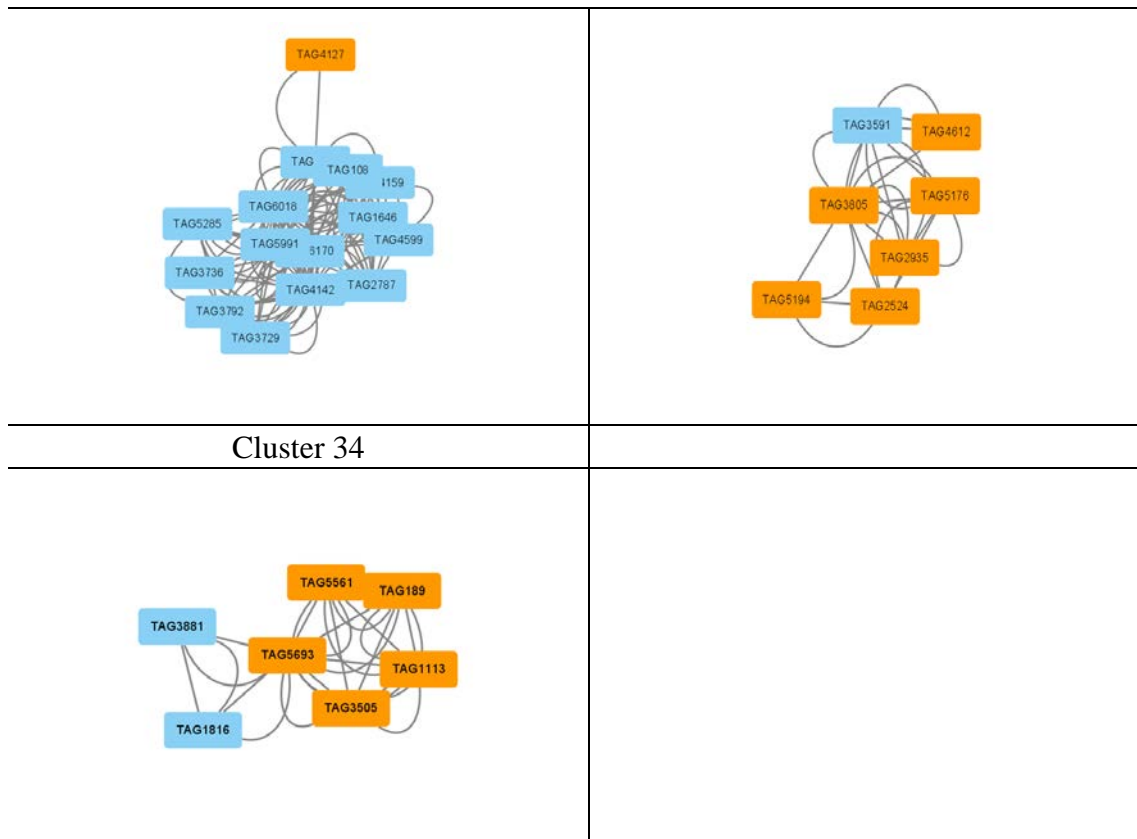
Cluster 29



Cluster 30




Cluster 32





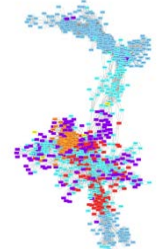
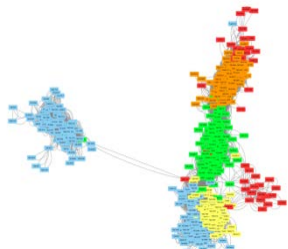


Cluster 34

Figure 33: Distribution of genes related to *SPROUT\_LS*.

The genes related to the nine traits all show two characteristics in the distribution of the gene network: most of them are concentrated in the same small gene network; a few genes are scattered in other clusters. The details of the nine traits are summarized in the table below:

Traits	Figure of the cluster where most genes are concentrated	the cluster where most genes are concentrated	Other clusters with gene distribution
CHIP		Cluster 9	Cluster 1, 2, 3, 4, 6, 7, 8, 11, 12, 14, 16, 18, 22, 26, 33

EMERGENCE		Cluster 3	Cluster 1, 2, 4, 6, 7, 8, 11, 13, 18, 23, 25, 27, 29, 30
FLOWER_COLOR		Cluster12	Cluster 4, 6, 7, 8, 11, 18, 28
FLOWER_TIME		Cluster 6	Cluster 1, 2, 3, 4, 7, 8, 9, 11, 14, 15, 16, 18, 20, 25, 28, 29, 30, 31
GLUCOSE		Cluster 9	Cluster 1, 2, 3, 6, 7, 8, 12, 14, 22, 26, 33, 34
MATURITY		Cluster 6	Cluster 1,2,3,4,7,8,9,11, 12,14,16,18,20,28,29,30,31
RED_Skin		Cluster 7	Cluster 1,2,3,4,6,8,11,16, 30

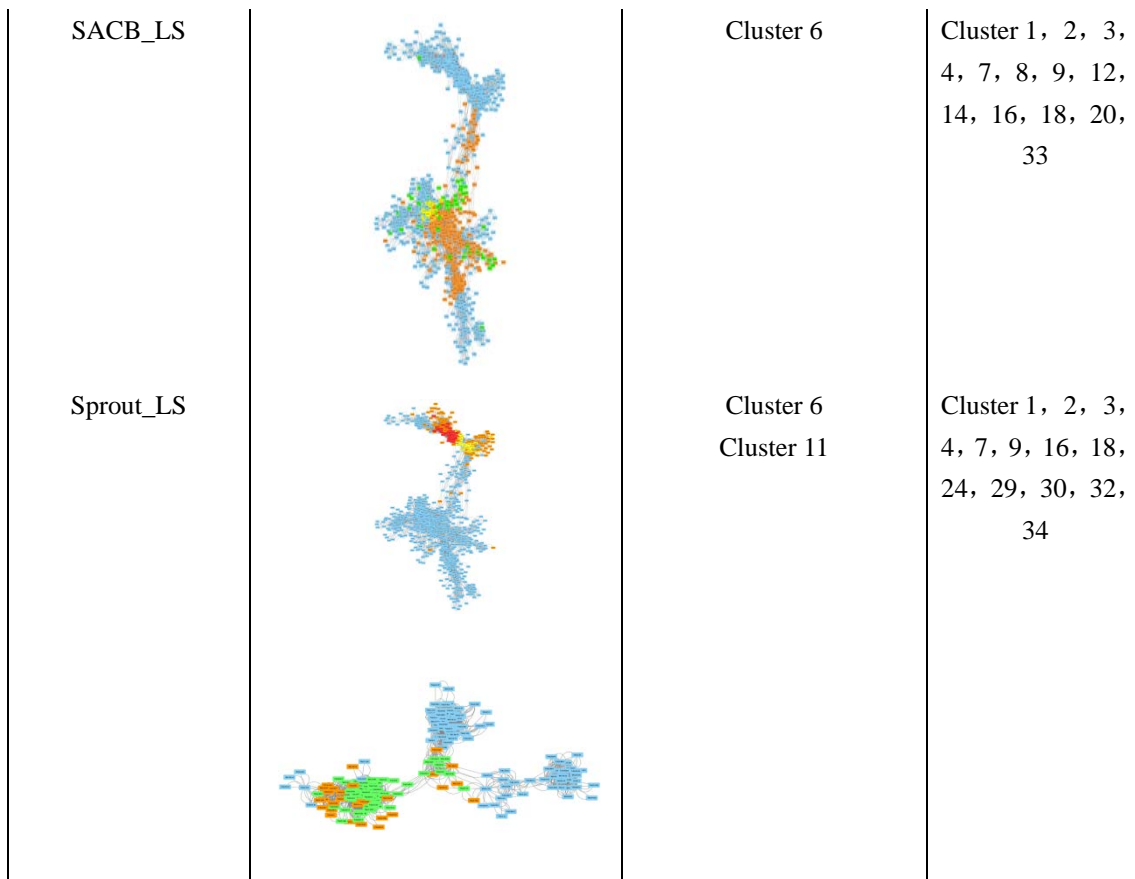


Figure 34 :The distribution of trait-related genes in the gene network

.Using Cytoscape to mark genes related to traits on the gene network, the results show that most of the genes related to CHIP and GLUCOSE are clustered on cluster 9 of the gene network; most genes related to EMERGENCE are clustered on cluster 3 of the gene network ; The genes related to FLOWER\_COLOR are mostly concentrated on cluster 12 of the gene network; the genes related to RED\_SKIN are mostly concentrated on cluster 7; the genes related to FLOWER\_TIME, MATUTRITY, SCAB\_LS and SPROUT\_LS are mostly concentrated on cluster 6. In addition, a large part of the genes related to SPROUT\_LS are concentrated on cluster 11. Although the entire gene network is clustered into 34 small gene networks, only a few gene networks have a large number of genes related to traits, and most of the gene networks are only scattered genes related to traits.

### 3.6 An additional work

#### 3.61 data distribution (written by Yuexu, checked by Jiachen)

Making clear the distribution of data can lay the foundation for the subsequent operation, by using R code to plot the histograms for genes which selected randomly

by R, the picture shows an unobvious normal distribution, and after taking logarithmic function for data, the plots looks like a normal distribution obviously.

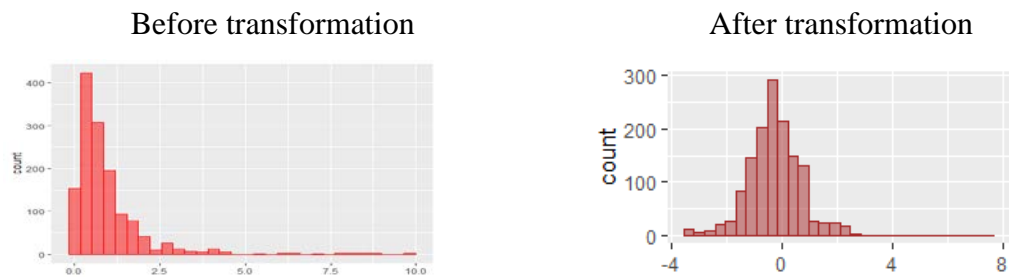


Figure 35: after log transformation, the LOD scores follow normal distribution.

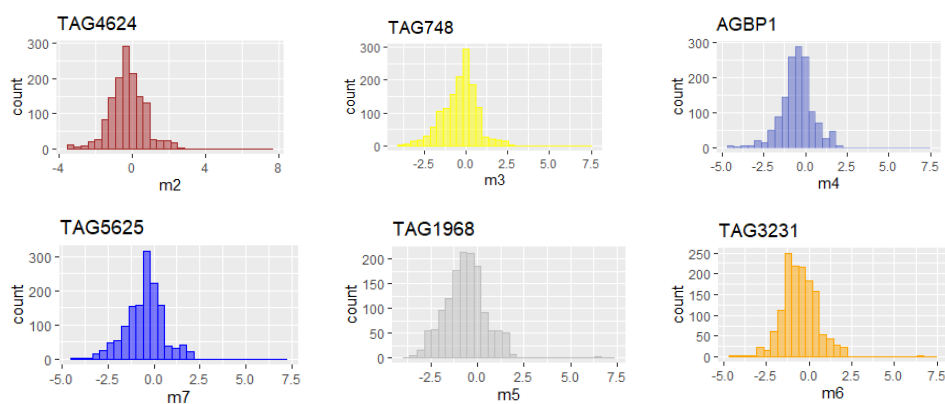
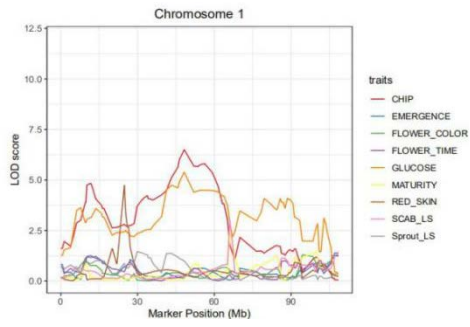


Figure 36: in order to avoid outliers, 6 genes were randomly choose to check their distribution, and the results show log (LOD scores) follow normal distribution.

In order to avoid special situations, 6 genes were randomly chose and plotted the histograms, and Figure 7 shows that after log transformation, LOD scores follow normal distribution.

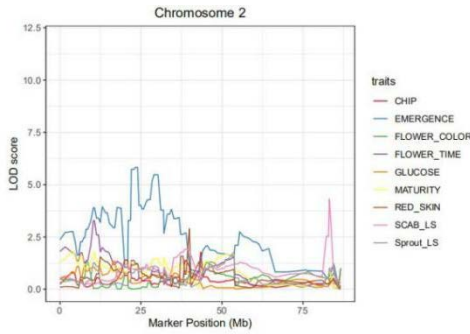
### 3.62 Distribution of karyotypes on chromosomes

As Finegold (2019) noted, there are from hundreds to thousands of genes in one chromosomes, and traits can be determined by any genes and in general determined by one than one genes. Therefore, more than one traits could be highly correlated to on chromosome. Figure 8 shows a line chart for the LOD scores of 9 traits corresponding to 12 chromosomes which will show the relationship between traits and chromosomes. As mentioned before, if LOD scores of gene larger than 3, it will be recognized as perform significantly, similarly, if the line of LOD scores of each trait is higher than 3, then the trait is highly correlated to one chromosome.



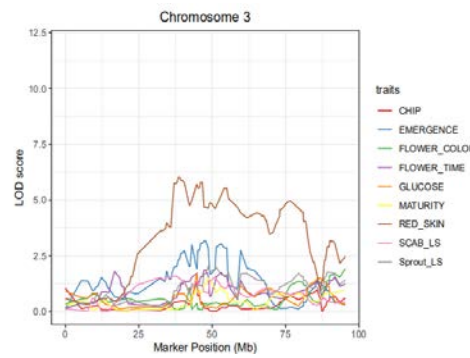
Significant traits:

CHIP  
GLUCOSE  
RED\_SKIN



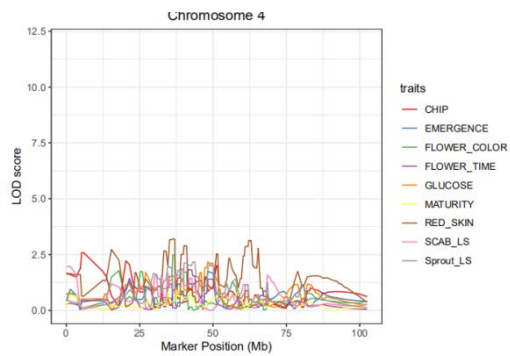
Significant traits:

EMERGENCE  
SCAB\_LS



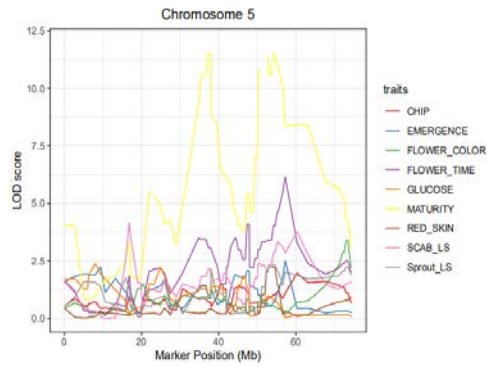
Significant traits:

RED\_SKIN  
EMERGENCE



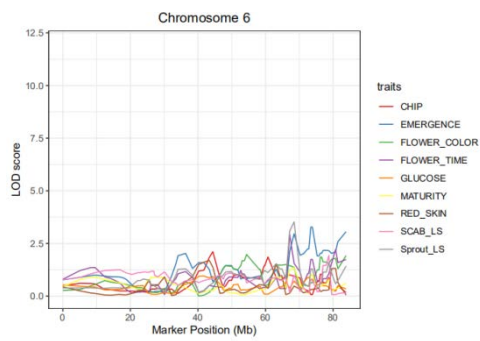
Significant traits:

RED\_SKIN



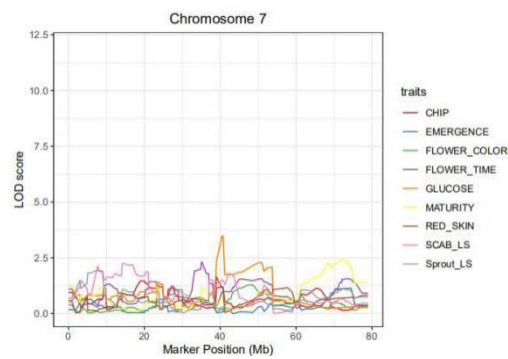
Significant traits:

MATURITY  
FLOWER\_TIME  
SCAB\_LS  
FLOWER\_COLOR



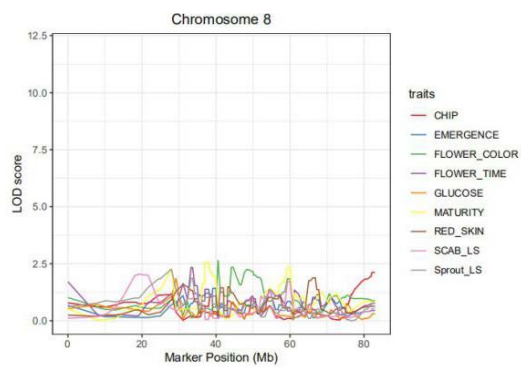
Significant traits:

Sprout\_LS  
EMERGENCE



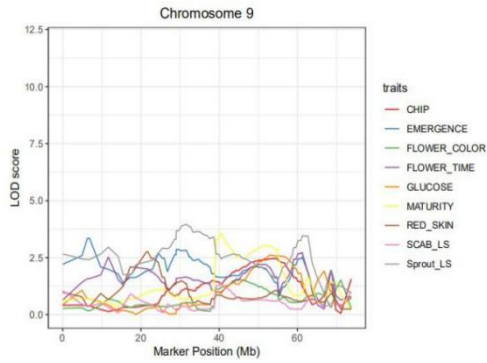
Significant traits:

GLUCOSE



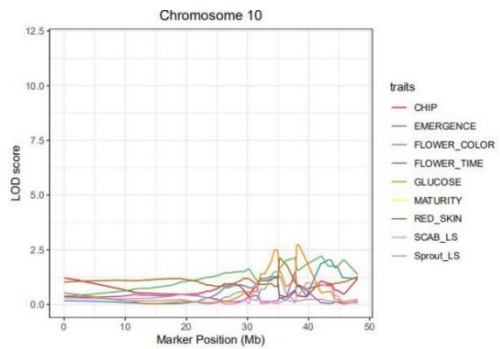
Significant traits:

NA



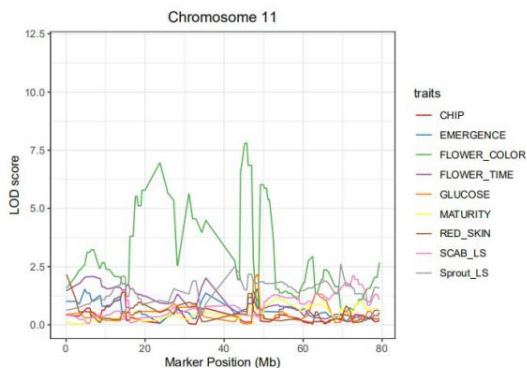
Significant traits:

EMERGENCE  
Sprout\_LS  
MATURITY



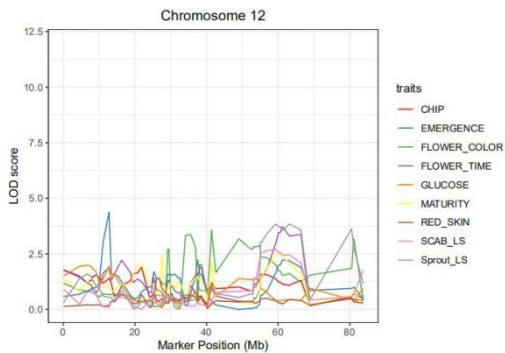
Significant traits:

NA



Significant traits:

FLOWER\_COLOR



Significant traits:

EMERGENCE  
FLOWER\_TIME  
FLOWER\_COLOR  
Sprout\_LS

Figure 37: For most chromosomes, there are more than one traits perform significantly, such as chromosome 1 is highly correlated with CHIP, GLUCOSE, and RED\_SKIN, chromosome 5 is highly correlated with FLOWER\_TIME, SCAB\_LS, and FLOWER\_COLOR, and especially MATURILTY. However, there also exists some chromosomes with no highly correlated trait like chromosome 8 and 10.



## **4. Discussion** (written by Yuexu & Jiachen, checked by Yuexu & Jiachen)

When using DBSCAN to make clusters, the selection of both parameters  $\epsilon$  and  $\text{minpts}$  is very important, however we only choose these two parameters by experience and it would be better to find a more accurate way to choose them in the future.

When doing NMF analysis, there are various ways to normalize the matrix  $W$ . This article uses the L1 norm to construct a diagonal matrix to update the  $W$  matrix and the  $H$  matrix. In future analysis, it may be possible to regularize the matrix to increase the sparsity of the matrix for more specific feature selection.

We hope to find genes related to traits by screening genes to improve the quality and yield of potatoes in a targeted manner. After the research in this paper, we have found the relationship between genes and genes and constructed a gene network; we have also found genes related to traits and found the position of these genes in the gene network. This is of progressive significance for improving the quality of potatoes and increasing the yield.

## Reference

1. Brunet, J., Tamayo, P., Golub, T., Mesirov, J., & Lander, E. (2004). Metagenes and Molecular Pattern Discovery Using Matrix Factorization. *Proceedings of the National Academy of Sciences of the United States of America*, 101(12), 4164-4169.  
<http://www.jstor.org/stable/3371580>
2. Finegold, N. D. (2019). Genes and Chromosomes. Retrieved May 4, 2021, from <https://www.msmanuals.com/home/fundamentals/genetics/genes-and-chromosomes#:~:text=A%20chromosome%20contains%20hundreds%20to%20thousands%20of%20genes.,are%20the%20result%20of%20a%20new%20gene%20mutation.>
3. Hahsler, M., Piekenbrock, M., & Doran, D. (2019). Dbscan: Fast density-based clustering with r. *Journal of Statistical Software*, 91(1). DOI: <https://doi:10.18637/jss.v091.i01>
4. Kriegel, H-P., Kröger, Peer., Sander, J., & Arthur, Z. (2011). Density-based Clustering. *WIREs Data Mining and Knowledge Discovery*. 1(3), 231–240. DOI: <https://doi: 10.1002/widm.30>.
5. Krishan. (2016). What is NMF and What Can You do With It? Retrieved from [Iksinc.online/2016/03/21/what-is-nmf-and-what-can-you-do-with-it/](https://www.ksinc.com/online/2016/03/21/what-is-nmf-and-what-can-you-do-with-it/)
6. Sander, J., Ester, M., Kriegel, HP. et al. (1998). Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. *Data Mining and Knowledge Discovery* 2, 169–194 (1998). DOI: <https://doi.org/10.1023/A:1009745219419>
7. Schubert, E., Sander, J., Martin, E., Kriegel, H-P., & Xu, X.W. (2017). DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Trans. Database Syst.* 42, 3, Article 19. DOI: <https://doi.org/10.1145/3068335>.
8. Weisstein, Eric W. (2021). "Normal Distribution." Retrieved from MathWorld--A Wolfram Web Resource.  
<https://mathworld.wolfram.com/NormalDistribution.html>
9. Zhang, CH., Huang, DS., Zhang, L. et al. (2009). Tumor Clustering Using Nonnegative Matrix Factorization With Gene Selection. *IEEE Transactions on Information Technology in Biomedicine*, 13(4), 599-607.