Xmotif Biclustering Analysis on Genes Expression Datasets of Maize Growth Stages

Maidah Maidah[1] and Husna Nugrahapraja[2,3]*

1. Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Islam Indonesia, Yogyakarta, INDONESIA
2. School of Life Sciences and Technology, Institut Teknologi Bandung, Bandung 40132, INDONESIA
3. University Centre of Excellence for Nutraceuticals, Biosciences and Biotechnology Research Center, Institut Teknologi Bandung, Bandung 40132, INDONESIA
*nugrahapraja@sith.itb.ac.id

Abstract

The maize growth stage consists of the vegetative growth stage and the reproductive growth stage. Based on genes expression datasets of the maize growth stages, we aimed to perform xMotif biclustering to classify the organ-specific expression. We performed two stages of filtering, first, removing the row line, which has an average equal to zero (0) and secondly, uses the t-test function (p-value < 0.0001), resulting in data with dimensions of 512 x 68.

As a result, we found 53 biclusters with the highest average expression value in one bicluster consisting of 22 genes. The highest expression value is Zm00001d034773 and the blade organ has the highest expression genes value. The blade is one of maize vegetative phases. In conclusion, bicluster analysis can be used as a classification study in maize genes expression datasets.

Keywords: Maize growth stages, xMotif, Blade, Bicluster, Vegetative Phase.

Introduction

The breakthrough on the maize genetics and genomics database is providing the genetic information with website-based interfaces[1]. Another database such as qTeller[2] is also providing genes expression datasets on maize growth datasets. Large-scale genes expression research and genome sequencing projects support information that can be used to identify or predict cellular regulation processes[3]. Genes are classified based on the similarity profiles or expression functions and tend to have a cluster of genes that are regulated by the same transcription factors.

Large-scale biological data in the matrices such as genes expression datasets make the researchers need a computational approach so-called data mining. Data mining uses statistical techniques, mathematics, artificial intelligence and machine learning to extract and identify useful information and detailed knowledge from various large databases. One of the data mining methods is biclustering, which is a new technique that develops from cluster techniques in general[4].

In biclustering, rows are grouped based on variables of rows and columns to detect local patterns in data. The biclustering method has many algorithms that have been used to group and identifies patterns in biological data, namely Cheng and church, Plaid, OPSM, ISA, Spectral, xMotifs and Bimax[4].

Biclustering using xMotif algorithm is based on coherent evolution[5]. This algorithm looks for rows with constant values above a set of columns. They call bicluster "motives for preserved gene expression," abbreviated as "xMotif" for genes expression data. For this application, it is essential to find an excellent preprocessing method because the primary purpose of this algorithm is to define the same gene (row) status in the selected sample (column), so-called the conserved gene (row). One way to deal with the state of genes is to utterly discrete data. This study aimed to analyze large-scale maize genes expression datasets from maize qTeller (www.qteller.com/qteller4/), so the genes can be classified based on similarity profiles expression on a specific organ.

Material and Methods

We used genes expression data of maize growth stages from qTeller (www.qteller.com/qteller4/, Accessed on April 14th, 2018) and downloaded all genes expression data from the qTeller database[9-14]. The data contains the 39,348 genes name variable and its expression level, 68 maize organs, and20 chromosomes. We used R programming with multtest package to conduct biclustering using xMotif (expression motif) algorithm analyses[6,7]. We performed feature selection filtering to eliminate or discard the row line, which has an average of zero (0) and used the t-test function (p-value < 0.0001). Filtered data is used for biclustering analysis using the xMotif algorithm.

Results and Discussion

We performed maize growth stages data before analysis was carried out by filtering stage to select a subset of probes available for exclusion or equalization in the investigation. We used an additional package multtest package in R programming [6,7].

We removed the row line, which has an average of zero (0), resulting in 37,479x68 matrix data. We calculated using mt.teststat function to calculate t-test statistics to compare each gene expression in reproductive and vegetative phases. The data assumed to normally distributed, and the variance is not the same. Applying a p-value of 0.0001 resulting data with dimensions of 512 x 68 will be used in the xMotif algorithm. The xMotif algorithm is a search algorithm to identify structures that are preserved. Determination of parameters before analysis is essential given the dependence of the xMotif algorithm on discretion methods. The use of different discretion is avoided because it can cause different results. In this study researchers used 10 discretion, ns=65, nd=100, sd=5 and alpha=0.01. Based on computation on 512 genes and 65 maize organs with the above parameters, we obtained 53 biclusters. The lowest p-value of 4 biclusters is displayed as in table 1.

The four biclusters showed a significant row and the column effect means that each gene's expressions in each cluster and organ are different. The number of genes and organs in each cluster is shown in table 2. Figure 1 shows the expression value of 22 gene names in maize organs and *Zm00001d034773* is the highest expression value. The highest value of gene expression in maize organs is P7_blade2_1. P7_blade2_1 is the stage of leaf development in maize. Figure 2 shows that most organs have relatively the same gene expression and *Zm00001d018356* showed the highest expression value in organs P7_blade_L12_1.

Organ P7_blade_L12_1 is the stage of leaf development in maize. Figure 3 informs the expression of each organ that has a high level of expression in the gene. Developing_Leaf has the highest gene expression value in the gene *Zm00001d013367*, which indicated by the red colour of the gene and the lowest expression on the gene *Zm00001d027323*.

Table 1
Representation Of P-Value from Four Clusters of Bicluster Analysis

Observation of FStat	***P-value***			
	BC1	*BC2*	*BC3*	*BC4*
Row Effect	5.4780e-96	3.8603e-61	2.6184e-19	2.3283e-43
Column Effect	1.4660e-02	7.7259e-01	3.9707e-02	1.9282e-05

Table 2
Representation of p-value from four clusters of bicluster analysis

Number of Clusters found: 53				
First 4 Cluster sizes				
	BC1	BC2	BC3	BC4
Number of genes	22	53	28	22
Number of organs	18	6	7	7

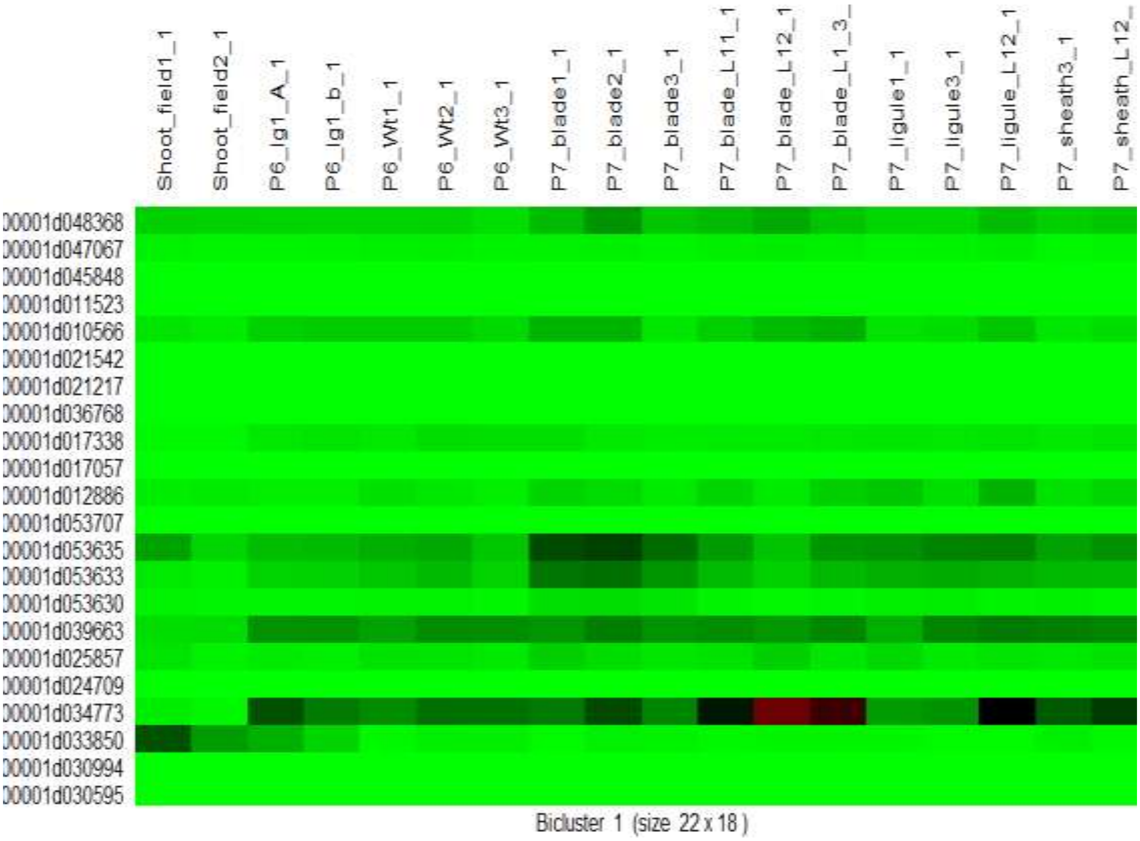

Figure 1: Bicluster 1 Visualization

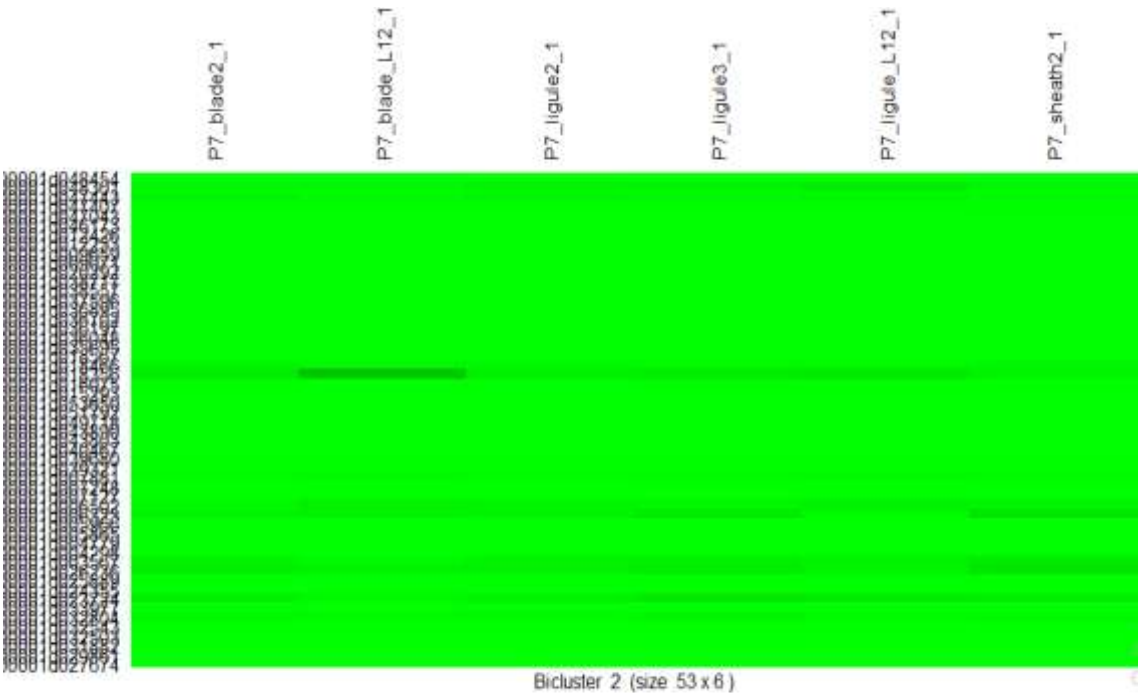

Figure 2: Bicluster 2 Visualization

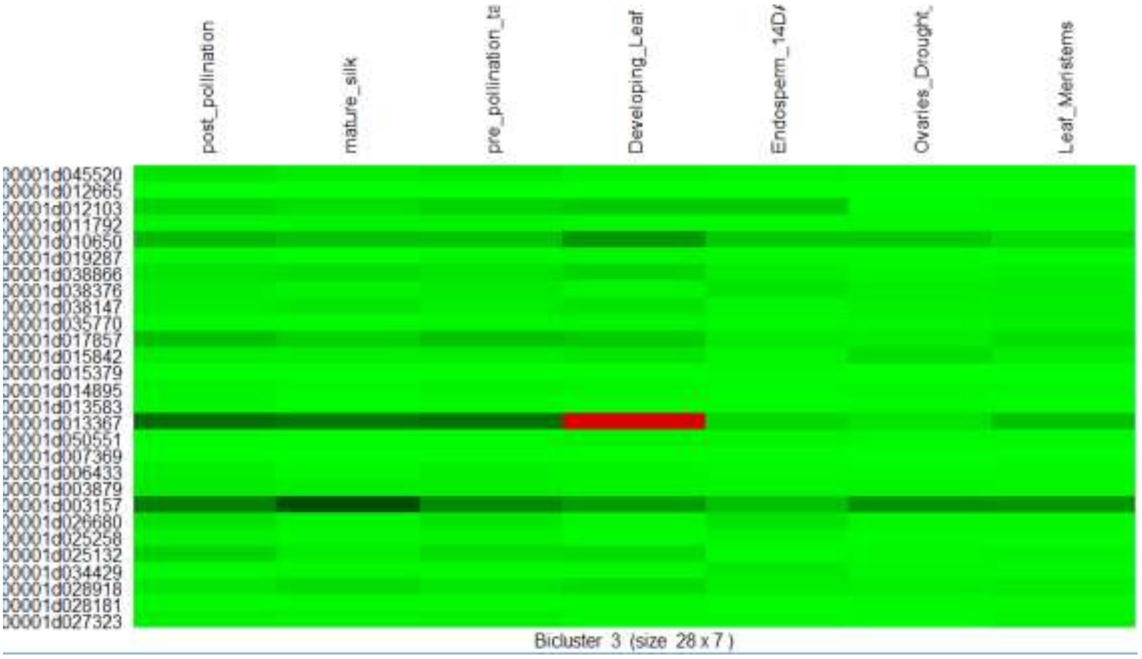

Figure 3: Bicluster 3 Visualization

In line with figure 3, figure 4 showed that the Pre_pollination_tassel organ has the highest gene expression level in the gene name *Zm00001d048585*. Pre_pollination_tassel is the reproductive stage before the male flower fertilizes the female flower (cob).

Here, we presented 19 biclusters from 53 biclusters. The distribution of bicluster members is obtained from heatmap analysis and showed that heatmap allows genes to overlap, and most colours gather on the right side. 3D plot displays information on the value of gene expression on maize organs (condition) and genes (genes). The expression of organs in the 5-10 genes was fluctuated, while the 20th gene was not expressed. As a result, the average gene expression in each organ is relatively high (Table 3).

Fifty-three biclusters are represented by the highest average organ expression value (Table 3). There are seven biclusters referred to in the reproductive phase. The highest expression value was 38.788 in the 41st bicluster, coming from the twenty-five DAP endosperm organ cluster where endosperm is a food reserve for embryos in maize kernels. This condition can be interpreted that expressed genes are used to build embryonic structures in maize. In contrast, in the vegetative phase, there are 43 biclusters.

There are four highest gene expression values with 220.968, 215.178, 200.534 and 114.522 in cluster 1 organ P7 blade 1, cluster 20 organs of bundle sheath cells, cluster 3 organ developing leaf and cluster 23 organs Mesophyll cells. In the last phase, namely the formation of individual maize, three biclusters with the highest gene expression value of 4,112 in cluster 47. Another study using KNN classification also showed that the data mining technique is quite powerful to discriminate specific expression based on organ characteristics[8].

We extended our analysis to check whether the bicluster analysis is in concordance with the genes expression profile. So, we confirmed using the qTeller Tools to visualize specific gene expressions, such as *Zm00001d034773, Zm00001d018356, Zm00001d013367, Zm00001d027323* and *Zm00001d048585* (Figure 7-11).

Surprisingly, all genes showed organ-specific expression means that from the bicluster analysis, we can find a gene marker during maize growth stages. The bicluster with xMotif algorithm can be used for the future method to reduce the complexity of data analysis, especially for genes expression datasets.

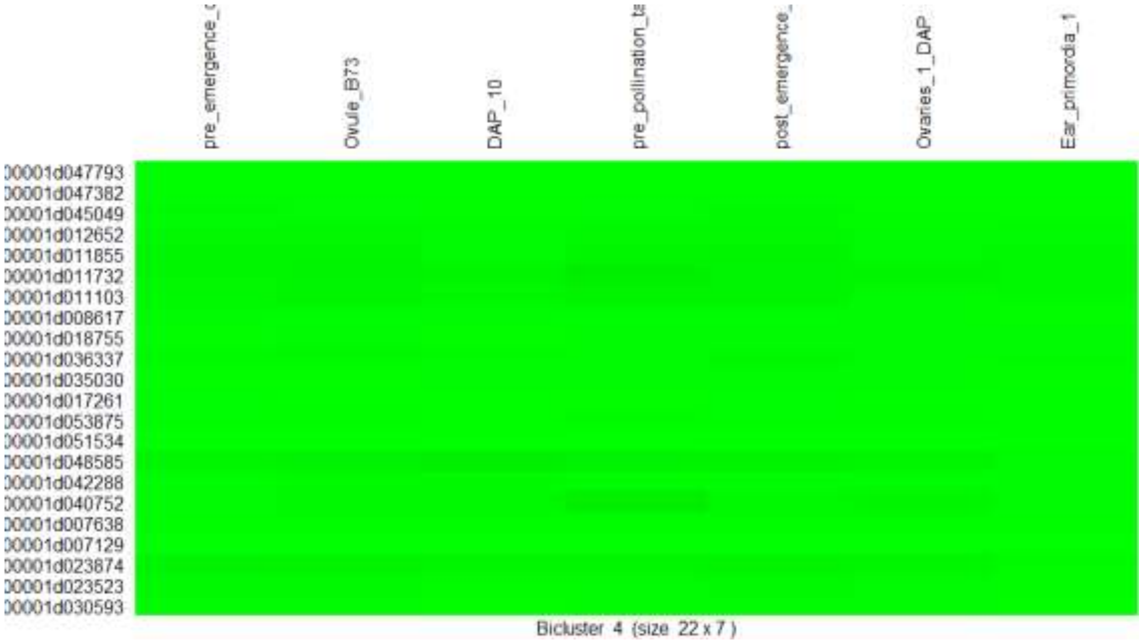

Figure 4: Bicluster 4 Visualization

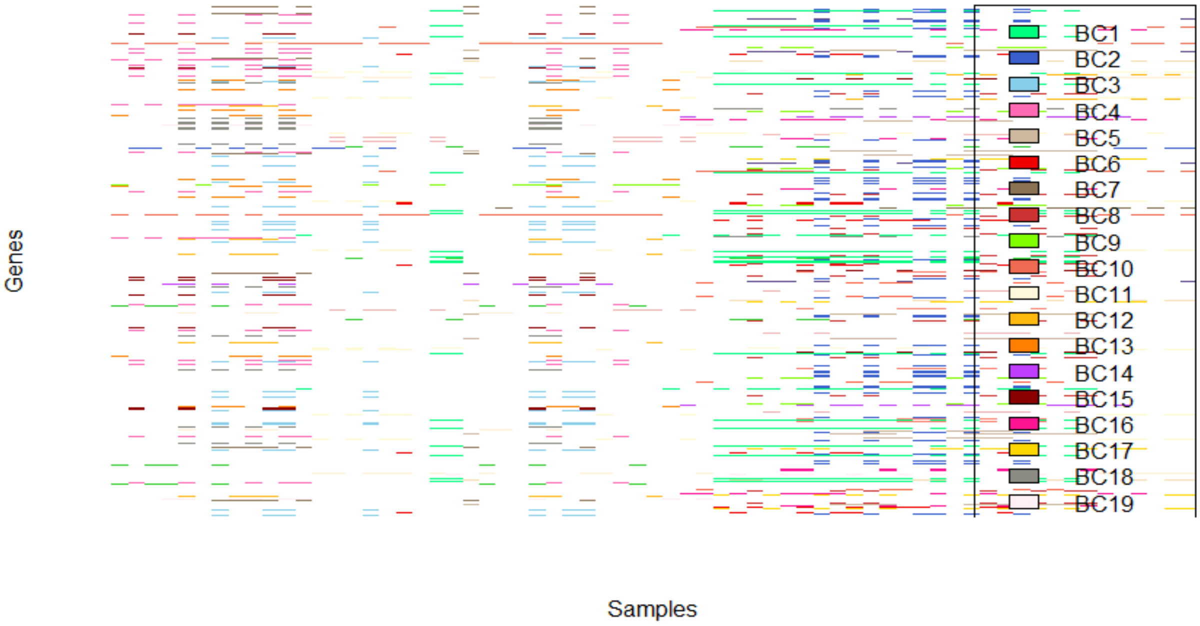

Figure 5: Bicluster Heatmap

Table 3
Bicluster Analysis and Mean of Genes Expression

Bicluster No	Maize Organ	Mean of Genes Expression	Maize Growth Stages
40	Whole_anthers	17.528	Reproductive Stage
41	twentyfive_DAP_endosperm	38.788	
31	twentyfive_DAP_endosperm	6.372	
21	twentyfive_DAP_endosperm	19.952	
46	Ovule_B73	11.494	
7	fiveDAP_Seed	23.334	
4	pre_pollination_tassel	30.392	
20	Bundle_sheath_cells	215.178	Vegetative Stage
16	Bundle_sheath_cells	0.707	
14	Bundle_sheath_cells	0.401	
11	Bundle_sheath_cells	32.142	
51	Developing_Leaf	8.846	
36	Developing_Leaf	12.628	
3	Developing_Leaf	200.534	
25	Shoot_field2_1	0.582	
50	Seedling_shoots	13.934	
18	Endosperm_14DAP	21.651	
15	Endosperm_14DAP	27.567	
12	Endosperm_14DAP	19.025	
30	Ovaries_1_DAP	12.478	
13	Ovaries_Drought_stressed	27.784	
53	Tassel_primordia_1	2.314	
23	Mesophyll_cells_400_1	114.522	
6	P6_lg1_b_1	0.343	
27	P6_lg1_b_1	1.231	
22	P6_Wt3_1	0.664	
37	P6_Wt2_1	40.538	
44	P6_Wt2_1	0.268	
28	P6_Wt2_1	97.062	
49	P6_Wt1_1	1.264	
5	P7_blade3_1	0.100	
1	P7_blade2_1	220.968	
52	P7_blade1_1	0.728	
33	P7_blade1_1	2.423	
8	P7_blade_L12_1	0.108	
38	P7_ligule3_1	0.831	
43	P7_ligule3_1	5.289	
17	P7_ligule3_1	1.801	
35	P7_ligule2_1	6.735	
26	P7_ligule2_1	0.503	
19	P7_ligule1_1	0.539	
32	P7_ligule_L13_1	0.416	
24	P7_ligule_L13_1	0.553	
48	P7_ligule_L12_1	1.176	
10	P7_ligule_L12_1	0.100	
45	P7_ligule_L12_1	25.648	
9	P7_sheath2_1	8.424	
2	P7_sheath2_1	14.833	
42	P7_sheath_L13_1	1.291	
39	P7_sheath_L13_1	0.111	
47	Maize_b1_control	4.112	
34	Maize_b1_control	0.001	
29	Maize_b1_control	0.020	

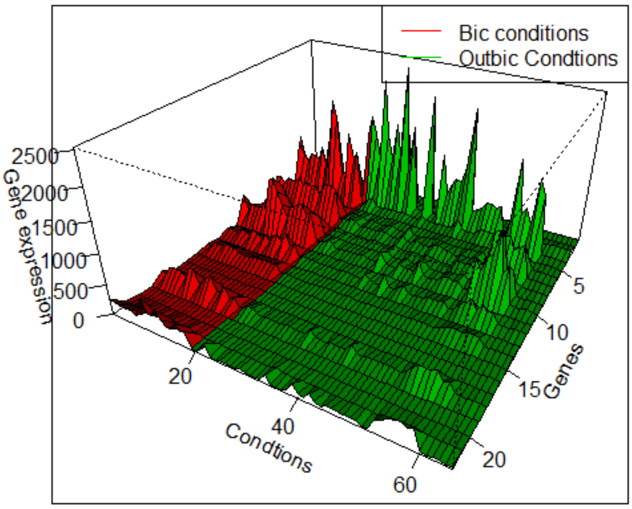

Figure 6: 3D Plot from Matrix Data

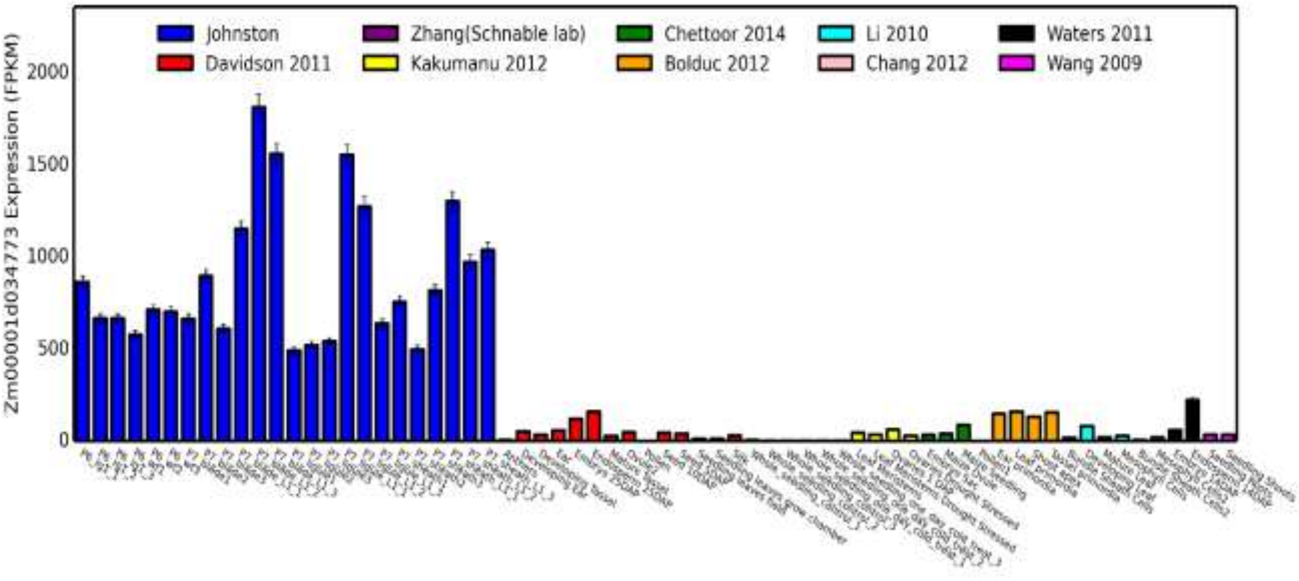

Figure 7: Gene expression of *Zm00001d034773* coming from different RNA-Sequencing libraries[9-14]. The barplot was generated automatically from the qTeller database (http://qteller.com/qteller4/generate_figures.php).

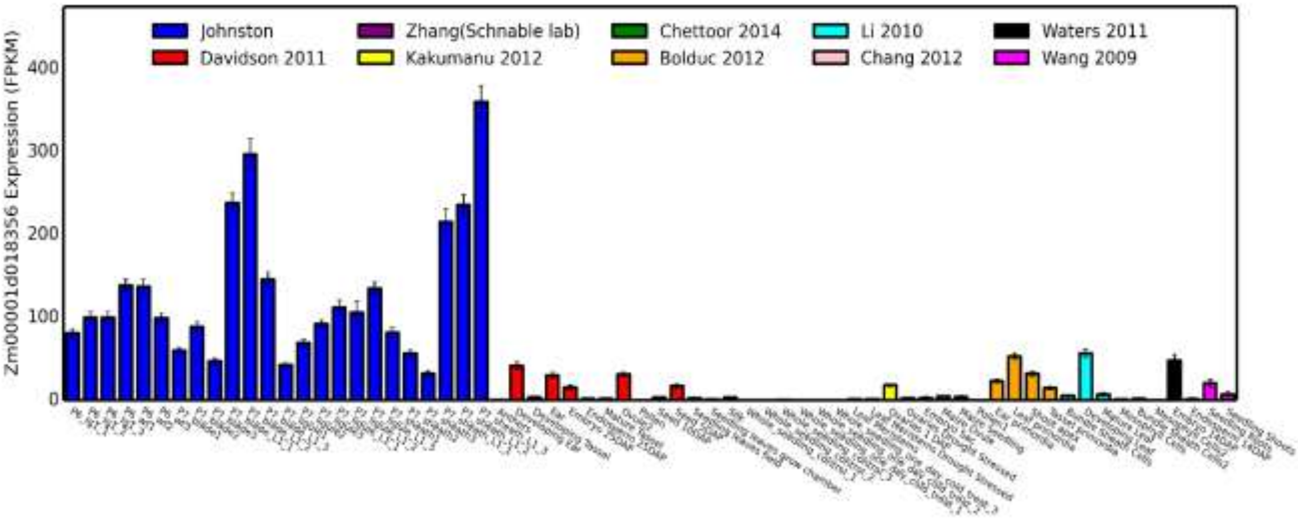

Figure 8: Gene expression of *Zm00001d018356* coming from different RNA-Sequencing libraries[9-14]. The barplot was generated automatically from the qTeller database (http://qteller.com/qteller4/generate_figures.php).

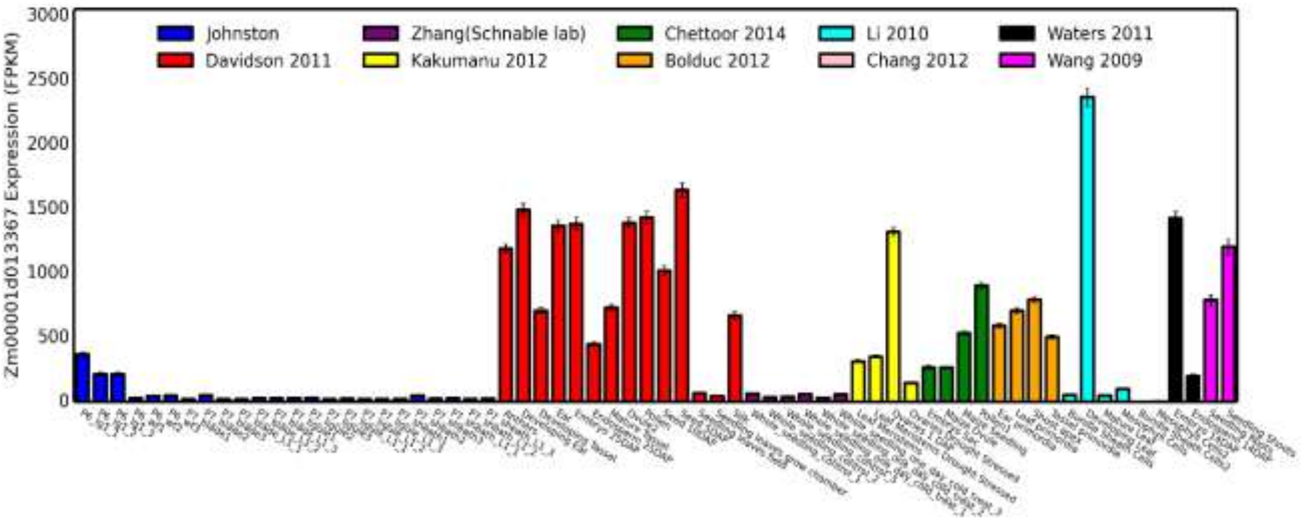

Figure 9: Gene expression of *Zm00001d013367* coming from different RNA-Sequencing libraries[9-14]. The barplot was generated automatically from the qTeller database (http://qteller.com/qteller4/generate_figures.php).

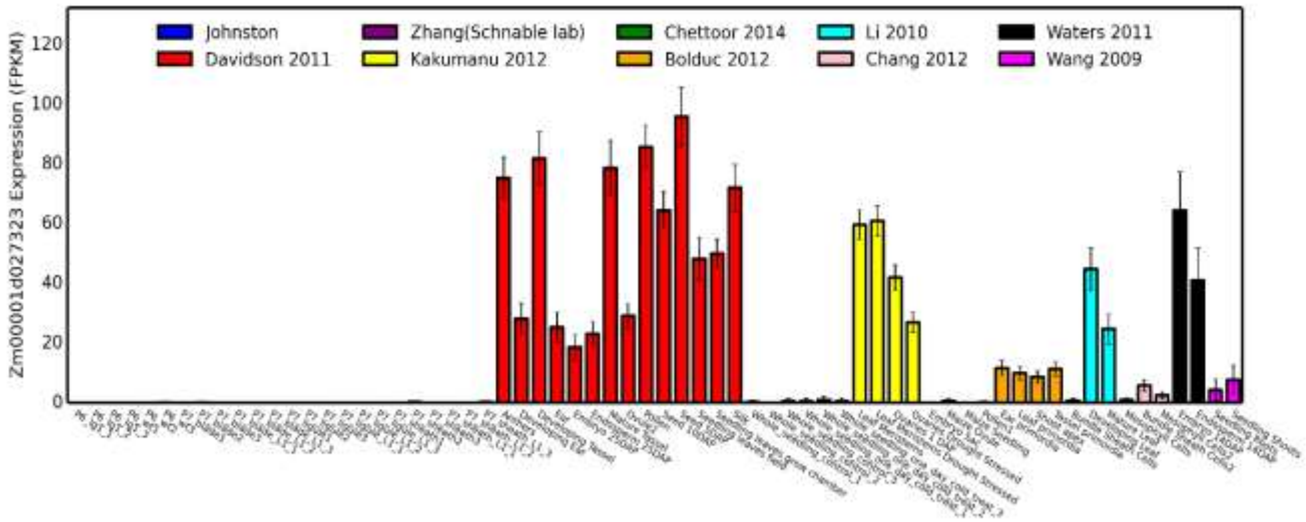

Figure 10: Gene expression of *Zm00001d027323* coming from different RNA-Sequencing libraries[9-14]. The barplot was generated automatically from the qTeller database (http://qteller.com/qteller4/generate_figures.php).

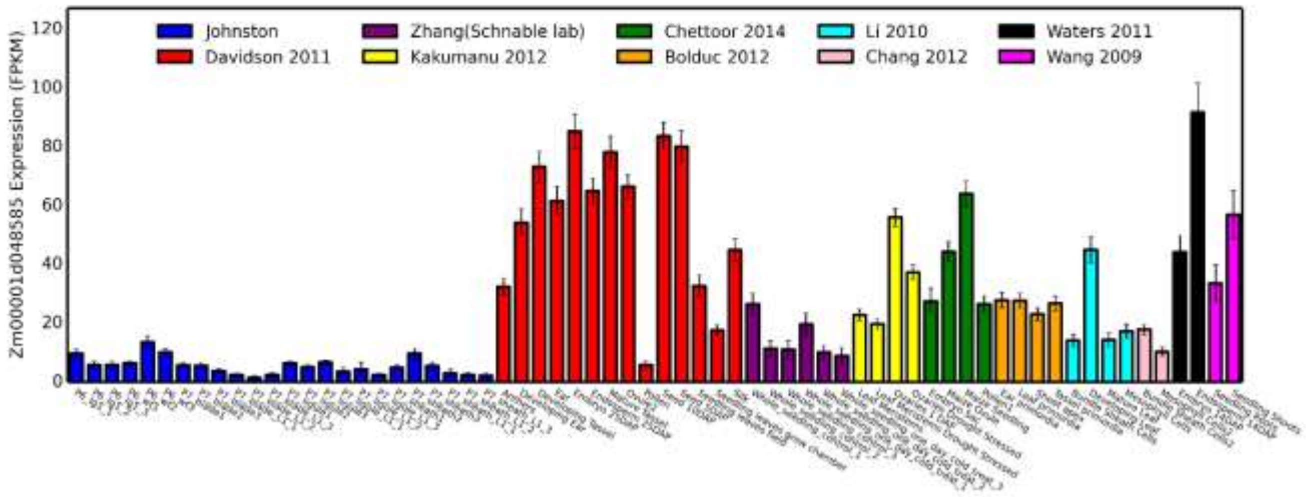

Figure 11: Gene expression of *Zm00001d048585* from different RNA-Sequencing libraries[9-14]. The barplot was generated automatically from the qTeller database (http://qteller.com/qteller4/generate_figures.php)

Conclusion

We performed bicluster analysis using genes expression datasets from qTeller containing 37,479x68 matrix data. Finally, we found 53 biclusters using the xMotif algorithm. Furthermore, we highlight four significant clusters based on the highest rank of statistical p-value to emphasize the gene function during maize growth stages. Genes play a role in controlling patterns of growth and development in plants.

The gene expression value influences the growth of maize. Here, we confirmed the results from different biclusters concordant with the gene expression profile generated from the qTeller database. The xMotif Bicluster algorithm can be used for future work in gene expression data analysis.

Acknowledgement

We thank Dr Rohmatul Fajriyah for the fruitful comments and discussion. Authors thank Riset Unggulan Pusat Penelitian Biosains dan Bioteknologi, Institut Teknologi Bandung.

References

1. Lawrence C.J., Seigfried T.E. and Brendel V., The Maize Genetics and Genomics Database, The community resource for access to diverse maize data, *Plant Physiol*, **138**, 55–58 **(2005)**

2. Schnable, qTeller, retrieved April 14th, 2018, from (www.qteller.com/qteller4) **(2018)**

3. Jakt L.M., Cao L., Cheah K.S.E. andSmith D.K., Assessing clusters and motifs from gene expression data, *Genome Res.*, **11**, 112–123 **(2001)**

4. Kusrini and Emha L.T., Algoritma Data Mining, Andi, Yogyakarta **(2009)**

5. Murali T.M. and Kasif S., Extracting Conserved Gene Expression Motifs from Gene Expression Data, Proceedings of the 8th Pacific Symposium on Biocomputing, 77–88 **(2003)**

6. R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/, 1 **(2013)**

7. Pollard K.S., Dudoit S. and Van Der Laan M.J., Multiple Testing Procedures: R multtest Package and Applications to Genomics, in Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Springer, 249-271 **(2005)**

8. Widiawati I.F., Nugrahapraja H. and Fajriyah R., K-Nearest Neighbor (KNN) Analysis on Genes Expression Datasets of Maize Nested Association Mapping (NAM) Showed Confident Classification on Organ-specific Expression, 2018 1st International Conference on Bioinformatics, Biotechnology andBiomedical Engineering - Bioinformatics and Biomedical Engineering, IEEE, 1-3 **(2018)**

9. Waters A.J. et al, Parent-of-Origin Effects on Gene Expression and DNA Methylation in the Maize Endosperm, *The Plant Cell*, **23(12)**, 4221-4233 **(2011)**

10. Davidson R.M. et al, Utility of RNA Sequencing for Analysis of Maize Reproductive Transcriptomes, *The Plant Genome*, **4**, 191–203 **(2011)**

11. Li P. et al, The developmental dynamics of the maize leaf transcriptome, *Nature Genetics*, **42**, 1060-1067 **(2010)**

12. Wang X. et al, Genome-Wide and Organ-Specific Landscapes of Epigenetic Modifications and Their Relationships to mRNA and Small RNA Transcriptomes in Maize, *Plant Cell*, **21**, 1053–1069 **(2009)**

13. Jia Y. et al, Loss of RNA Dependent RNA Polymerase 2 (RDR2) Function Causes Widespread and Unexpected Changes in the Expression of Transposons, Genes and24-nt small RNAs, *PLoS Genet*, **5(11)**, e1000737 **(2009)**

14. The Maize Gametophyte Project: Unpublished Dataset SRP006965, NCBI Bioproject, 1 **(2011)**

(Received 15th October 2020, accepted 20th December 2020)