Life Science Informatics Publications

Research Journal of Life Sciences, Bioinformatics, Pharmaceutical and Chemical Sciences

Journal Home page http://www.rjlbpcs.com/

RJLBPCS

ISSN 2454-6348

Original Research Article **DOI: 10.26479/2019.0501.25**

IDENTIFYING GENES RESPONSIBLE FOR SHOOT DEVELOPMENT IN *ARABIDOPSIS THALIANA* USING MICROARRAY TECHNIQUE

Bhupendra Prasad[1], Sunny Karodia[2], Jitendra Malviya*[3]

1. Department of Microbiology, Career College Bhopal Govindpura BHEL Bhopal, Madhya Pradesh, India.

2. Department of Biological Science and Engineering, Maulana Azad National Institute of Technology (MANIT), Bhopal, Madhya Pradesh, India.

3. Department of Biotechnology and Bioinformatics Centre, Barkatullah University, Bhopal, Madhya Pradesh, India

ABSTRACT: Gene expression analysis of *Arabidopsis thaliana* is becoming more and more important in many areas of biomedical research. cDNA microarray technology is one very promising approach for high throughput analysis and provides the opportunity to study gene expression patterns on a genomic scale. Microscope slide were determined by measuring the fluorescence intensity of labeled mRNA hybridized to the arrays. The Unscrambler and Genesis tools have been used to simultaneously visualize and analyze a whole set of gene expression experiments. Several graphical data show matrix of genes and then compared with each other. Fluorescence ratios have been normalized and best possible representation of the data is used for statistical analysis. Non hierarchical algorithms have been implemented to identify similar expressed genes and expression patterns, including: k-means clustering and principal component analysis. Finally the gene expression data was analyzed and the cluster responsible for the Shoot development in *Arabidopsis thaliana* is determined.

KEYWORDS: *Arabidopsis thaliana,* Genesis, Unscrambler, microarrays, cluster analysis, principal component analysis, genomics, bioinformatics, shoot development.

Corresponding Author: Jitendra Malviya* Ph.D.

Department of Biotechnology and Bioinformatics Centre, Barkatullah University, Bhopal, Madhya Pradesh, India. Email Address:jitmalviya123@gmail.com

1. INTRODUCTION

Microarray is a type of gene expression profiling. Gene expression is of three types mainly Global gene expression; expression pattern of all the genes present in the genome at the same time. Co-expression of genes; expression pattern of similar genes present in the genome at the same time, respectively differential expression of genes; expression pattern of dissimilar genes present in the genome at the same time.[10]

DNA Microarray

After genome sequencing, DNA microarray analysis has become the most widely used source of genome scale data in the life sciences. Microarray expression studies are producing massive quantities of gene expression and other functional genomics data, which promise to provide an insight into gene function and interactions within and across metabolic pathways. Unlike genome sequence data, however, which have standard formats for presentation and widely used tools and databases, much of the microarray data generated so far remain inaccessible.

Table 1: Microarray versus Microchips.

Types	Production	Substrate	Density (probes/cm^2)
High density array	Spotting of oligonucleotide or PCR fragments	Membranes	Up to 64
Microarray	Spotting of oligonucleotide or PCR fragments	Glass	Up to 10^4
Chip (Affymetrix Agilent)	Synthesis on substrate	Glass	Up to 2.5 *10^5

Microarray Technique

Microarray technology began about a quarter century ago, with Ed Southern's key insight that labeled nucleic acid molecules could be used to interrogate nucleic acid molecules attached to a solid support.[2] Today, thousands or even tens of thousands of genes can be spotted on a microscope slide and relative expression levels of each gene can be determined by measuring the fluorescence intensity of labeled mRNA hybridized to the arrays, facilitating the measurement of RNA levels for the complete set of transcripts of an organism. Applied to the functional genetics and mutation screening, microarrays give us the opportunity to determine thousands of expression values in hundreds of different conditions, allowing the contemplation of genetic processes on a whole genomic scale to determine genetic contributions to complex polygenic disorders and to screen for important changes in potential disease gene. cDNA microarrays exploit the preferential binding of complementary, single stranded nucleic acid sequences. Basically, a microarray is a specially coated glass microscope slide to which cDNA molecules are attached at fixed locations, called spots. [3, 4, 5] With up to date computer controlled high-speed robots 19200 and more spots can be printed on a single slide, each representing a single gene. RNA from control and sample cell is extracted. Fluorescently labeled cDNA probes are prepared by incorporating either cye-3 or cye-5 d'UTP using

a single round of reverse transcription, usually taking the red dye for RNA from the sample cells and the green dye for that from the control population. Both extracts are simultaneously incubated on the microarray, enabling the gene sequences to hybridize under stringed conditions to their complementary clones attached to the surface of the array. [6] Laser excitation of the incorporated targets yield an emission with characteristic spectra, which is measured using a scanning confocal laser microscope. Monochrome images from the scanner are then imported into the software in which they are pseudo-colored and merged. [7] A spot for instance will appear red, if the corresponding RNA from the sample population is in greater abundance and green if the control population is in greater abundance. If both are equal, the spot will appear yellow. If neither binds, the spot will appear black. Thus the relative gene expression levels of the sample and the reference populations can be estimated from the fluorescence intensities and colors emitted by each spot during scanning. The production and hybridization of slides is just one pace in a pipeline of many steps necessary to gain meaningful information from microarray experiments. Because of the vast amount of data produced by a microarray experiment, sophisticated software tools are used to normalize and analyze the data. [7]

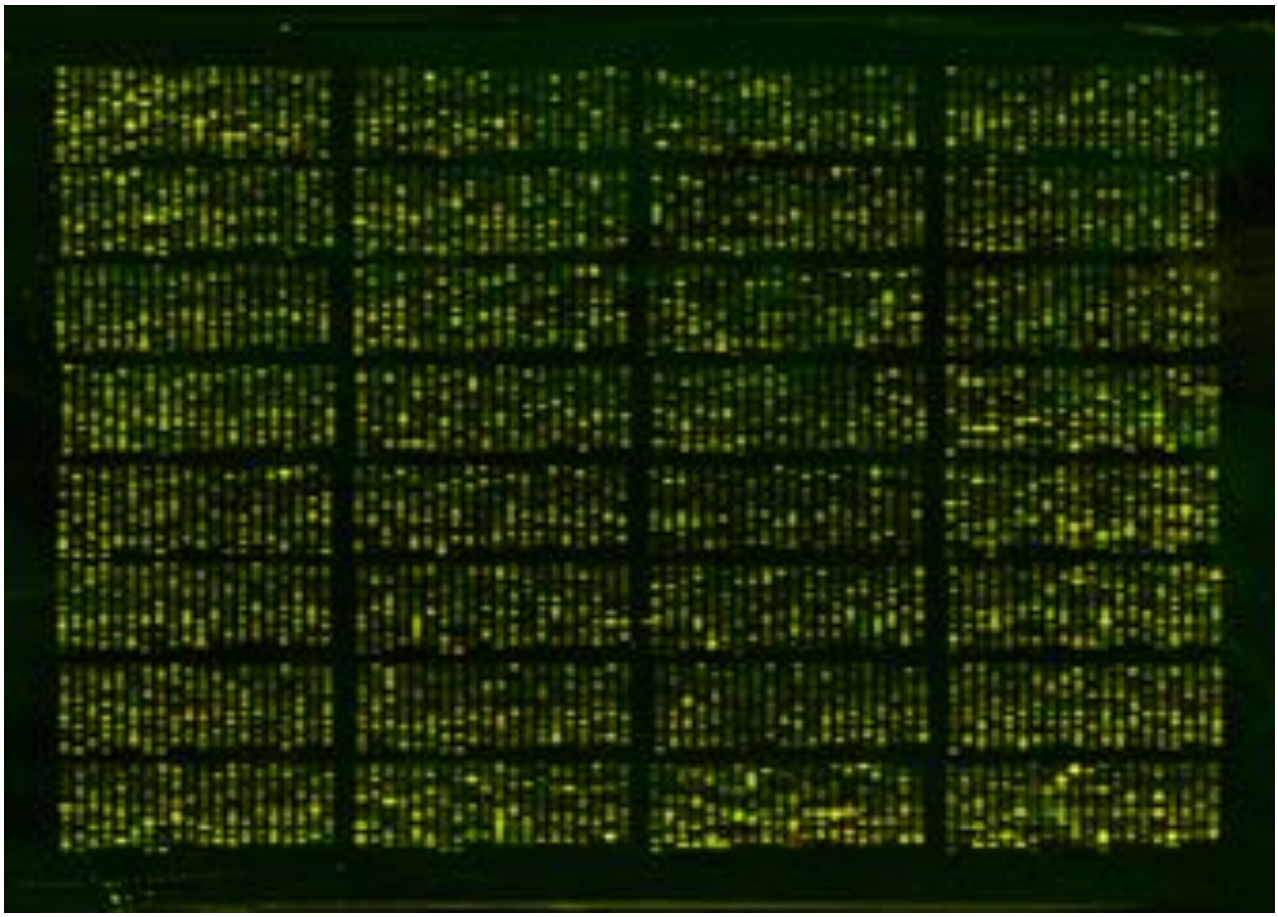

Figure 1.1 Profiling genes expressed during *Arabidopsis* shoot development in tissue culture

Figure 1.2 *Arabidopsis thaliana* –Model Organism

This little plant has become to plant biology what *Drosophila melanogaster* and *Caenorhabditis elegans* are to animal biology. [8] *Arabidopsis* is an angiosperm, a dicot from the mustard family (*Brassicaceae/Crusiferae*). [9, 3] It is popularly known as *thale cress* or *mouse-ear cress*. *Arabidopsis* is not an economically important plant such as turnip, cabbage, broccoli, and canola. Despite this, it has been the focus of intense genetic, biochemical and physiological study for over 40 years because of several traits that make it very desirable for laboratory study. [9]

Scientific classification

Kingdom	:	Plantae
(unranked)	:	Angiosperms
(unranked)	:	Eudicots
(unranked)	:	Rosids
Order	:	Brassicales
Family	:	Brassicaceae
Genus	:	*Arabidopsis*
Species	:	***A. thaliana***
Binomial name	:	*Arabidopsis thaliana* (L.) Heynh
Synonyms	:	Arabis thaliana

K-means Clustering

It's an unsupervised Clustering approach. [10, 11] It's a clustering algorithm which is widely used because of its simple implementation. The algorithm takes the number of cluster (k) to be calculated as an input. The number of clusters is usually chosen by the user. [12, 13]The k-means algorithm is one of the simplest and the fastest clustering algorithms. However, it has a major drawback. The results of the k-means algorithm may change in successive runs because the initial clusters are chosen randomly. As a result, the researcher has to assess the quality of the obtained clusters. [14,

13]The researchers may measure the size of the clusters against the disease of the nearest cluster. This may be done to all clusters. If the distance between the clusters is greater than the sizes of the clusters for all the clusters then the results may be considered as reliable. [15]

Principle Component Analysis

In experiments each gene and each experiment may represent one dimension. For example, a set of 10 experiments involving 20,000 genes may be conceptualized as 20,000 data points (genes) in a space with 10 dimensions (experiments) or 10 points (experiments) in a space with 20000 dimensions (genes). Both the situations are beyond the capabilities of current visualization tools and beyond the visualization capabilities of our brains. [16]

2. MATERIALS AND METHODS

SMD: The Stanford Microarray Database

The whole data has been downloaded from SMD (Stanford Microarray Database). http://genomewww5.stanford.edu/The goals for SMD are to serve as a storage site for microarray data from ongoing research at Stanford University, and to facilitate the public dissemination of that data once published, or released by the researcher.[2] SMD make use of many public resources to connect expression information to the relevant biology, including SGD and can be accessed at http://genomewww.stanford.edu/microarray.

Genesis

It is a platform independent Java suite, which integrates tools for analyzing gene expression data. [17] High throughput gene expression analysis is becoming more and more important in many areas of biomedical research. Genesis visualization of the gene expression and clustering results is user friendly. The flexibility, the variety of analysis and data visualization tools as well as the transparency and portability, provides Genesis software suite with the potential to become a valuable tool in functional genomics studies. [14]

The Unscrambler X

The software was originally developed in 1986 by HARALD MARTENS and later by CAMO software. The Unscrambler is a commercial software product for more than one data analysis, used primarily PCA and PCA projection was done with the help of Unscrambler X performed k-means Clustering with the help of Genesis. The software used for calibration in the application of near infrared spectroscopy and development of predictive models for use in real-time spectroscopic analysis of materials.

1. Data Collection from STANFORD MICROARRAY DATABASE
2. Removing gaps and errors from the SMD file
3. Converting the excel file into tab limited format
4. Performing k-means clustering in Genesis
5. Performing Principal Component Analysis in Unscrambler

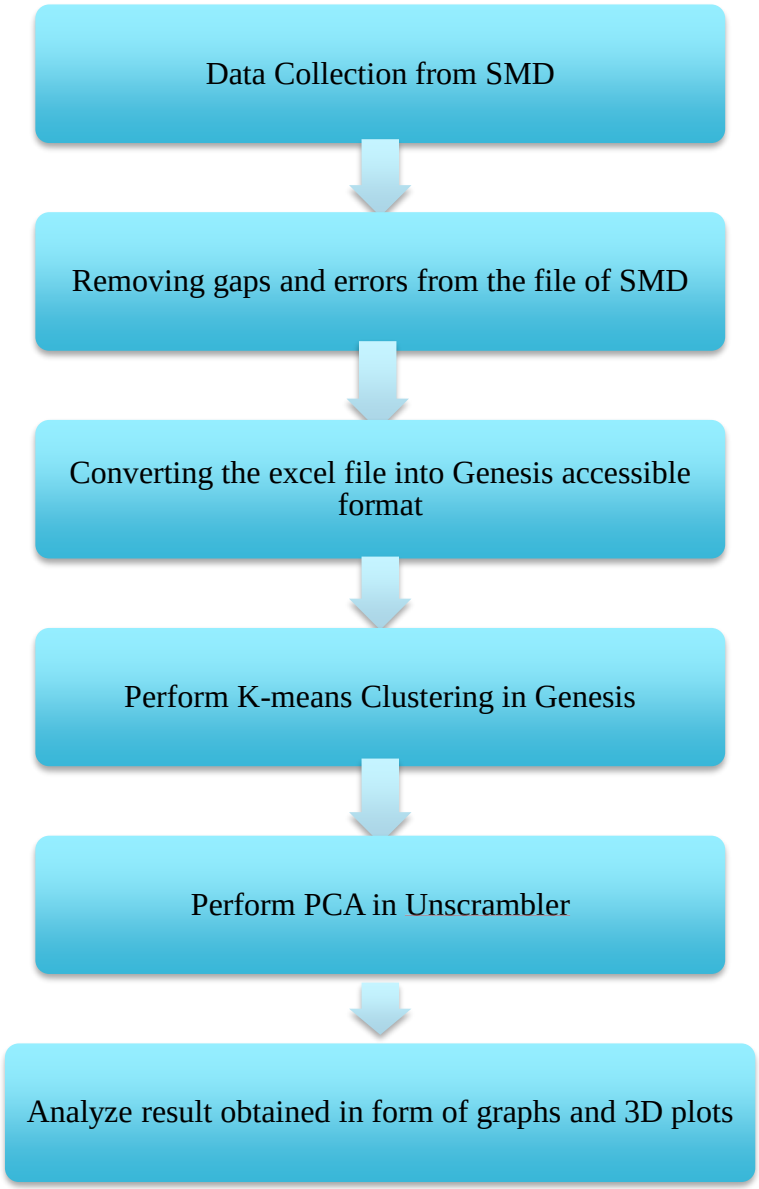

Figure 2.1 Methodology chart

3. RESULTS AND DISCUSSION

Genesis

In this work, 'Genesis' a versatile and transparent software suite for large-scale gene expression cluster analysis was used. The Genesis software was used to enable data import, visualization, data normalization, and clustering via: k-means and Self organizing maps. Also the Unscrambler software was used as an additional support for the work. It also enabled the data import, visualization and interpretation using Principal Component Analysis.[15]

Centroid View

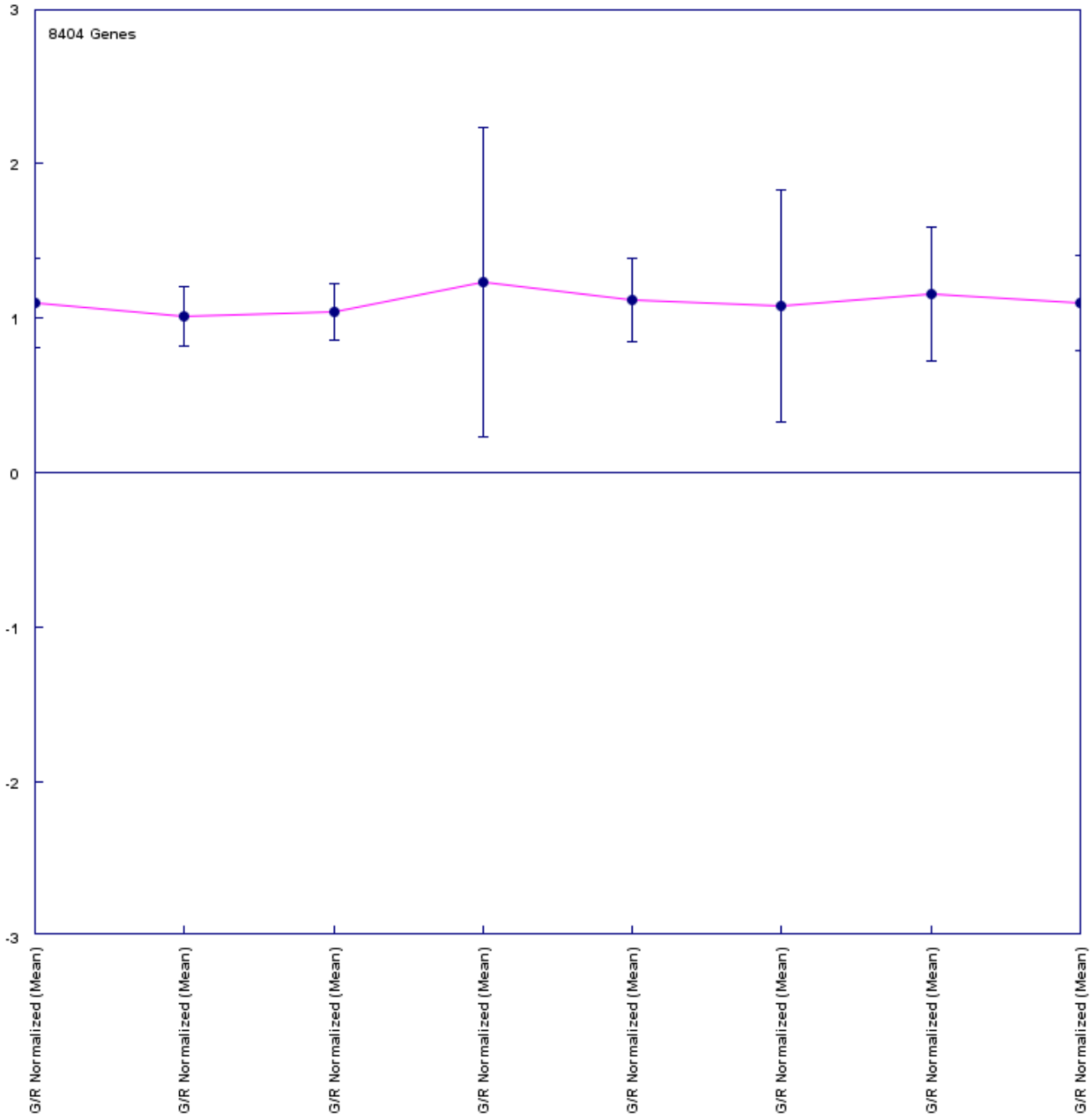

Figure 3.1 Centroid View of total data in Genesis

Here we can clearly see that the Centroid view of the fourth G/R Normalized (Mean) is having a clear higher projection of the expression. So the probability of the genes expressing the desired value is more in fourth data file.

Expression View

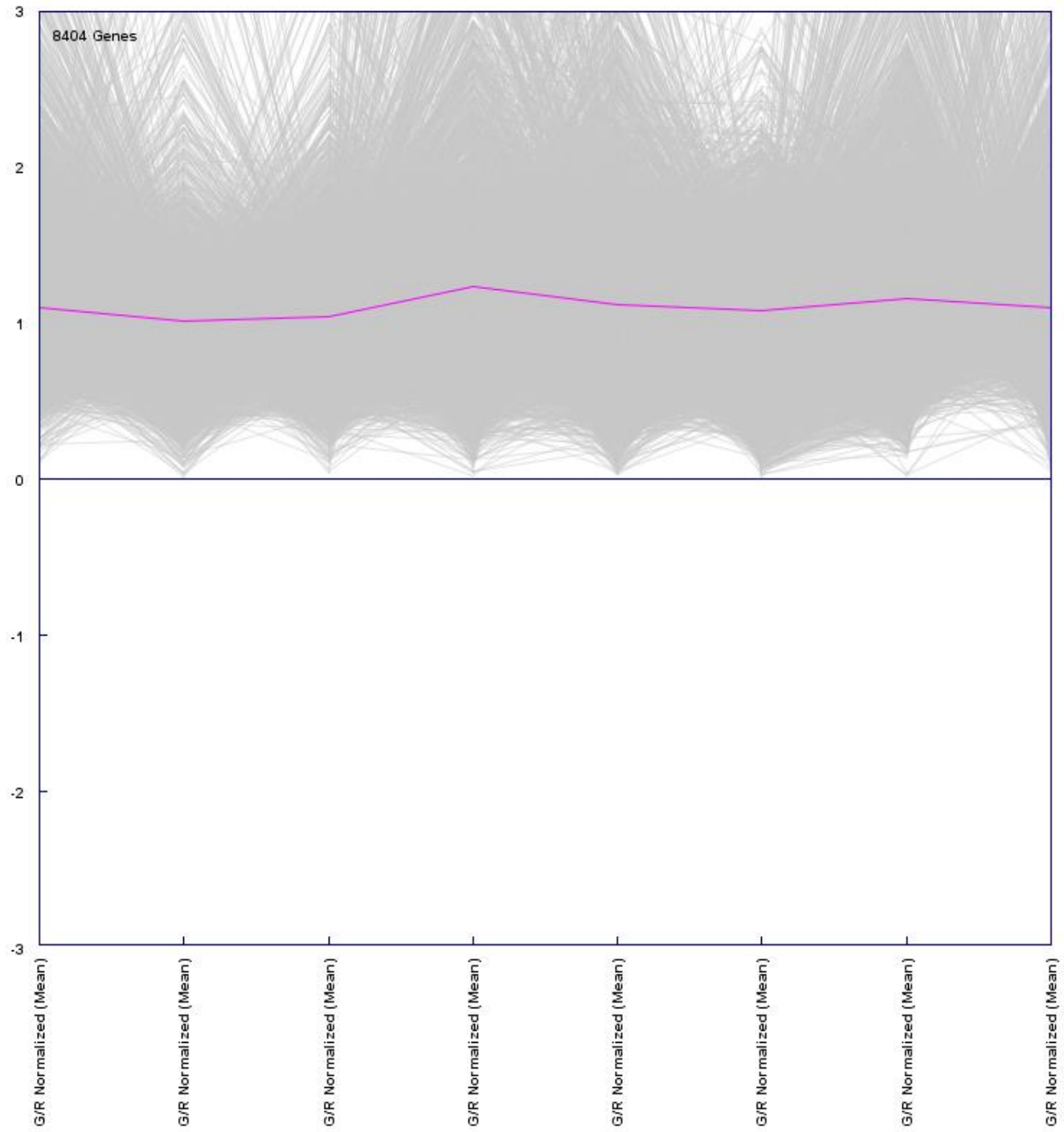

Figure 3.2 Expression View of total data projected in Genesis

Even the expression of the fourth G/R Normalized (Mean) data table is showing a peak in graph clearly. So this adds too for our prediction of presence of genes having desired property in the fourth data file.

Centroid View of Ten clusters

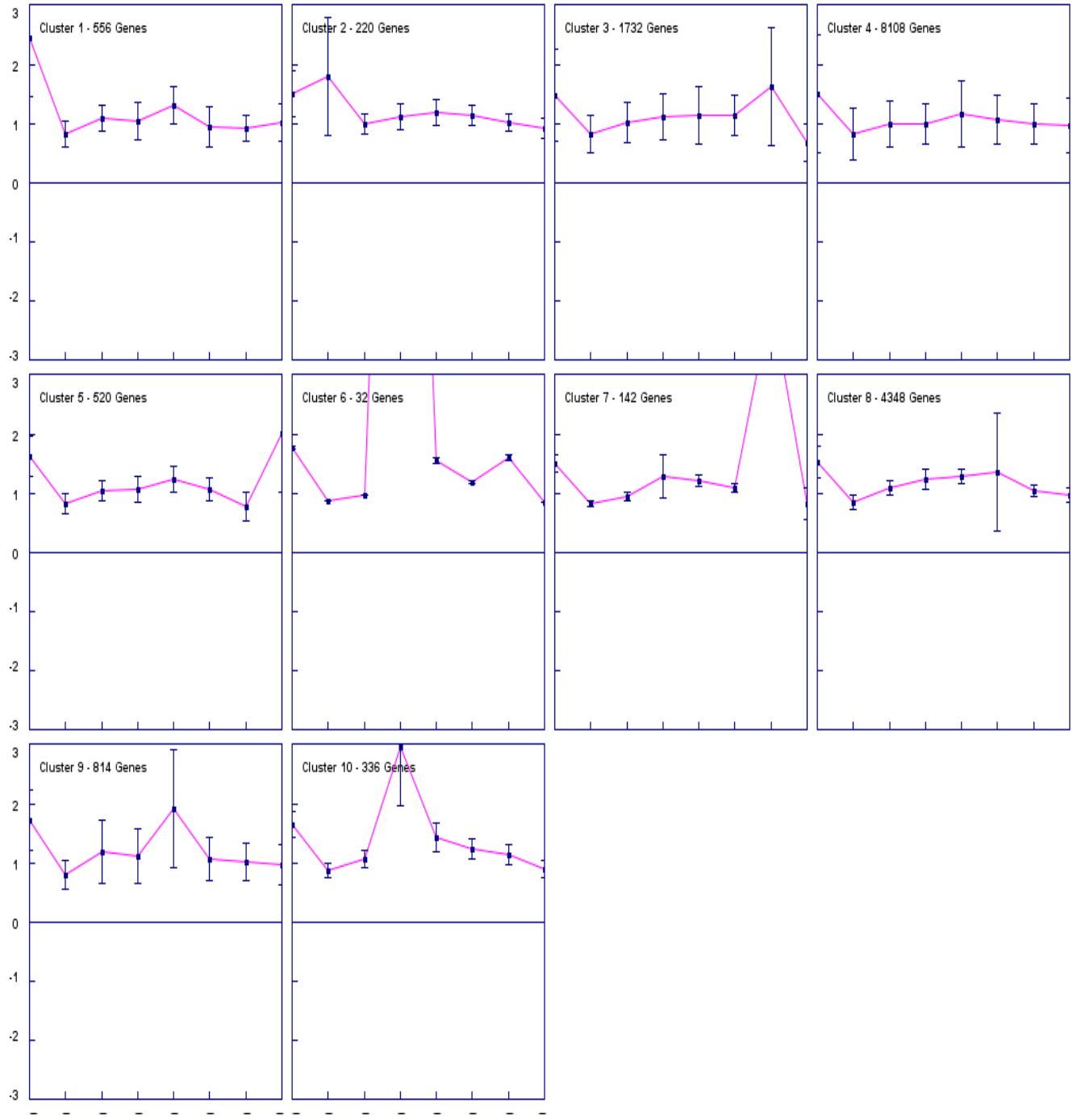

Figure 3.4 Centroid View of 10 clusters projected using Genesis through k-means clustering algorithm

After the k-means Clustering algorithm was applied on the data the Centroid View was obtained as shown above. Here we can see that the cluster 4 with 8108 genes is having a uniform peak values in its Centroid and the number of data is sufficiently large. Hence the genes of desired property can be present in the fourth cluster.[18]

Expression Views of Ten clusters

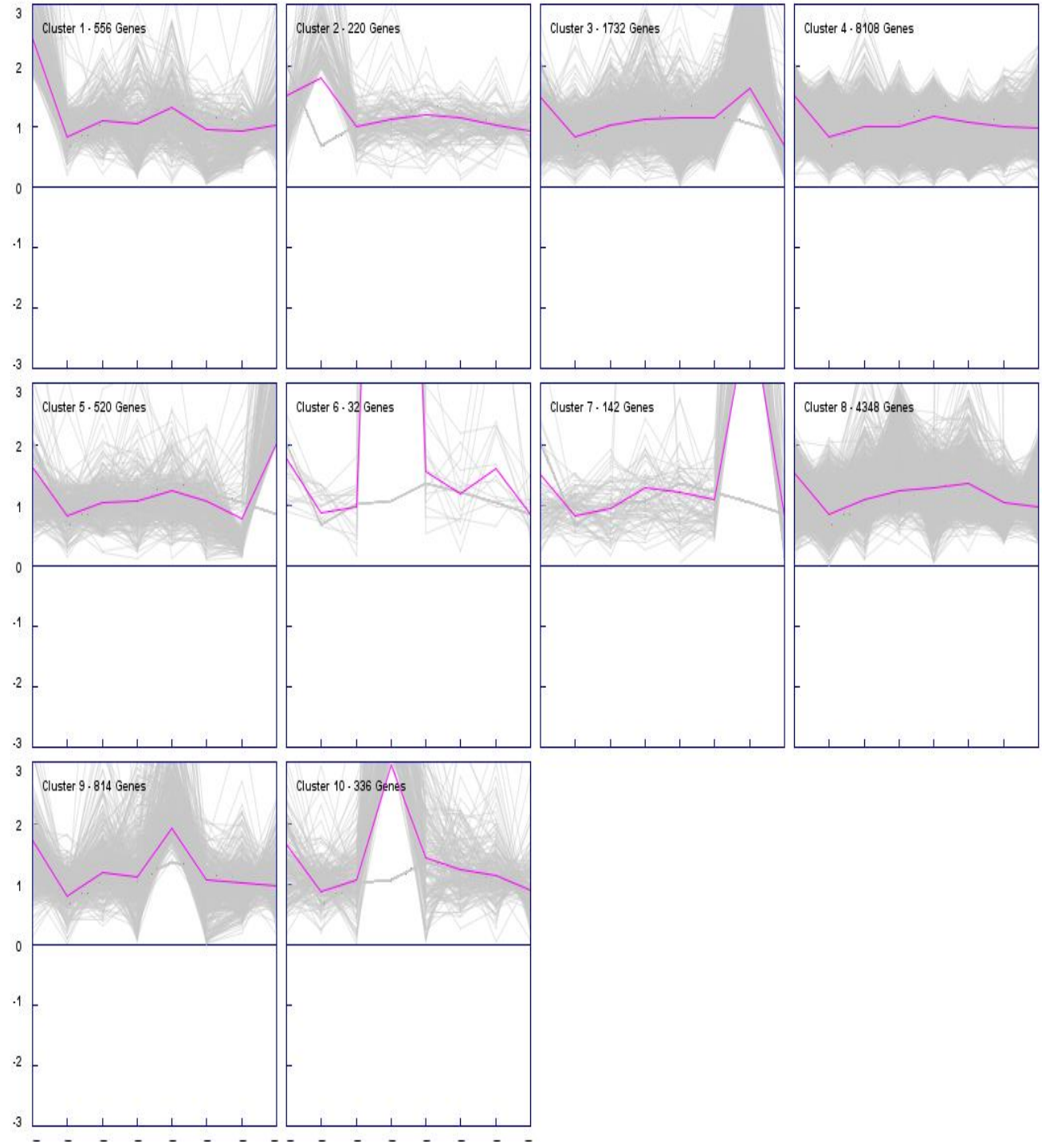

Figure 3.5 k- means clustering result using Euclidian Distance in ten clusters Expression plots of all genes and the median of a cluster.

After the k-means Clustering algorithm was applied on the data the Expression View was obtained as shown above. Here we can see that the cluster 4 with 8108 genes is having a uniform expression of the value and has no oscillating pattern. Hence this evidence too boosts the presence of genes of desired property in the fourth cluster. [16]

Expression Image of a Cluster

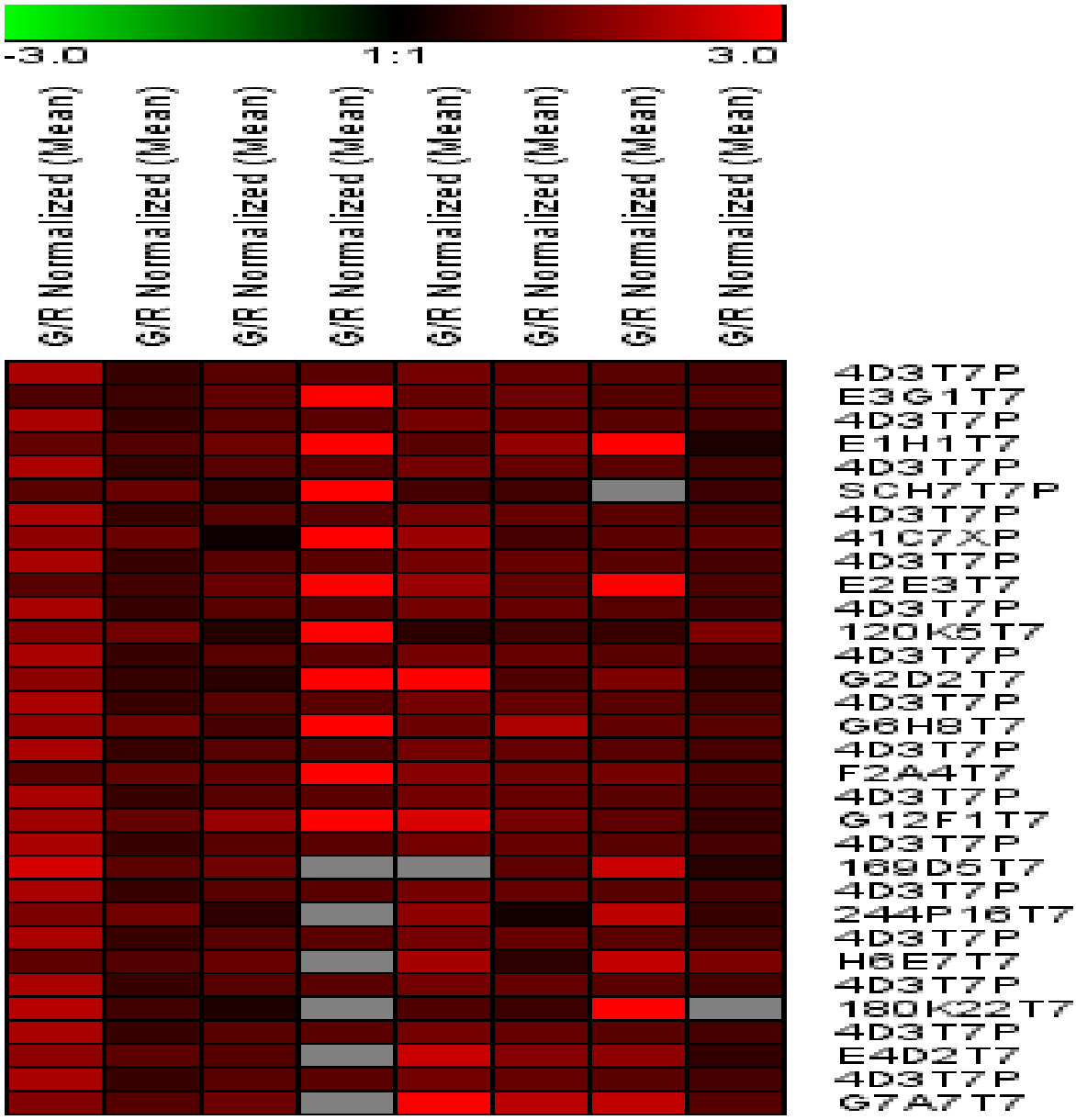

Figure 3.6 Expression image of the cluster. The color scale ranges from saturated green for log ratios-3.0and below to saturated red for log ratios3.0andabove. Eachgeneisre presented by a single row of colored boxes; each column represents a single G/R Normalized (Mean).

After the k-means Clustering algorithm was applied on the data the basic Cluster information was obtained as shown above. Here we can see that the cluster 4 with 8108 genes is having some important values as below: Share of 96% of Genes in Cluster. [19]. Average distance from cluster mean = 0.12918513 which is lowest compared to all the available clusters. Next nearest neighbor variance = 0.028671337 which shows that the genes are almost identical in their properties within cluster variance = 0.06459144 which is lowest among all the clusters. After the k-means Clustering algorithm was applied on the data the basic Gene information was obtained as shown above. Here we can clearly see that the larger value occurrences of NNR values of Cluster 4 are maximum

compared to all the other Clusters. Hence from the Genesis we concise our point of view to Cluster number 4. Now we will verify this result in Unscrambler by performing PCA and PCA projection.

Cluster Information

Table 3.1 The basic Cluster information of all 10 clusters obtained in Genesis

k-means Result

Cluster	Number of Genes in Cluster	Share of Genes in Cluster	Average distance from cluster mean	next nearest neighbor variance	within cluster variance
Cluster 1	556	7 %	0.7742823	0.036890898	1.9302913
Cluster 2	220	3 %	0.65044343	0.04769829	0.5784114
Cluster 3	1732	21 %	0.65642047	0.036462884	0.20110764
Cluster 4	8108	96 %	0.12918513	0.028671337	0.06459144
Cluster 5	520	6 %	0.36314118	0.04421689	1.203546
Cluster 6	32	0 %	0.60086673	0.08682262	688.2634
Cluster 7	142	2 %	0.74638313	0.060880214	10.346304
Cluster 8	4348	52 %	0.1557256	0.024325997	7.638902
Cluster 9	814	10 %	0.55867404	0.027451806	0.7584053
Cluster 10	336	4 %	0.63540477	0.042794477	1.8799174

Table 3.2 The Gene Information obtained after performing k-means Clustering in Genesis.

k-means Result

	UniqueID	Min	Mean	Max	SDev	CV	NNR	Cluster
7910	118M6XP	0.391	0.92	1.833	0.46	49.91 %	0.9969518	8
1811	E8A3T7	0.51	1.16	1.79	0.5	43.17 %	0.9969696	4
265	148N23T7	0.378	1.16	2.929	0.83	71.47 %	0.9970114	4
7795	177B21T7	0.706	1.01	1.272	0.23	22.99 %	0.9970121	4
8204	62A4XP	0.73	1.21	1.867	0.43	35.46 %	0.99705887	4
5174	121I14T7	0.941	1.18	1.724	0.23	19.82 %	0.9970778	4
3035	225L16T7A	0.543	1.1	1.502	0.3	27.4 %	0.9972767	4
2528	82F12T7	0.902	1.15	1.361	0.17	14.51 %	0.99729186	4
4058	164D12T7	0.807	1.1	1.498	0.23	20.63 %	0.99743366	4
3845	123B20T7	0.729	1.17	1.95	0.37	31.65 %	0.9975406	2
3890	E4H9T7	0.88	1.09	1.404	0.19	17.07 %	0.99763507	4
867	66D8XP	0.898	15.2	113.0	39.52	259.9 %	0.99764115	8
1558	119H4T7	0.641	1.01	1.593	0.32	31.85 %	0.9976971	4
3696	184A2T7	0.732	1.3	2.104	0.5	38.17 %	0.9979429	1
68	121K12XP	0.679	1.16	1.484	0.24	20.93 %	0.99796665	4
4352	290B4T7	0.676	1.25	1.798	0.4	32.27 %	0.998036	3
250	SCD3T7P	0.275	0.45	0.71	0.15	34.06 %	0.9980426	3
940	103J5T7	0.741	0.96	1.693	0.35	36.45 %	0.9982047	4
2801	104F1T7	0.132	0.97	1.817	0.56	57.39 %	0.9982553	1
7605	94E7T7	0.766	1.03	1.401	0.26	24.72 %	0.9982568	4
1582	145P7T7	0.779	1.2	1.711	0.27	22.14 %	0.99833816	4
3056	42E10T7	0.801	1.1	1.565	0.26	23.49 %	0.9984317	8
7285	244K9T7	0.782	1.05	1.318	0.21	20.23 %	0.9985792	8
905	225G22T7	0.521	0.86	1.432	0.3	34.87 %	0.99867684	4
2144	ATERD2	0.685	1.05	1.568	0.27	25.37 %	0.9986985	4
455	241I13T7	0.822	1.15	1.77	0.32	27.57 %	0.99872607	4
2526	38B1T7	0.631	1.22	1.846	0.4	32.96 %	0.99877787	4
5480	315B11T7	0.682	1.09	1.326	0.21	19.21 %	0.99879134	4
4132	178P23T7	0.359	1.19	1.901	0.49	41.04 %	0.99883705	8
5217	H6F10T7	0.743	1.57	2.447	0.67	42.71 %	0.99888873	8
4314	108I3T7	0.586	1.1	1.787	0.44	39.93 %	0.9989457	8
6494	156F1T7	0.859	1.05	1.206	0.14	13.84 %	0.999043	4
6809	88I23T7	0.524	1.12	1.81	0.4	35.77 %	0.99919534	8
5206	240K17T7	0.827	1.58	2.429	0.61	38.38 %	0.9995301	8
2705	206H8T7	0.826	1.06	1.401	0.2	18.56 %	0.9998443	4
5492	180K22T7	0.354	1.13	4.068	1.23	108.93 %	0.999981	6

Unscrambler

Line Plot

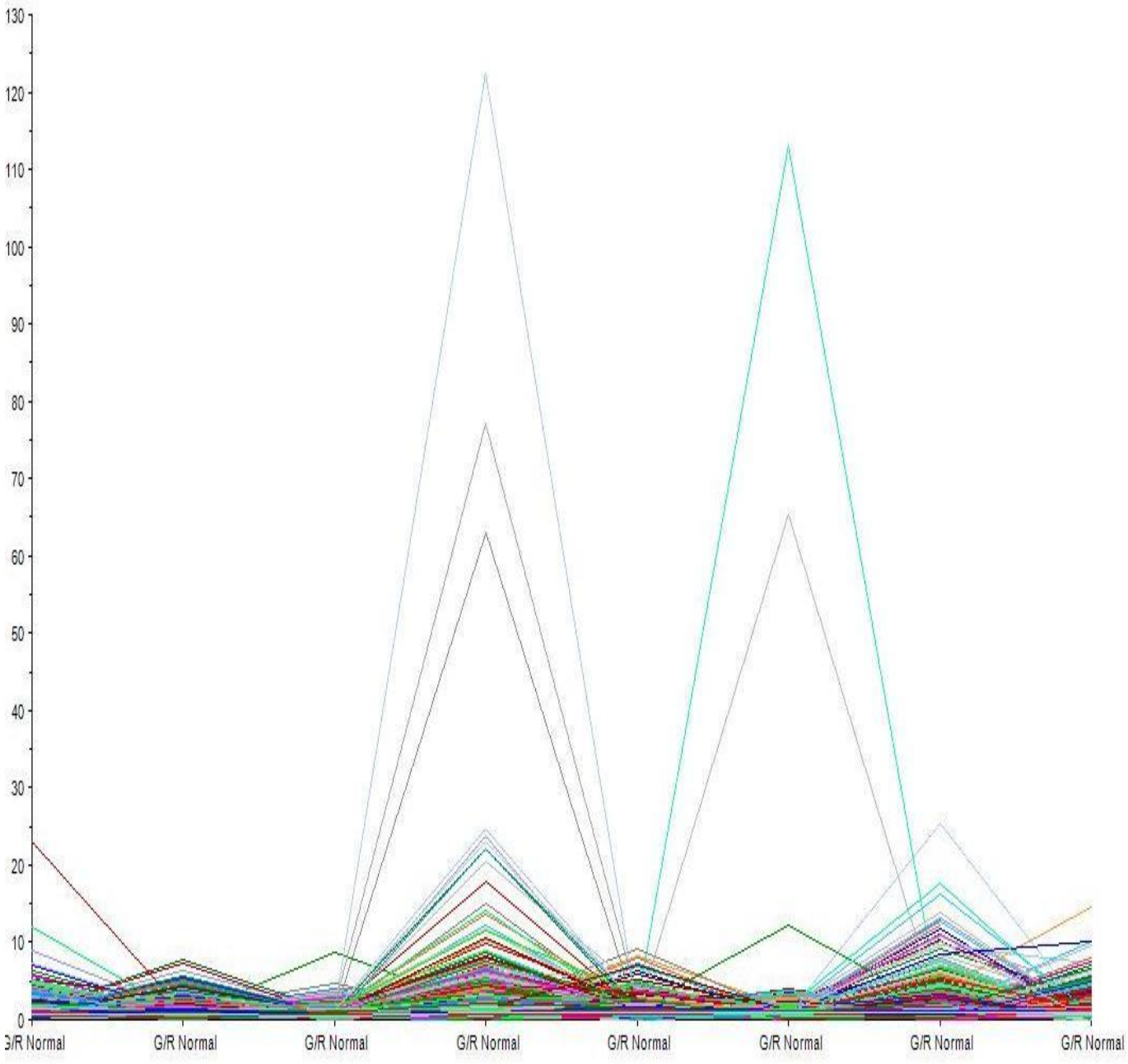

Figure 3.7 The Line Plot projected in Unscrambler

After the saved file form Genesis was imported in Unscrambler we projected the line plot of the data. Here we can clearly see that the expression of the fourth G/R Normalized (Mean) is having a clear higher projection of the expression. So the probability of the genes expressing the desired value is more in fourth data file. [7]

Bar Plot

After the Cluster 4 was analyzed in the form of Bar graph we can clearly see that the expression of the Fourth Cluster is having lesser number of noises.

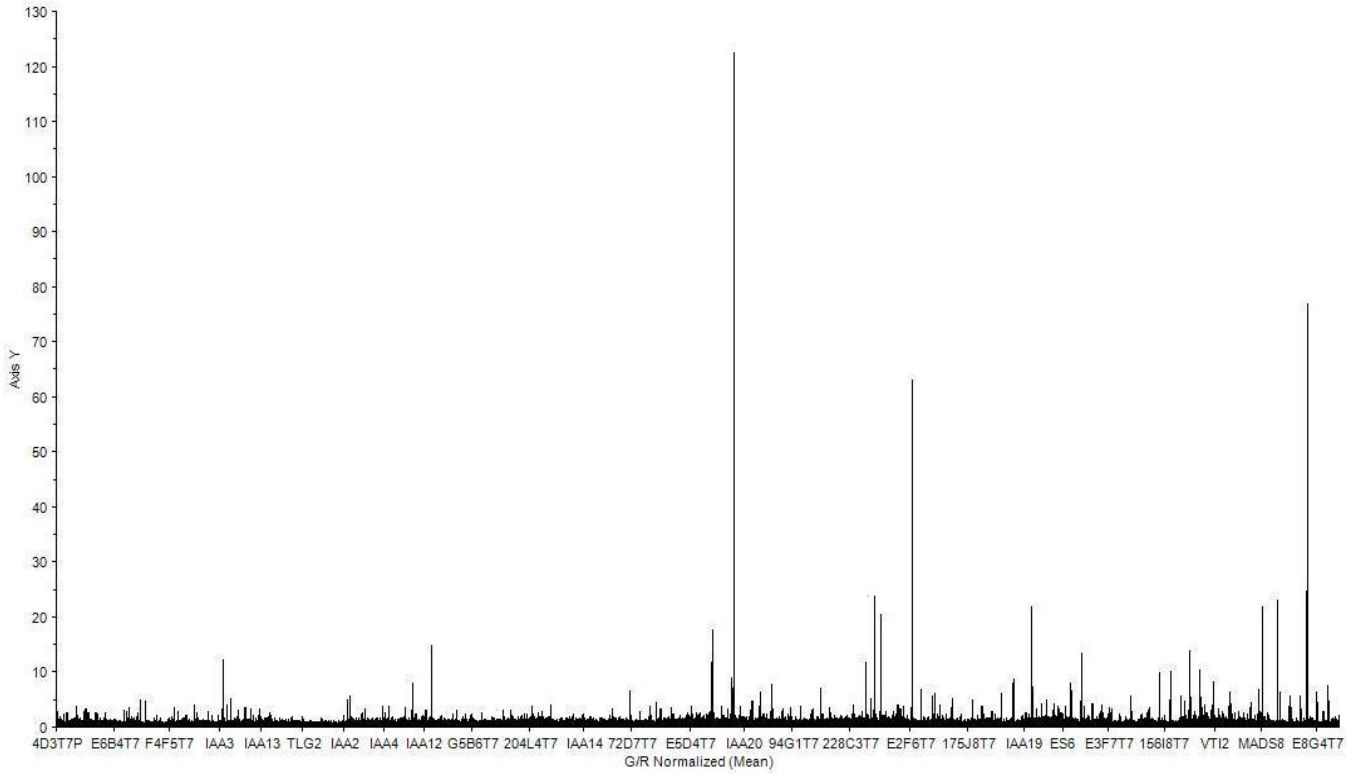

Figure 3.8The Bar Plot projected in Unscrambler of Cluster 4 of Genesis

3D-Scatter Plot

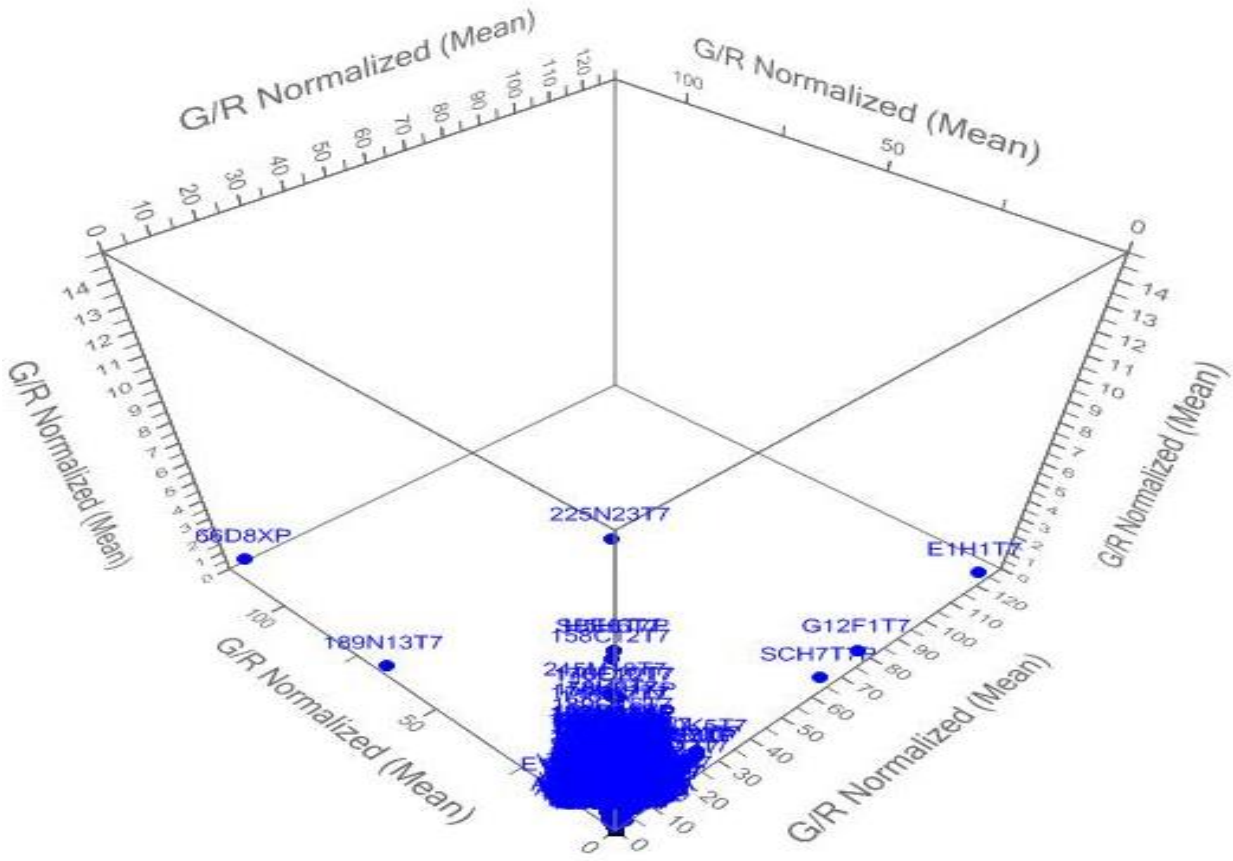

Figure 3.93D Scatter Plot obtained by Unscrambler

For more precise observation we did a 3D projection of the 4th, 6th and 7th G/R Normalized (Mean) data. And what we observed was a uniform cluster with less amount of variation in data.

Scatter Plot

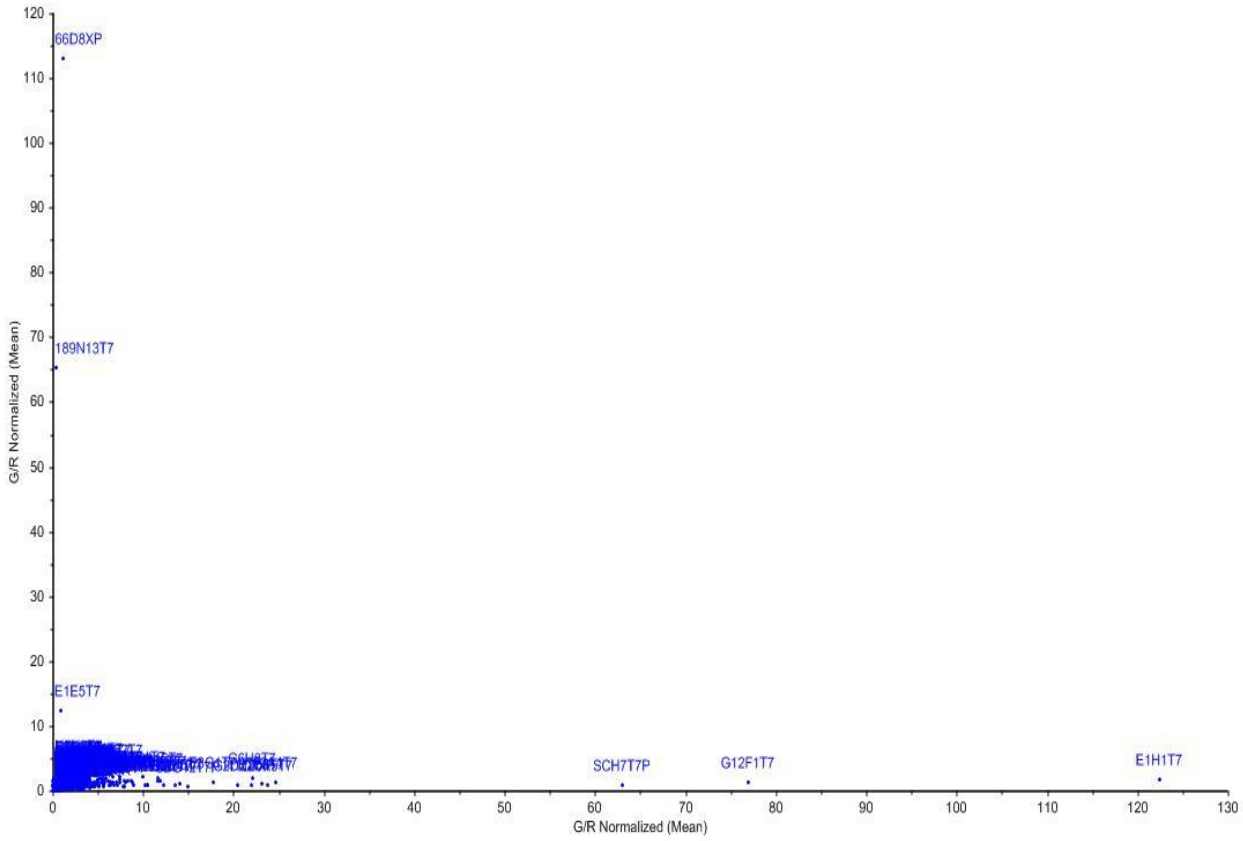

Figure 3.10 Scatter Plot obtained by Unscrambler

For more precise observation we plotted the single dimension of the 4th and 6thG/R Normalized (Mean), and what we observed was a uniform cluster of the data in initial stage of graph.

Table 3.3 Eigen values and Eigen value distribution

Principal Component	Value	Variation Explained	Cumulative Variation Explained
Principal Component 1	00.156	22.293 %	22.293 %
Principal Component 2	00.136	19.476 %	41.768 %
Principal Component 3	00.113	16.182 %	57.950 %
Principal Component 4	00.089	12.672 %	70.622 %
Principal Component 5	00.077	10.990 %	81.612 %
Principal Component 6	00.070	10.071 %	91.684 %
Principal Component 7	00.058	08.316 %	100.000 %
Principal Component 8	00.000	00.000 %	100.000 %

PCA

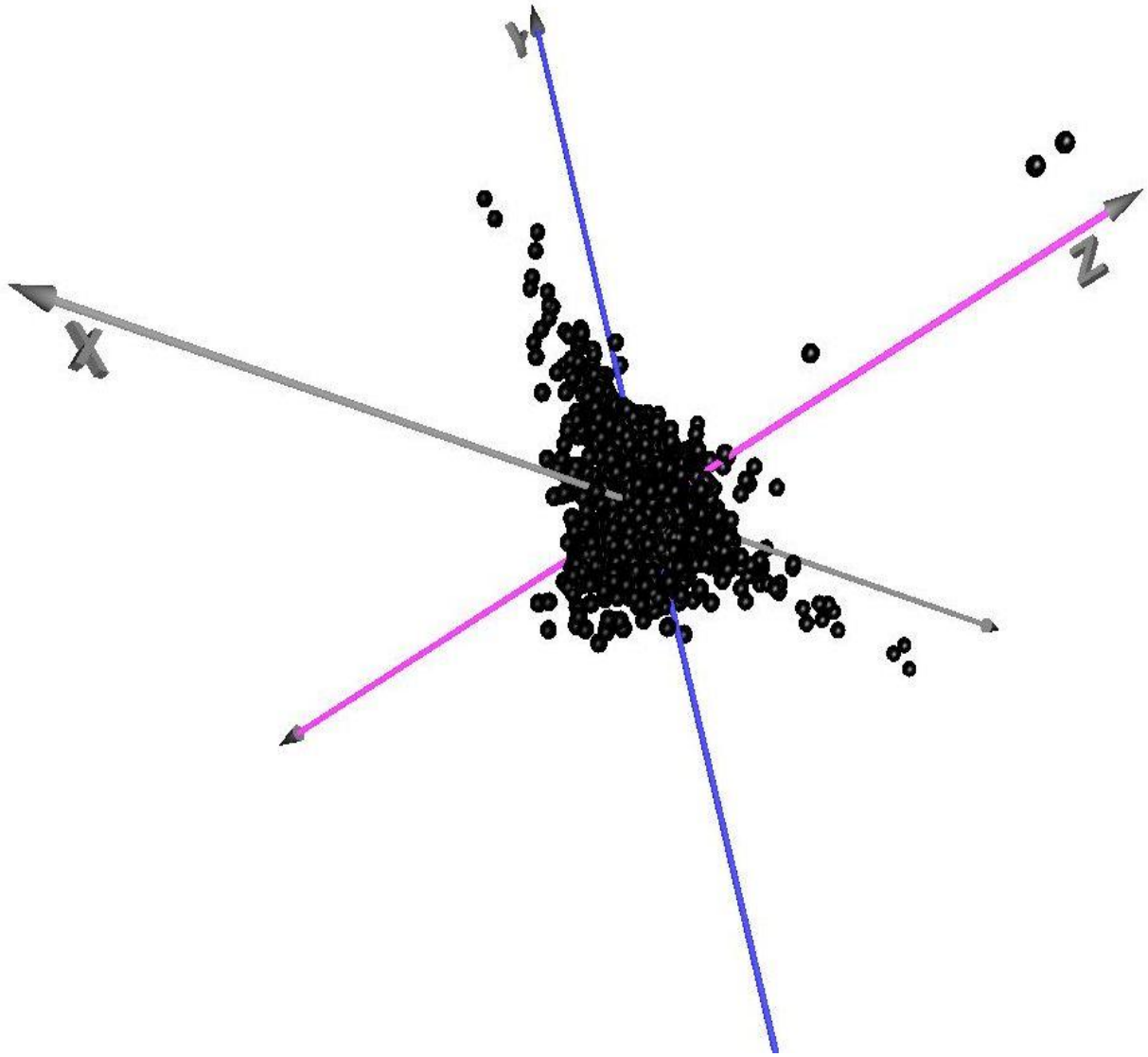

Figure 3.11 The 3-dimensional view of the data in Eigenvector-space. Clusters from different clustering calculations can be visualized using colored points in the 3D-view.After the PCA was performed on the data the above 3D graph was obtained. Here some clusters generated a compact data cloud in space. PCA helped to determine how compact and self-contained a cluster of genes was.

Eigen Values

Here the first 3 PCs combine more than 57% of the variance, so that the components 4 to 8 have all together less than 43% of the information.[20,21] Their patterns mainly describe the noise component in the dataset; PC 8 has less than 5 ppm of the information!

Explained Variance

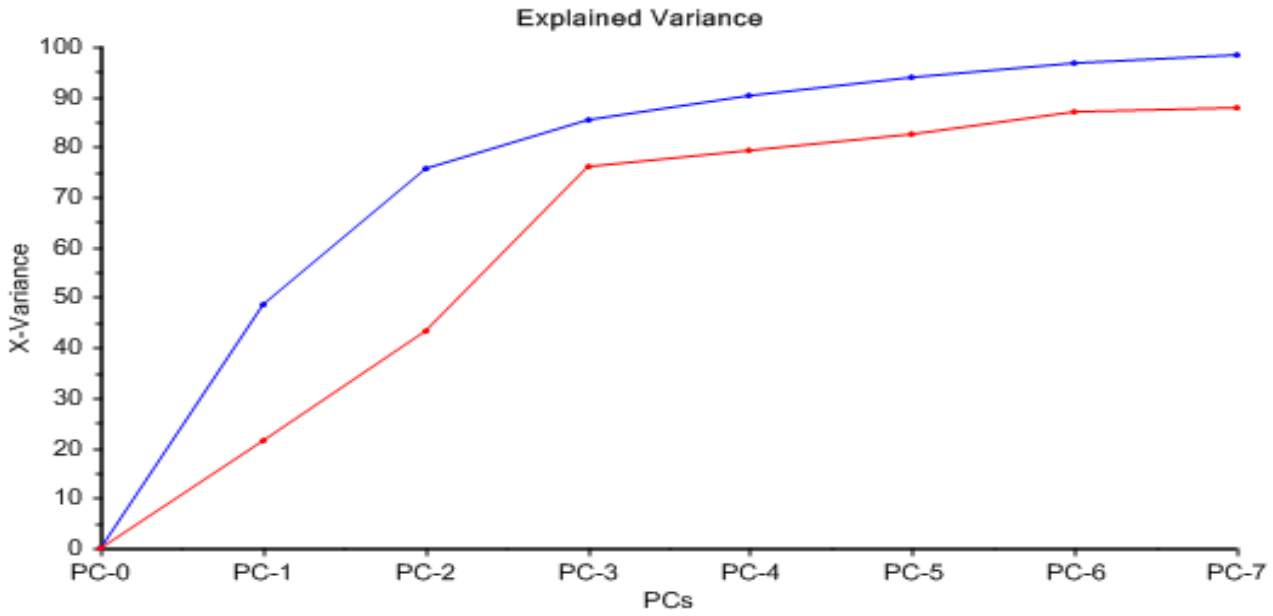

Figure 3.12 Explained Variance obtained by Unscrambler having x-axis representing PCs and y-axis representing the variation of the data from the mean.

Here both the expected and derived lines move in almost parallel fashion hence there is less variation in data.

PCA Projection: Residual Variance

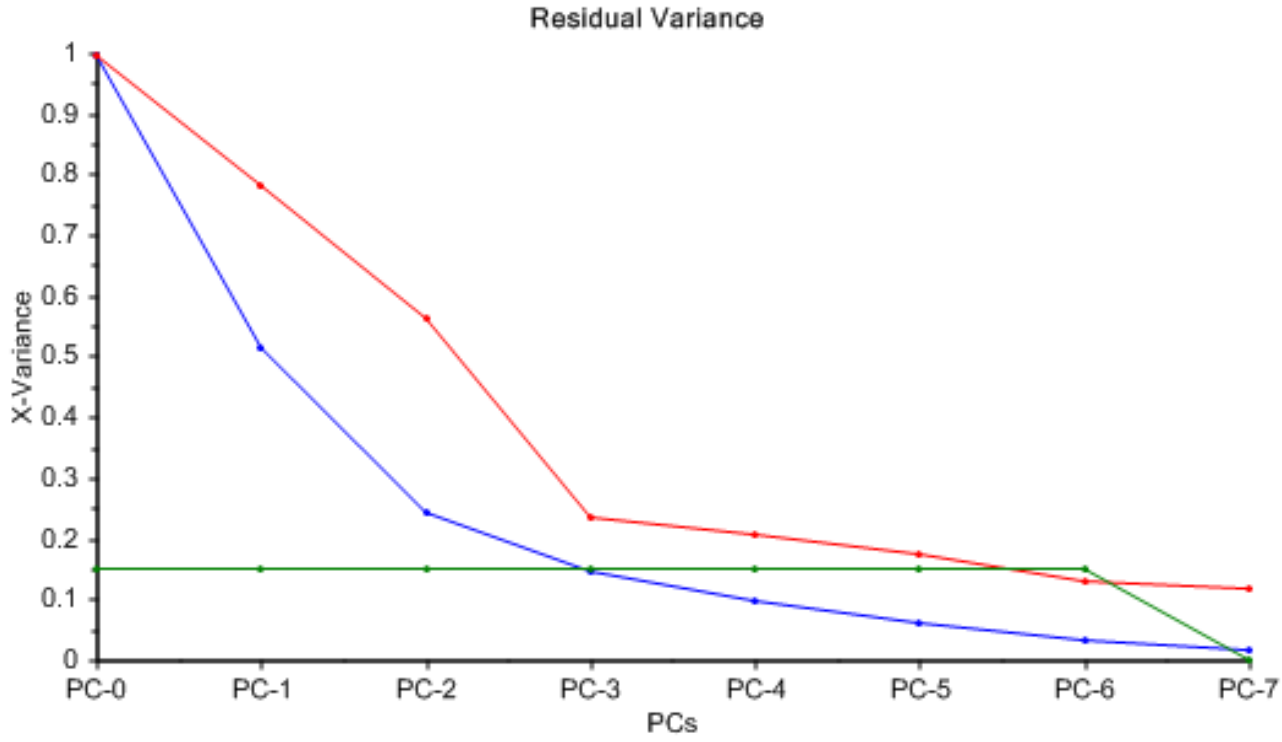

Figure 3.13 Residual Variance obtained by Unscrambler having x-axis representing PCs and y-axis representing the variation of the data from the mean.

The residual variance decreases until PC 7 is reached. The lowest residual variance is found with 7 PCs. The residual variance tells us a lot about how the model performs. By and large models, it should diminish persistently. [22] An expansion demonstrates that there is an issue which ought to be distinguished and evacuated. You may see that the lingering change diminishes consistently from PC 0 to PC 7. Consequently our information is less uproarious and accordingly the nearness of the qualities in charge of the Shoot improvement is affirmed in Cluster 4. The sub-atomic component that could clarify the tso1 freak phenotype is missing. Through a hereditary screen, we recognized 32 silencers that guide to the MYB3R1 quality, encoding a preserved cell cycle regulator.[23] Further investigation demonstrates that TSO1 transcriptional subdues MYB3R1, and the ectopic MYB3R1 action intervenes the tso1 freak phenotype. Since creature homolog's of TSO1 and MYB3R1 are parts of a cell cycle administrative complex, the DREAM complex, we tried and demonstrated that TSO1 and MYB3R1 immune-precipitated in tobacco leaf cells. [26] The work uncovers a monitored cell cycle administrative module, comprising of TSO1 and MYB3R1, for appropriate plant development.[24, 25] Plant postembryonic advancement depends on a little pool of immature microorganisms at the shoot and root tip. The subject of how the cell cycle administrative exercises are coordinated into the explicit undifferentiated organism setting isn't surely knew. [27, 28] This examination distinguishes a formerly obscure administrative module in the blooming plant comprising of two administrative qualities, TSO1 and MYB3R1.[26] TSO1 adversely manages MYB3R1 to control cell division action, keep up appropriate undifferentiated organism pool size, and offset cell expansion with separation in shoot and root. Essentially, creature homologs of TSO1 and MYB3R1 are individuals from a cell cycle administrative complex, proposing this moderated module works in the two plants and animals. [29]

4. CONCLUSION

We conclude from the Residual Variance Curve that seven PCs were optimal. Thus the Cluster 4 of the Genesis data is responsible for the Shoot Development in *Arabidopsis thaliana*. [30]In this work, 'Genesis' a versatile and transparent software suite for large-scale gene expression cluster analysis was used. The Genesis software was used to enable data import, visualization, data normalization, and clustering via: k-means and Self Organizing Maps. Also the Unscrambler software was used as an additional support for the work. It also enabled the data import, visualization and interpretation using Principal Component Analysis. Here we calculated and compared clustering results from different algorithmic approaches. One of the challenges in analyzing microarray data is the fact that there is no biological definition of a gene cluster. Moreover, due to the different underlying assumptions for the clustering techniques and the necessity to adjust various parameters, the clustering results can differ substantially. Thus, it is an imperative to apply several clustering techniques on the same data set and to compare the results. The comparison of clusters obtained using several clustering techniques enabled us to identify genes that have been rated similar in all

clustering results. Here PCA was used to visualize these clusters in 3D space and get an impression of cluster size, integrity, and distribution, and helped to retrieve the most significant patterns in a study.[31] It also revealed some information about the number of clusters in the dataset. All these clustering and classification procedures enabled us to get an impression of subset of genes which are responsible for the Shoot Development in *Arabidopsis thaliana* and thus provided an opportunity for us to concentrate on a particular aim. [11, 15]

ACKNOWLEDGEMENT

Thanks for MANIT provide Bioinformatics Lab for Unscrambler and Genesis software was used as an additional support for the work.

CONFLICT OF INTEREST

There is no any conflict of interest exists.

REFERENCES

1. Emmert-Streib, F. and Dehmer, M. Analysis of Microarray Data A Network-Based Approach. Wiley-VCH. ISBN 3-527-31822-4. 2008.
2. D.W. Meinke, J.M. Cherry, C. Dean, S.D. Rounsley, M. Koornneef. "Arabidopsis thaliana: A Model Plant for Genome Analysis". Science. 1998; 282 662–682.
3. Kulesh DA, Clive DR, Zarlenga DS, Greene JJ. "Identification of interferon-modulated proliferation-related cDNA sequences". Proc Natl Acad Sci USA. 1987; 84 (23): 8453–8457.
4. Lausted C et al. "POSaM: a fast, flexible, open-source, inkjet oligonucleotide synthesizer and microarrayer". Genome Biology. 2004; 5 (8):
5. Más P. "Circadian clock signaling in Arabidopsis thaliana: from gene expression to physiology and development". Int. J. Dev. Biol. 2005; 49 (5–6)
6. GA.Churchill, "Fundamentals of experimental design for cDNA microarrays" (– Scholar search). Nature genetics supplement. 2002; 32: 490–5.
7. Vattani A. "K-means requires exponentially much iteration even in the plane". Discrete and Computational Geometry. 2011; 45 (4): 596–616.
8. NASC-Nottingham Arabidopsis Stock Center-Background Lines-Description-http://arabidopsis.info/CollectionInfo?id=94
9. David P. Horvath, Robert Schaffer, Mark West and Ellen Wisman Arabidopsis microarrays identify conserved and differently expressed genes involved in shoot growth and development form distantly related plant species. The Plant Journal. 2003; 34, 125-134.
10. H. Zha, C. Ding, M. Gu, X. He and H.D. Simon. "Spectral Relaxation for K-means Clustering", http://ranger.uta.edu/~chqding/papers/Zha-Kmeans.pdf, Neural Information Processing Systems vol.14 (NIPS 2001). pp. 1057–1064, Vancouver, Canada. Dec. 2001.

11. Hamerly, G. and Elkan, C. "Alternatives to the k-means algorithm that find better clusterings". Proceedings of the eleventh international conference on Information and knowledge management (CIKM). 2002
12. Arthur; Abhishek Bhowmick. "A theoretical analysis of Lloyd's algorithm for k-means clustering".2009
13. Chris Ding and Xiaofeng He. "K-means Clustering via Principal Component Analysis". Proc. of Int'l Conf. Machine Learning ICML 2004; 225–232.
14. Ding C. and X. He. "K-means Clustering via Principal Component Analysis". Proc. of Int'l Conf. Machine Learning ICML. 2004; 225–232.
15. Hartigan, J. A.; Wong, M. A.. "Algorithm AS 136: A K-Means Clustering Algorithm". Journal of the Royal Statistical Society, Series C (Applied Statistics). 1979; 28 (1): 100–108. JSTOR 2346830.
16. Alexander Sturn. “Cluster Analysis for Large Scale Gene Expression Studies”. 2000.
17. Bammler T, Beyer RP; Consortium, Members of the Toxicogenomics. "Standardizing global gene expression analysis between laboratories and across platforms". Nat Methods. 2005; 2 (5): 351–356.
18. Kanungo, T.; Mount, D. M.; Netanyahu, N. S.; Piatko, C. D.; Silverman, R.; Wu, A. Y.. "An efficient k-means clustering algorithm: Analysis and implementation". IEEE Trans. Pattern Analysis and Machine Intelligence. 2002; 24: 881–892.
19. Priness I., Maimon O., Ben-Gal I. "Evaluation of gene-expression clustering via mutual information distance measure"]. BMC Bioinformatics. 2007; 8 (1): 111.
20. Schena M, Shalon D, Davis RW, Brown PO. "Quantitative monitoring of gene expression patterns with a complementary DNA microarray". Science. 1995; 270 (5235): 467–470.
21. Shalon D, Smith SJ, Brown PO (). "A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization". Genome Res. 1996; 6 (7): 639–645.
22. The Arabidopsis Biological Resource Center (ABRC), http://abrc.osu.edu The Arabidopsis Genome Initiative. "Analysis of the genome sequence of the flowering plant Arabidopsis thaliana". Nature .2000; 408 (6814): 796–815.
23. Litovchick L, et al. Evolutionarily conserved multi subunit RBL2/p130 and E2F4 protein complex represses human cell cycle-dependent genes in quiescence. Mol Cell. 2007; 26:539–551.
24. Kobayashi K, et al. Transcriptional repression by MYB3R proteins regulates plant organ growth. EMBO J. 2015; 34.
25. Sijacic P, Wang W, Liu Z. Recessive antimorphic alleles overcome functionally redundant loci to reveal TSO1 function in Arabidopsis flowers and meristems. PLoS Genet. 2011; 7:e1002352
26. Wanpeng Wang, Paja Sijacic, Pengbo Xu, Hongli Lian, and Zhongchi Liu Arabidopsis SO1 and

MYB3R1 form a regulatory module to coordinate cell proliferation with differentiation in shoot and root. PNAS. 2018;115, 13 3045–3054.

27. Fischer M, Müller GA (2017) Cell cycle transcription control: DREAM/MuvB and RB-E2F complexes. Crit Rev Biochem Mol Biol 52:638–662.
28. Sarkar AK, et al. Conserved factors regulate signalling in Arabidopsis thaliana shoot and root stem cell organizers. Nature. 2007; 446:811–814.
29. Litovchick L, et al. Evolutionarily conserved multisubunit RBL2/p130 and E2F4 protein complex represses human cell cycle-dependent genes in quiescence. Mol Cell. 2007; 26:539–551.
30. Yuk Fai Leung and Duccio Cavalieri, Fundamentals of cDNA microarray data analysis. TRENDS in Genetics. 2003; 19, 11
31. Tang T, François N, Glatigny A, Agier N, Mucchielli MH, Aggerbeck L, Delacroix H. "Expression ratio evaluation in two-colour microarray experiments is significantly improved by correcting image misalignment". Bioinformatics. 2007; 23 (20): 2686–2691.