# GENE ACTIVITY ANALYSIS OF DISEASE GLIOBLASTOMA MULTIFORME WITH DATA MINING METHODS

1st Fatma Tanrikulu
*Department of Computer Engineering*
*Yildiz Technical University*
İstanbul, Turkey
fatmaatanrikulu@gmail.com

2th Mustafa Aydin
*Department of Computer Engineering*
*Yildiz Technical University*
İstanbul, Turkey
82mustafa82@gmail.com

***Abstract*—Glioblastoma Multiforme dünya üzerinde yaygın görülennn ve bir kaç çeşidi olan en tehlikeli beyin tümörlerindendir. Henüz kesin bir tedavisi bulunamayan ve kısa sürede yüksek hayati tehlike barındıran bu tümör, insan hücrelerindeki birçok protein tarafından desteklenebilmektedir. İnsan hücrelerinde yaşamsal önem taşıyan ve organizmanın devamlılığında kritik rol oynayan genler mutasyon sonrası tamamen tümör lehine aktivite gösterebilmektedir. Araştırmalara göre gerçekleşen mutasyonlar sonucunda tümör hücrelerini tetikleyici ve hareket etkinliğini artırıcı yönde hareket etmeye başlayan pek çok gen ve sinyal yolu mevcuttur. [2], [5], [8]**

***Index Terms*—Glioblastoma, data mining, R language, gbm**

## I. Introduction

Glioblastoma Multiforme, also known as GBM, GBM-IV or 4th degree tumor, is the most common type of malignant brain tumor. Glioblastoma consists of astrocytomies which is the highest-degree (4th degree) variety of gliomas. Gliomas are groups of tumors formed in glia cells that protect and support nerve cells. GBM is a primary malignant brain tumor and does not have a definitive treatment besides its rapid progression. The goal of the project is to research and analyze the gene activity of Glioblastoma Multiforme patients from the literature, web and data sets. [1], [3], [6], [9]

The biological data related to GBM which obtained from data banks such as NCBI, Broad GDAC is analyzed in the scope of project. Also these data are classified with learning algorithms, and based on this classification, the process of making predictions and analyzing the differences between data is implemented in line with the project objectives. These applications to be carried out during the project process is concluded by observing the classification,clustering, association rules and statistical methods in data mining. The data of sick individuals are compared with the data of normal individuals and gene activity is determined. Process of these analyses R programming language that has "Bioconductor" libraries(eg Affy, Limma, Deseq2, Illumio) is used. [4]

Within the scope of the projects 6 data sets from different information banks are used. These data sets were subjected to different classification, clustering and association methods and gene activity of the sick and normal individuals' was examined.

## II. Classification & Clustering Methods

Various of classification and clustering methods were applied on the different data sets which have been obtained from the data banks such as NCBI.

### A. gbm - iDINGO

In order to analyze in the project, the iDINGO package of the R language was chosen because it met the need for the data set for glioblastoma disease. There are 156 samples of 18 proteins in this data set. Covariates constitute the first column of the data set, while the remaining herds contain gene expressions. By applying different classification and clustering methods on this data set, it was tried to understand, analyze and analyze the data set. For each classification or clustering method studied, different classes available in the package were used, depending on the requirements of the method.

*1) Gini Method:* Gini classification is applied on gbm - iDINGO data set and as a result PIK3R2 gene's Gini coefficient below 0.97 in %85 of the samples. The number of samples in which the Gini coefficient of the PIK3R1 protein, which is similar to the PIK3R2 proteins, is less than -0.13, is %47, that is, almost half of the samples. The sample rate where the value of PIK3C2B is above 0.08 is %23, that is, where the value of PIK3R1 is low, the value of PIK3C2B is also low. In this way, we can conclude that PIK3R1 and PIK3C2B genes act together for the gbm data set and Gini classification. Following this, it was concluded that the proteins PIK3R1, PTEN and PIK3C2B genes were large in coefficients. It was concluded that the genes with the highest activity in Glioblastoma patients in the Gbm data set were those with a high Gini coefficient.

*2) Regression Analysis:* As a result of the regression analysis applied to the gbm data set, a regression tree was created. According to the regression tree obtained, it was concluded

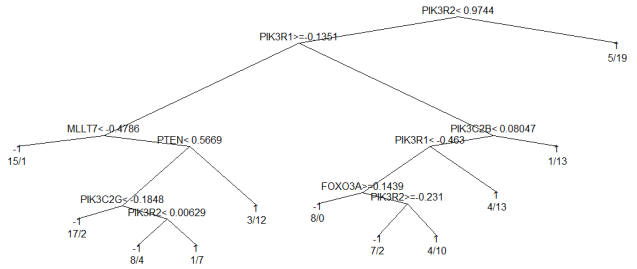


Fig. 1. Result Tree Graph after Gini Classification on gbm - iDINGO

that the most determining genes were Pık3R2, PIK3R1, followed by PIK3C2B, MLLT7. The data extracted from the gbm data set as a result of the regression analysis application overlaps with the data obtained in the Gini classification, and it was assumed that it was applied correctly because it was in line with the results of the literature and web research.

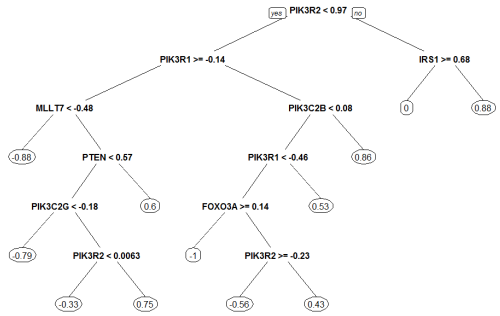


Fig. 2. Regression Analysis on gbm - iDINGO

*3) Random Forest Classification:* In the random forest method applied on the gbm data set, 500 decision trees were used and predictions were made for each sample.

While the error rate of the random forest method applied in the first graph was 1.7 at first, then this rate decreased to around 1. This shows that a correct machine learning is done by using trees.

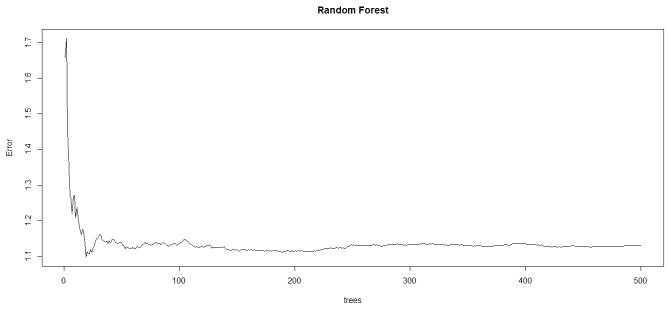


Fig. 3. Random Forest Error Rate on gbm iDINGO

According to the importance order graph of the genes obtained as a result of the random forest algorithm, it is seen that the PIK3R2 gene is ahead of the proteins in the gbm data set.

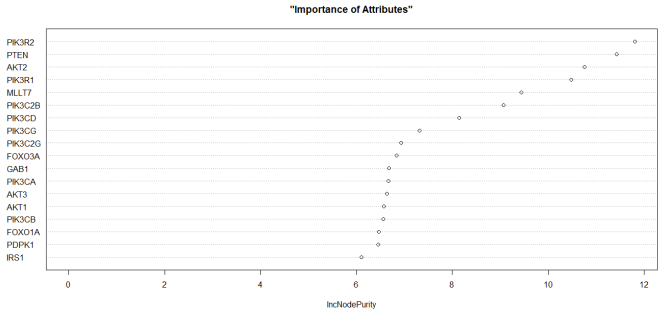


Fig. 4. Random Forest importance ofder graph on gbm iDINGO

*4) Pam Clustering Method:* When PAM clustering method was applied to the Gbm data set, it was observed that the data set was divided into 3 separate clusters. This cluster, which has a total of 156 samples, was calculated as a result of the calculations made by the algorithm itself, and as a result, it was reported that 39 samples converged to the mean value of 0.09, 69 samples to 0.06 and 48 samples to 0.11.

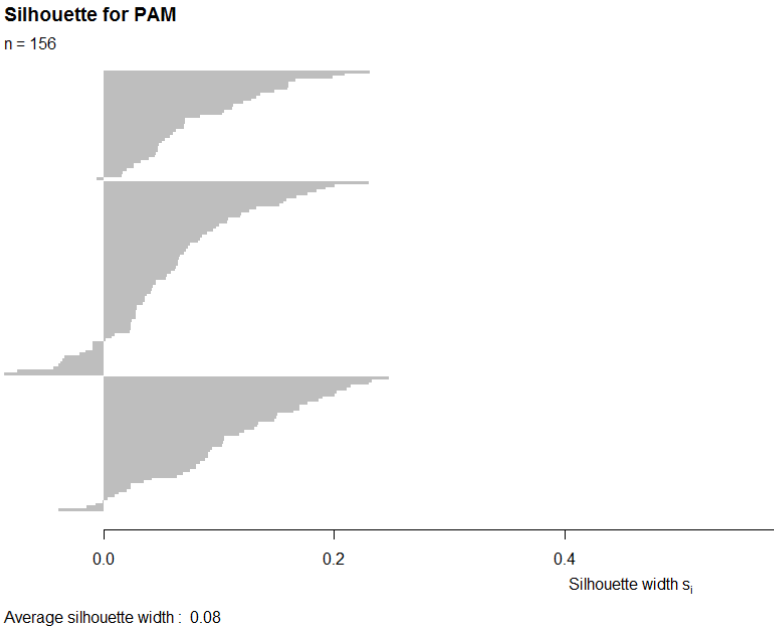


Fig. 5. Pam clustering on GBM iDINGO

## B. snv - expands

The snv data set examines mutations and loss of heterozygosity in GBM on the basis of nucleotides. In each row in this data set, which consists of a total of 773 rows and 7 rows, belonging to 23 chromosome pairs; the chromosome found, the examination start site, the end site, the reference nucleotide, the b-allele nucleotide, the frequency of the b-allele and the number of that b-allele in the normal cell are found.

*1) Regression Analysis:* The regression analysis performed on the Snv data package was carried out over the allele gene frequencies in the data set. Allele gene frequencies on chromosomes were classified from 0.82 nodes in accordance with the applied regression formulas and rules. According to the regression tree obtained, it was learned that the samples with an allele gene frequency less than 0.82 or between 0.82 and 1 had tumors, while samples with a frequency greater than 1 were healthy.

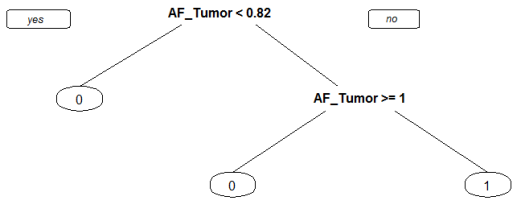


Fig. 6. Regression Analysis on snv - expands

*2) Random Forest Classification:* A decision forest was created by using 500 decision trees on the Snv data set. As a result of classification, results were obtained about whether each chromosome region was tumorous or not.

While the error rate of the random forest algorithm applied on the Snv data set was initially 0.05, it was observed in the graph that it decreased to 0.01 later on.

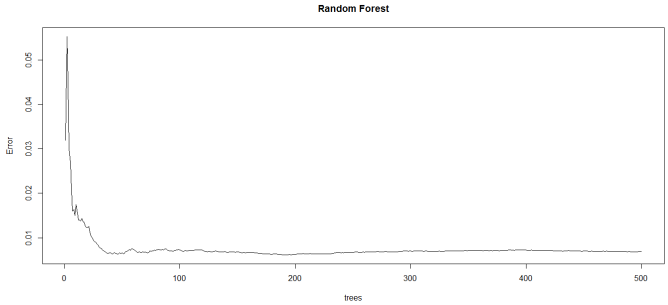


Fig. 7. Random Forest Error Rate on snv - expands

As a result of the applications, it was concluded that 0 value represented tumor and 1 value represented healthy individuals and %99.87 success was achieved.

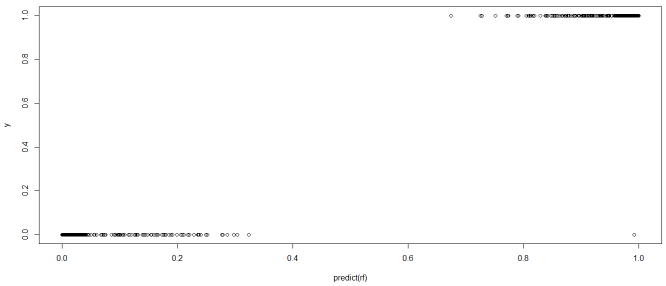


Fig. 8. Random Forest Analysis on snv - expands

### *C. cbs - expands Data set*

Similar to the Snv data set, this data set is a 120-row and 4-column data set from a glioblastoma tumor. It includes the chromosome of interest, the start and end location, and the average copy number of the segment among all cells.

*1) Random Forest Analysis:* A decision forest was created by using 500 decision trees on the GIS data set. While 120 samples were represented horizontally, 23 chromosome pairs were represented vertically. As a result of the random forest algorithm applied in the bottom graph, the graph moving in the z-axis showed that the real chromosomes and the chromosomes predicted as a result of the algorithm mostly overlapped. In this way, it was concluded that the applied random forest algorithm achieved a high rate of success.

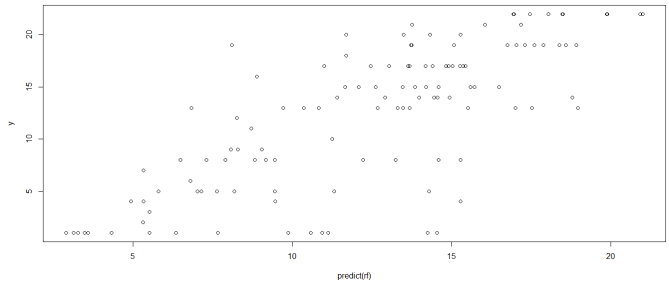


Fig. 9. Random Forest Analysis on cbs - expands

### *D. Methylmix Data set*

The METnormal and METcancer data sets belong to the "MethylMix" package, which aims to identify methylation-related cancer genes.

METnormal: DNA methylation data of normal tissue from glioblastoma patients.

METcancer: DNA methylation data of cancerous tissue from glioblastoma patients.

*1) Agglomerative Clustering:* Agnes method was applied to the data of normal people in the METnormal data set and it was concluded that FNDC3, OAS1 and SOX10 genes act together, apart from other genes that affect methylation. [6]

When Agnes method was applied on METcancer, it was seen that FNDC3, HOXD1, ME1 and SOX10 genes worked

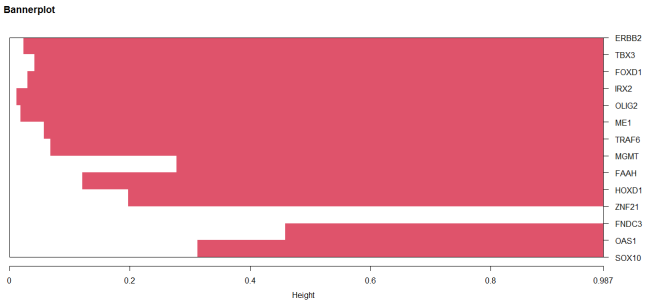


Fig. 10. Agnes Method on METnormal data set in MethylMix

together. The methylation cycle, which is disrupted after Gbm disease, was explained in this way by the Agnes method.

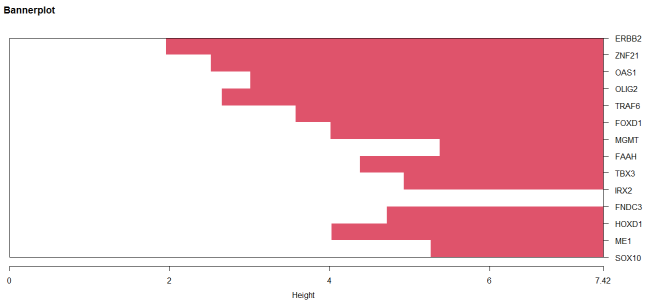


Fig. 11. Agnes Method on METcancer data set in MethylMix

*2) Average Linkage Clustering:* When the Average Link clustering method was applied to the METnormal data set, it was seen that the FNDC3, OAS1 and SOX10 genes formed a cluster in normal patients. In this way, the Agnes clustering method applied within the METnormal data set was verified, and it was understood that the methods applied for the joint movement of the genes in question provided a high success rate.

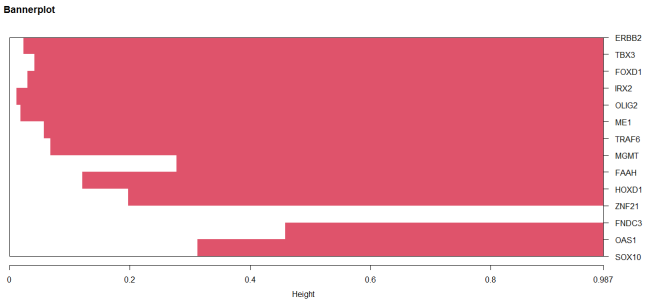


Fig. 12. Average Linkage Method on METnormal MethylMix

Average Link clustering method was applied to METcancer data set and it was seen that FNDC3, HOXD1, ME1 and SOX10 genes formed a cluster as in Agnes method.

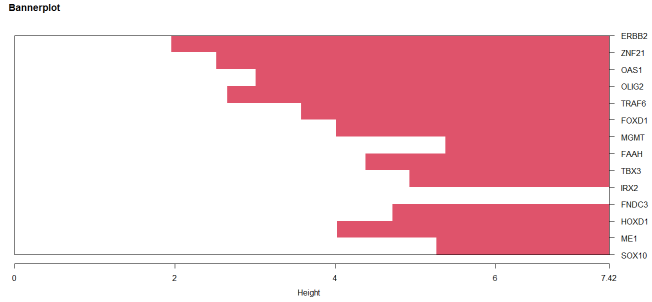


Fig. 13. Average Linkage Method on METcancer MethylMix

*3) Ward Clustering:* When we applied the Ward clustering method to the METnormal data set, it was concluded that the FNDC3, OAS1 and SOX10 genes formed a cluster, as in the Agnes and Average Link methods. Apart from these, it was seen that ZNF217, FAAH and HOXD10 genes formed a separate cluster, while MGMT, ERBB2 and TBX3 genes formed a separate cluster.

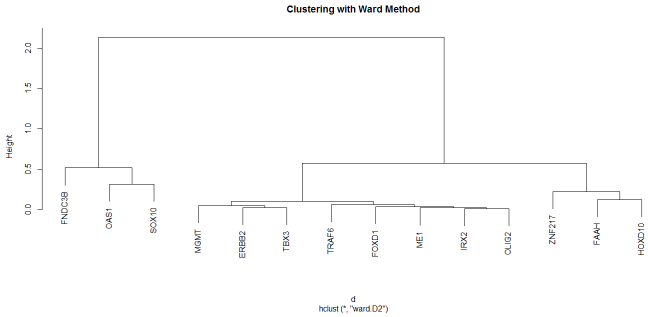


Fig. 14. Ward Method on METnormal MethylMix

When we applied the Ward clustering method on diseased data in the METcancer data set, HOXD10 and ME1 and FNDC3B and SOXD10; OAS1, ERBB2 and ZNF217 with FOXD1, OLIG2 and TRAF6; It was seen that IRX2 and MGMT and FAAH and TBX3 genes formed clusters.

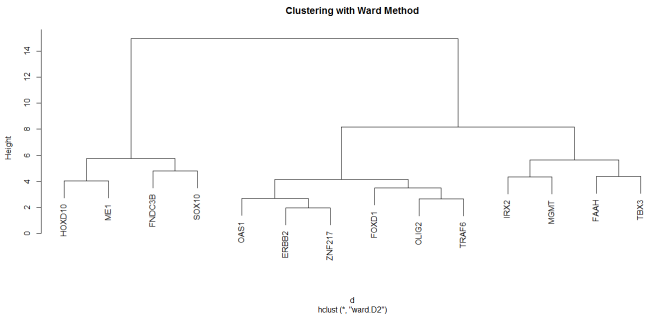


Fig. 15. Ward Method on METcancer MethylMix

*4) Centroid Clustering:* When the Centroid clustering method was applied to the METnormal data set, it was learned that FNDC3, OAS1 and SOX10 genes acted together, as in

Agnes, Average Link and Ward methods. It was also observed that ZNF217, FAAH and HOXD10 genes formed a separate cluster.

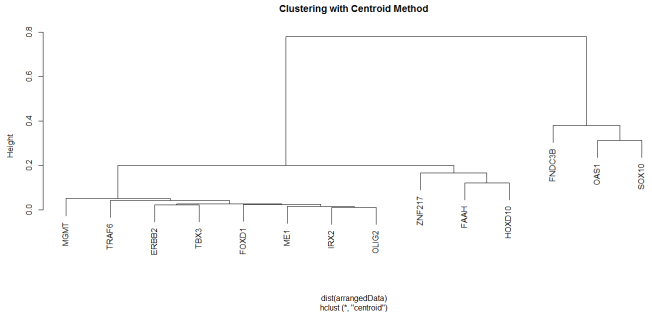


Fig. 16. Centroid Method on METnormal MethylMix

When the Centroid clustering method is applied to diseased data, the other FAAH, TBX3 and IRX2 are a cluster; It was observed that HOXD10 and ME1 formed a separate cluster and FND3B and SOX10 formed a separate cluster.

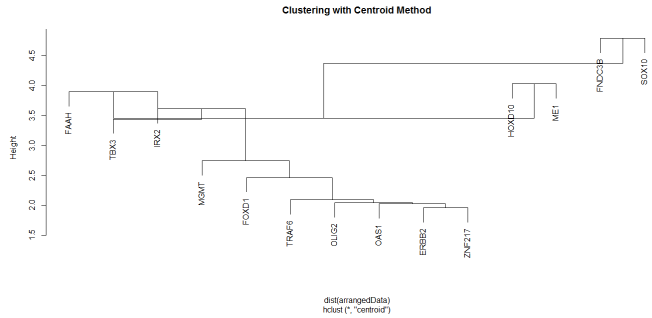


Fig. 17. Centroid Method on METcancer MethylMix

*5) Diana Clustering:* When the DIANA clustering method was applied to the METnormal data set, it was seen that the FNDC3, OAS1 and SOX10 genes formed a cluster, as in the Agnes, Average Link, Ward and Centroid methods.

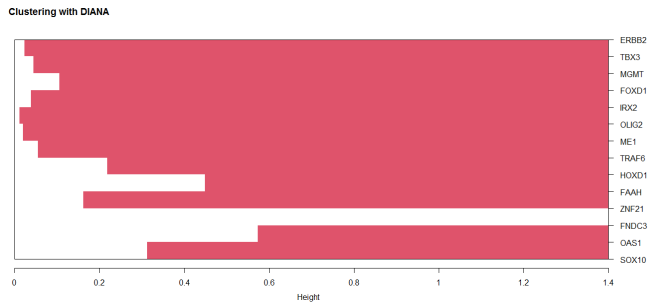


Fig. 18. Diana Method on METnormal MethylMix

The METcancer data set containing the diseased data was analyzed by Diana clustering method and it was measured that the common activity of the FNDC3, HOXD1, ME1 and SOX10 methylation genes was higher than the other genes.

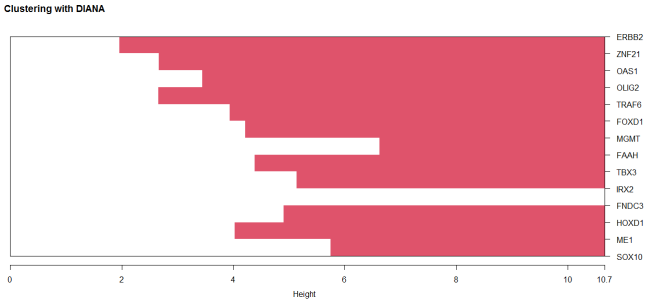


Fig. 19. Diana Method on METcancer MethylMix

*6) K-Means Clustering:* It was observed that FNDC3, OAS1 and SOX10 genes formed a cluster as in Agnes, Average Link, Ward, Centroid and DIANA methods, and ZNF217, FAAH and HOXD10 genes also formed a separate cluster.

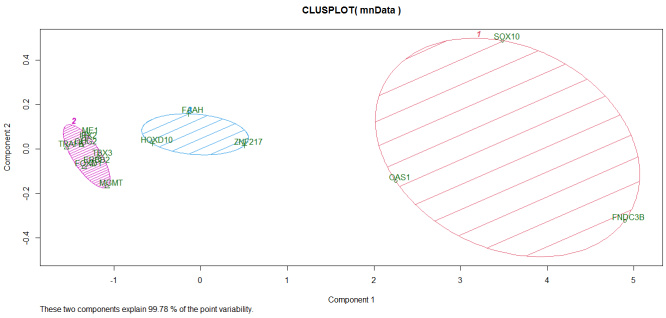


Fig. 20. K-Means Method on METnormal MethylMix

*7) Pam Clustering:* As in Agnes, Average Link, Ward, Centroid, DIANA and K-Means methods, it was observed that FNDC3, OAS1 and SOX10 genes acted together and ZNF217, FAAH and HOXD10 genes also showed separate joint activity.

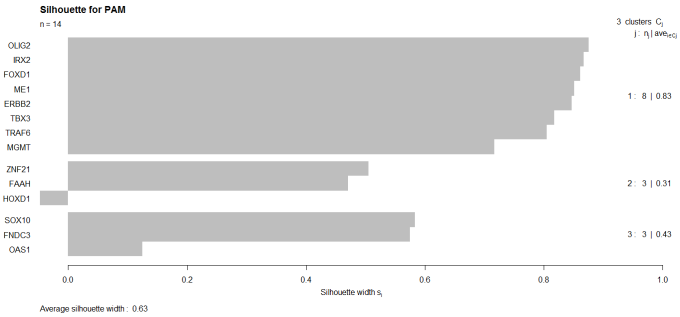


Fig. 21. Pam Method on METnormal MethylMix

With the PAM clustering method, it was learned that FNDC3, SOX10, ME1, HOXD1 and IRX2 genes formed one cluster, and FAAH and TBX3 genes formed a separate cluster in this data set.

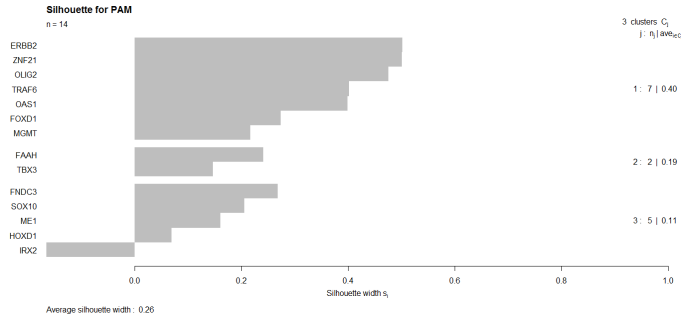


Fig. 22. Pam Method on METcancer MethylMix

*8) Clustering Large Applications Method:* The CLARA clustering method was applied on the METnormal data set and information was obtained that FNDC3, OAS1 and SOX10 genes formed a cluster, as in Agnes, Average Link, Ward, Centroid, DIANA, K-Means and PAM methods. ZNF21, FAAH and HOXD10 genes were also observed to form a separate cluster.

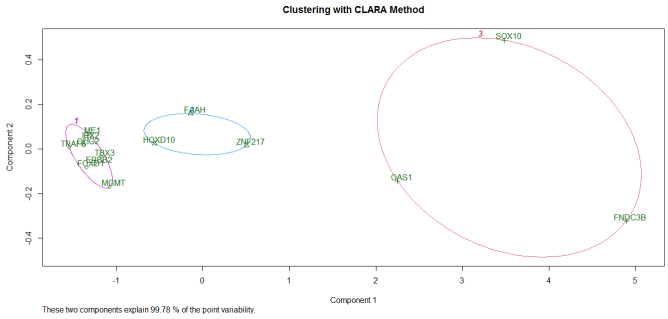


Fig. 23. Clara Method on METnormal MethylMix

## *E. gbm - iClusterPlus Data set*

The iClusterPlus gbm package, a set of data set from the TCGA study of glioblastoma disease, has copy number, methylation and mRNA data for a total of 82 different samples.

gbm.exp: GBM gene expression data. In this cluster, 1740 proteins are analyzed on 84 samples.

gbm.mut GBM mutation data. In this cluster, 306 mutations are examined on 84 samples.

In the gbm.exp data set, which examines the gene activity of gbm patients over 1740 genes, the data were divided into 2 separate clusters as a result of the Pam algorithm. When this set, which has 84 samples in total, was observed, it was seen that 37 samples converged to 0.14 and 47 samples to 0.10.

In this data set, which examined the mutations of 84 samples in gbm.exp, it was seen that the genes formed a single cluster after being mutated. As a result of this research on 306 genes, it was found that 82 samples converged to the value of 0.66, and thus, the high gene frequency was proportional to the severity of the gbm disease.

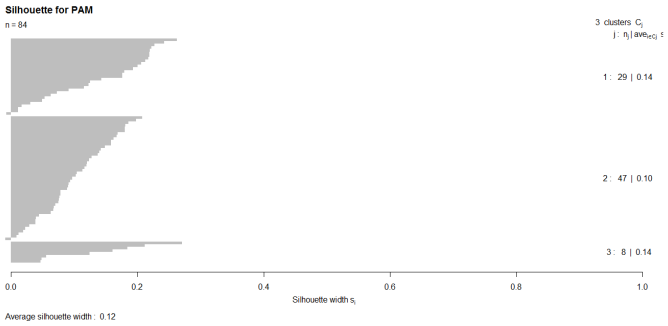


Fig. 24. Pam Method on gbm.exp iClusterPlus

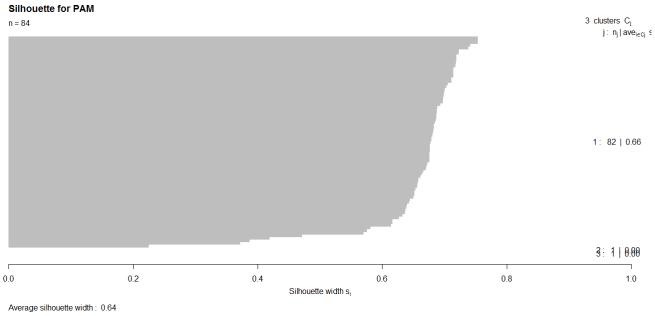


Fig. 25. Pam Method on gbm.mut iClusterPlus

## III. Result

During the project process, many classification, clustering and statistical methods were applied on the data sets and some estimation algorithms were tested. It was observed that the severity of the disease was higher in samples with high frequencies of genes under the influence of glioblastoma disease. Then, it was examined which genes had the highest frequencies and it was seen that the PIK3R, PTEN and AKT gene families had higher frequencies, respectively. Since the AKT, PIK3R and PTEN gene families have extremely vital functions responsible for cell survival, the diagnosis of Glioblastoma as fatal was supported in the project. [**?**], [12], [13], [16]

As a result of the studies carried out within the scope of the project, it was concluded that the most effective gene in Glioblastoma is PIK3R2. It became clear why Glioblastoma is fatal, as these genes, which are known to play a role in vital actions such as cell growth, survival, protein and insulin synthesis, widen the range of action of Glioblastoma disease and support tumorigenesis as a result of mutations. In the literature research, the knowledge that PIK3R2 and PIK3R1 act together was supported as a result of the processes carried out in the project, and it was observed that these two genes acted almost in common in the project and were the genes most affected by Glioblastoma. The PIK3C family was found where the PIK3R genes had a high frequency, and it was found that the PIK3C2G gene helped the recurrence of Glioblastoma. Apart from these genes, it was observed that PTEN, which is a tumor suppressor gene that takes part in the same vital

activities, has a higher frequency in areas where the PIK3R gene family is weak. [14], [15]

In the clustering processes implemented within the project, it was observed that the gbm data of the iDINGO package formed 3 clusters within the framework of 0.09, 0.06 and 0.11. As a result of applying the same clustering methods on the gbm.exp data set belonging to the iClusterPlus package and examining gene activity, it was seen that 2 clusters were formed based on the values of 0.14 and 0.10. Finally, the same clustering methods were applied on the gbm.mut data set, which is another data set of the iClusterPlus package and contains the mutated data, and it was seen that it formed a single cluster based on the 0.66 value. As a result of the same clustering methods applied on 3 different data sets; It has been reported that gene frequencies are higher in cancerous cells mutated by the effect of Glioblastoma disease.

In order to examine the effect of glioblastoma on the methylation cycle, comprehensive classification and clustering methods were applied in the MethylMix package, which includes both cancer data and normal data. According to the analysis results of the METnormal data set containing healthy data; It was observed that ERBB2, MGMT and TBX3 genes, which encode protein kinase and are involved in mutation repair, cell development and regeneration, have high frequencies and act together in healthy individuals.

FND3C3B, OAS1 and SOX10 genes, which play an important role in protein coding, virus immunity, embryonic cell development and cell cycle, were low-frequency and also acted together.

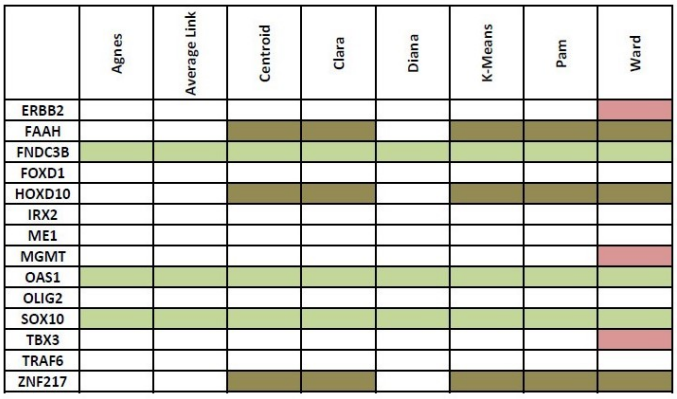


Fig. 26. METnormal MethylMix - Graph of Which Genes Joint Action in Healthy Cells as a Result of the Project

In order to compare the changes that occur as a result of tumor formation and progression, the same classification and clustering methods were applied to the METcancer data set, which belongs to the MethylMix package and contains the data of cancer cells. It was concluded that the FAAH, IRX2 and TBX3 genes, which are involved in fatty acid production, embryonic development and regeneration, have a high frequency in cancer cells and act together. It was concluded that they showed joint activity.

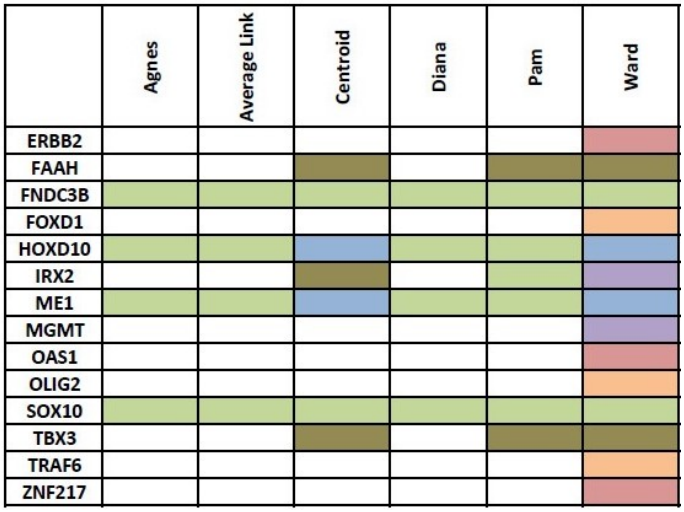


Fig. 27. METcancer MethylMix - Graph of Which Genes Joint Action in Tumorous Cells as a Result of the Project

## REFERENCES


[1] Hiroko Ohgaki, Paul Kleihues, Genetic Pathways to Primary and Secondary Glioblastoma, The American Journal of Pathology, Volume 170, Issue 5, 2007, Pages 1445-1453, ISSN 0002-9440, https://doi.org/10.2353/ajpath.2007.070011.

[2] Romana Richterová and Branislav Kolarovszki (July 13th 2016). Genetic Alterations of Glioblastoma, Neurooncology - Newer Developments, Amit Agrawal, IntechOpen, DOI: 10.5772/63127.

[3] Olar, A., & Aldape, K. D. (2014). Using the molecular classification of glioblastoma to inform personalized treatment. The Journal of pathology, 232(2), 165–177.

[4] Xu, H., Hu, Y., & Qiu, W. (2017). Potential mechanisms of microRNA-129-5p in inhibiting cell processes including viability, proliferation, migration and invasiveness of glioblastoma cells U87 through targeting FNDC3B. Biomedicine & Pharmacotherapy, 87, 405–411. doi:10.1016/j.biopha.2016.12.100

[5] Das KK, Kumar R. Pediatric Glioblastoma. In: De Vleeschouwer S, editor. Glioblastoma [Internet]. Brisbane (AU): Codon Publications; 2017 Sep 27. Chapter 15. Available from: https://www.ncbi.nlm.nih.gov/books/NBK469983/

[6] Durmaz Ramazan, Vural Murat, Primer ve Sekonder Glioblastoma Multiforme Genetiği, Eskişehir Osmangazi Üniversitesi, Tıp Fakültesi, Nöroşirürji Anabilim Dalı Eskişehir 2007

[7] Sanchez-Vega et al., 2018, Cell 173, 321–337 April 5, 2018 ª 2018 Elsevier Inc doi.org/10.1016/j.cell.2018.03.035.

[8] Rasras Saleh, Zibara Kazem, Vosughi Tina, Zayeri Zeinab Deris, Genetics and Epigenetics of Glioblastoma: Therapeutic Challenges. Year : 2018 Volume: 7 Issue : 2 Page no: 43-49

[9] Liu, A., Hou, C., Chen, H., Zong, X.,& Zong, P. (2016). Genetics and Epigenetics of Glioblastoma: Applications and Overall Incidence of IDH1 Mutation. Frontiers in oncology, 6, 16. doi.org/10.3389/fonc.2016.00016

[10] Carpten JD, Faber AL, Horn C, Donoho GP, Briggs SL, Robbins CM, Hostetter G, Boguslawski S, Moses TY, Savage S, Uhlik M, Lin A, Du J, Qian YW, Zeckner DJ, Tucker-Kellogg G, Touchman J, Patel K, Mousses S, Bittner M, Schevitz R, Lai MH, Blanchard KL, Thomas JE. A transforming mutation in the pleckstrin homology domain of AKT1 in cancer. Nature. 2007 Jul 26;448(7152):439-44. Epub 2007 Jul 4.

[11] Showler K, Nishimura M, Daino K, Imaoka T, Nishimura Y, Morioka T, Blyth BJ, Kokubo T, Takabatake M, Fukuda M, Moriyama H, Kakinuma S, Fukushi M, Shimada Y. Analysis of genes involved in the PI3K/Akt pathway in radiation- and MNU-induced rat mammary carcinomas. J Radiat Res. 2017 Mar 1;58(2):183-194. doi: 10.1093/jrr/rrw097. PMID: 27738081; PMCID: PMC5571612.

[12] Osaki M, Oshimura M, Ito H. PI3K-Akt pathway: its functions and alterations in human cancer. Apoptosis. 2004 Nov;9(6):667-76. doi: 10.1023/B:APPT.0000045801.15585.dd. PMID: 15505410.

[13] Fresno Vara JA, Casado E, de Castro J, Cejas P, Belda-Iniesta C, González-Barón M. PI3K/Akt signalling pathway and cancer. Cancer Treat Rev. 2004 Apr;30(2):193-204. doi: 10.1016/j.ctrv.2003.07.007. PMID: 15023437.

[14] Mazloumi Gavgani, F., Smith Arnesen, V., Jacobsen, R. G., Krakstad, C., Hoivik, E. A., & Lewis, A. E. (2018). Class I Phosphoinositide 3-Kinase PIK3CA/p110 and PIK3CB/p110 Isoforms in Endometrial Cancer. International journal of molecular sciences, 19(12), 3931. doi.org/10.3390/ijms19123931

[15] Cheng, C. K., Fan, Q. W., & Weiss, W. A. (2009). PI3K signaling in glioma–animal models and therapeutic challenges. Brain pathology (Zurich, Switzerland), 19(1), 112–120. https://doi.org/10.1111/j.1750-3639.2008.00233.x

[16] Emdad, L., Hu, B., Das, S. K., Sarkar, D., & Fisher, P. B. (2015). AEG-1-AKT2: A novel complex controlling the aggressiveness of glioblastoma. Molecular & cellular oncology, 2(3), e995008. doi.org/10.4161/23723556.2014.995008