İstanbul Şehir University

Computational Genomics Project
Exploratory and Predictive Analysis of TCGA data on Glioma

Prepared by

Kanza Batool, Ömer Kaan Karahan

Abstract

In this project, 691 instances of glioma cases are examined to find out important genes which would be responsible for glioma cases by carrying out t-test for normal and tumor cells. We also examined which genes would be responsible for Glioblastoma (grade IV glioma) via carrying out t-test for grade IV and grade II and III patients. Data is derived from TCGA project. Finally we tried to find out whether there exist a linear relationship between age and expression level of genes that would be responsible for cancer cases via carrying out a regression analysis for every single gene and age.

1. Introduction

At the beginning of 20th century microbes, viruses and contagious diseases were the most important causes of human death tolls on earth. In 21st century cancer leads the way as second cause of death in developed countries [1]. Cancer is a group of diseases characterized by the uncontrolled growth and spread of abnormal cells. If the spread is not controlled, it can interfere with essential life-sustaining systems and result in death. According to center for health statistics [2] nearly 600 thousand people die because of cancer solely in the USA. Therefore, it is a huge area of research through out the world. There are plenty of projects and researches carried out by scientists, doctors and engineers supported by both governmental and private funds.

1.1 TCGA (The Cancer Genome Atlas).

One of these projects is TCGA (The Cancer Genome Atlas). In their review article about TCGA Tomczak [3,4] et al. defines TCGA as follows:

> *The Cancer Genome Atlas (TCGA) is a public funded project that aims to catalogue and discover major cancer-causing genomic alterations to create a comprehensive "atlas" of cancer genomic profiles. So far, TCGA researchers have analysed large cohorts of over 30 human tumours through large-scale genome sequencing and integrated multi-dimensional analyses. Studies of individual cancer types, as well as comprehensive pan-cancer analyses have extended current knowledge of tumorigenesis. A major goal of the project was to provide publicly available datasets to help improve diagnostic methods, treatment standards, and finally to prevent cancer.*

In our project we took advantage of publicly available dataset for glioma patients.

1.2 Glioma

It is also known as brain cancer. Since cancer occurs in glial cells which supports neurons this type of cancer is named as Glioma. There are four grades of **glioma**, and each has different types of cells present and different treatment strategies. A glioblastoma is the most aggressive form ranked as grade IV Glioma. Glioblastoma (World Health Organization grade IV) was the first cancer studied by TCGA in a pilot study. This program led to the development of important principles in biospecimen banking and collection, and the establishment of the highly organized infrastructure that served similar efforts in further studies.

1.3 RNA sequencing

RNA sequencing (RNAseq) is a technology used for profiling extracts information from RNA strands with high precision and high throughput. RNAseq is used to identify and quantify rare and common transcripts, isoforms, novel transcripts, gene fusions, and non-coding RNAs, among a wide range of samples in a rapid and efficient way [(5)]. For transcriptome analysis TCGA uses a platform based on the Illumina system. The TCGA deposited data contains information about both nucleotide sequence and gene expression.

1.4 t-test as an exploratory data analysis

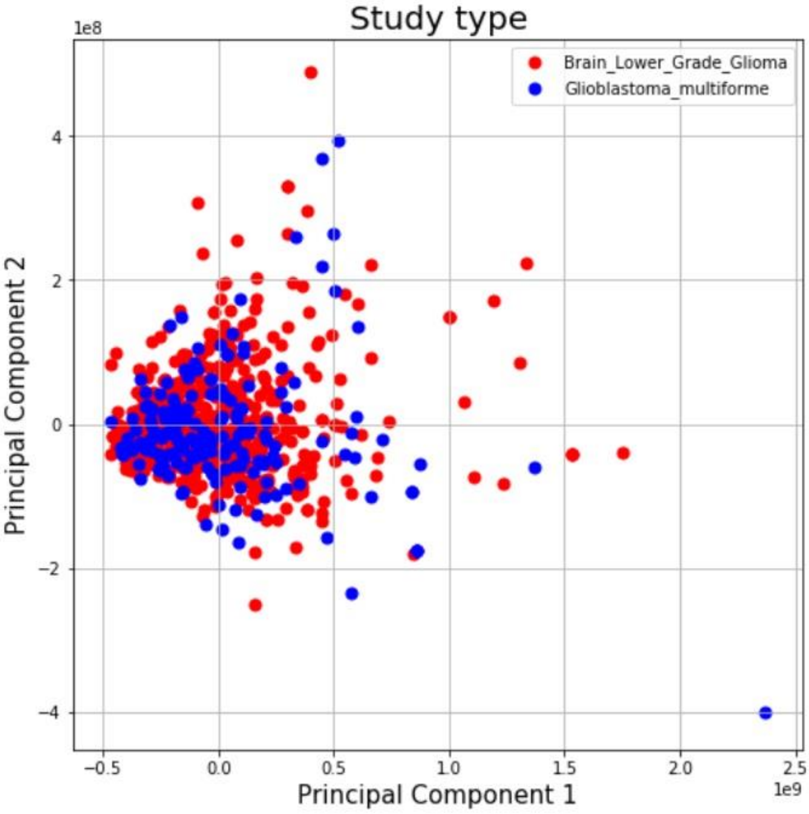
The t test compares two averages (means) of different data sets to clarify whether the sample datasets in concern are different from each other. It also provides information about significance levels of these differences. It measures it by using t values and every t value is accompanied by a p value which is a measure of the probability that the results from your sample data occurred by chance. Lower p-values are better and indicators of diisimilarities among data sets. There are used to identify thresholds of confidence interval.

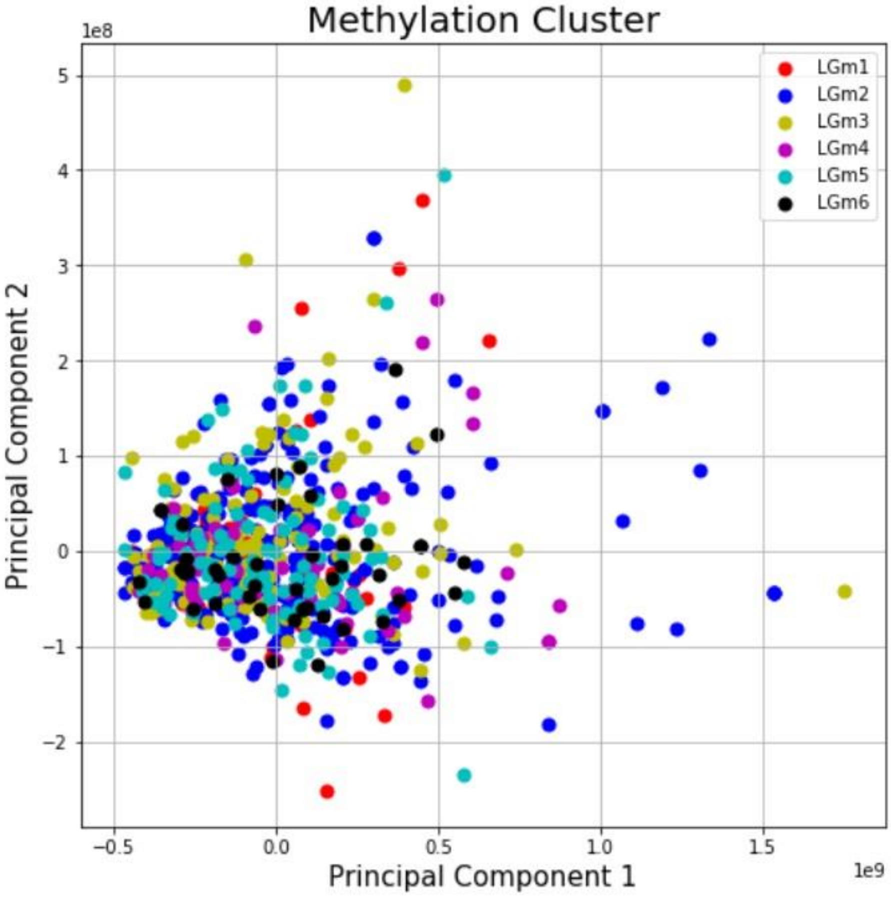
1.5 Regression as a predictive data analysis

Regression is a technique used to model and analyze the relationships between variables and often times how they contribute and are related to producing a particular outcome together. A linear regression refers to a regression model that is completely made up of linear variables. Beginning with the simple case, Single Variable Linear Regression is a technique used to model the relationship between a single input independent variable (feature variable) and an output dependent variable using a linear model i.e a line.

2. Data Exploration

Since we have more than five thousand gene expression values we wanted to minimize number of features we carried out a principle component analysis for gene expressions and derived the following graph below. (graph 1.1) We also clustered methylation data under two principle components in graph 1.2.


Study type
1e8
Brain_Lower_Grade_Glioma
Glioblastoma_multiforme
4
2
0
−2
−4
Principal Component 2
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
1e9
Principal Component 1


Methylation Cluster
1e8
LGm1
LGm2
LGm3
LGm4
LGm5
LGm6
5
4
3
2
1
0
−1
−2
Principal Component 2
−0.5
0.0
0.5
1.0
1.5
1e9
Principal Component 1

3. *Results*

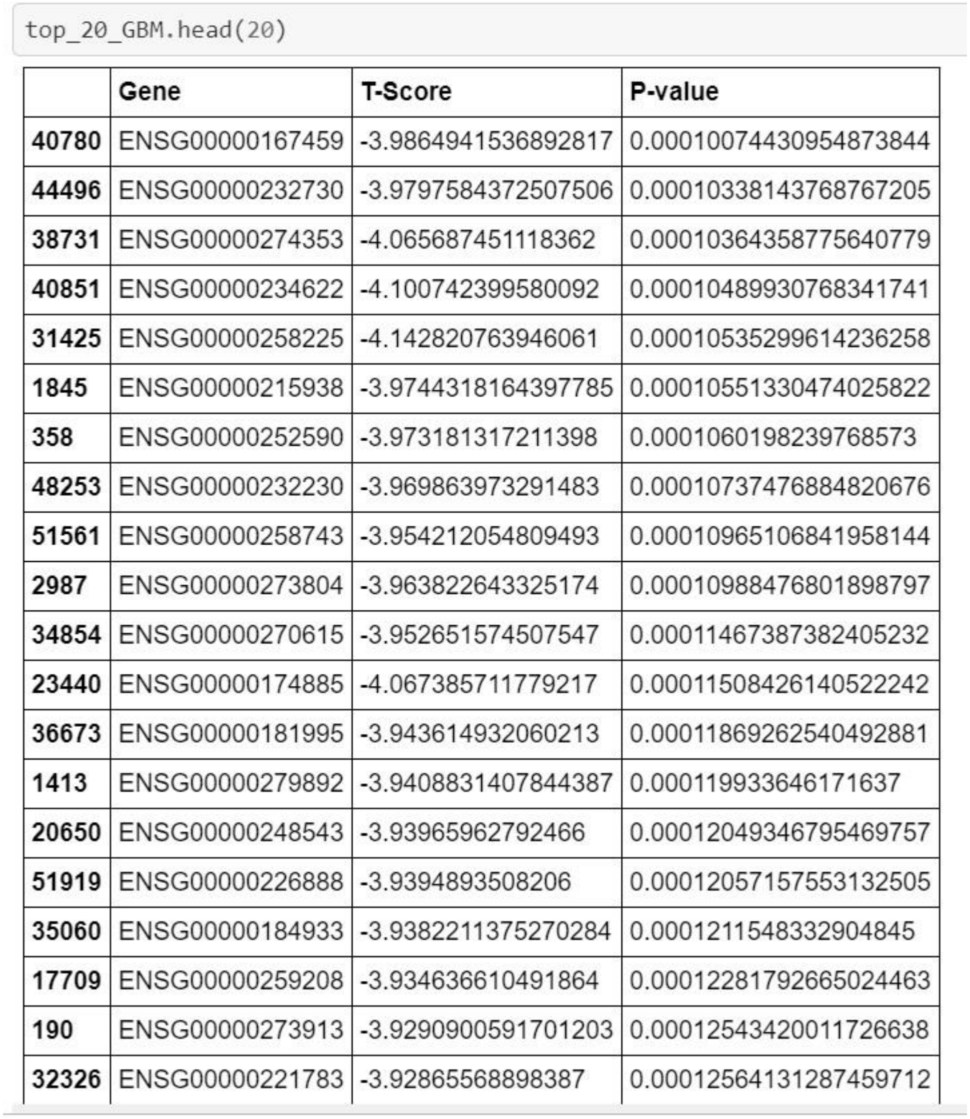
In our project we used data from TCGA project. We downloaded data from the following link: https://portal.gdc.cancer.gov/repository . And we carried out two exploratory data analysis and one regression analysis on the dataset. Exploratory data analysis covers applying t-test to normal and tumor cells and applying t-test on grade [2,3] and grade [4] cancer patients.

3.1 t-test for normal and tumor cells

First of all, we sort glioma patients under two categories LGG and GBM then we annotate samples by disease status (tumor positive and tumor negative). Both data sets have normal and tumor samples.

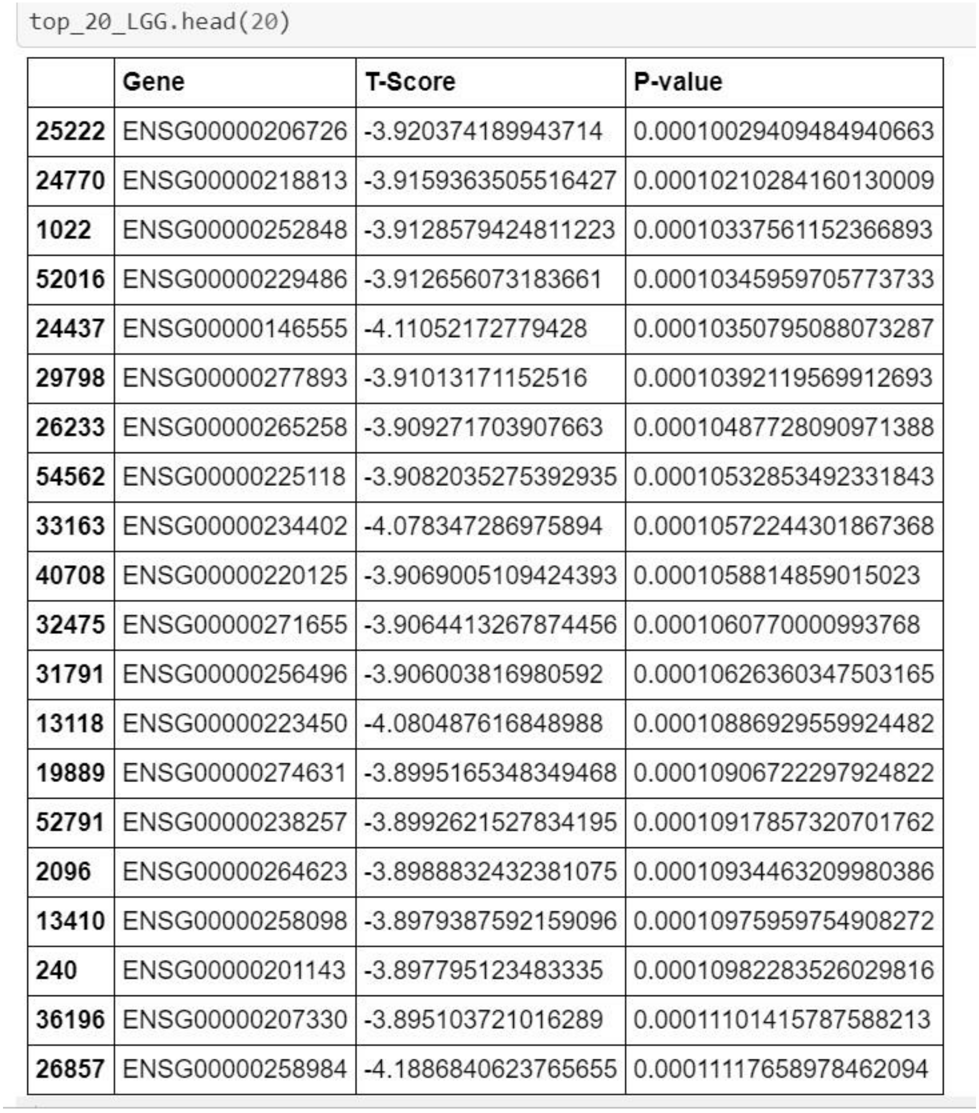
For each gene expression value, we compared two groups (normal, tumor) with t-test and we generated p-values to be able to find out which genes have higher expression values in other words to find out which genes would be responsible for deviation from normal case. Then we ordered genes in increasing order of p-values. First 20 genes with the lowest p-values are shown in Table 1.1

Table1.1



```
top_20_GBM.head(20)
```

	Gene	T-Score	P-value
40780	ENSG00000167459	-3.9864941536892817	0.00010074430954873844
44496	ENSG00000232730	-3.9797584372507506	0.00010338143768767205
38731	ENSG00000274353	-4.065687451118362	0.00010364358775640779
40851	ENSG00000234622	-4.100742399580092	0.00010489930768341741
31425	ENSG00000258225	-4.142820763946061	0.00010535299614236258
1845	ENSG00000215938	-3.9744318164397785	0.00010551330474025822
358	ENSG00000252590	-3.973181317211398	0.0001060198239768573
48253	ENSG00000232230	-3.969863973291483	0.00010737476884820676
51561	ENSG00000258743	-3.954212054809493	0.00010965106841958144
2987	ENSG00000273804	-3.963822643325174	0.00010988476801898797
34854	ENSG00000270615	-3.952651574507547	0.00011467387382405232
23440	ENSG00000174885	-4.067385711779217	0.00011508426140522242
36673	ENSG00000181995	-3.943614932060213	0.00011869262540492881
1413	ENSG00000279892	-3.9408831407844387	0.000119933646171637
20650	ENSG00000248543	-3.93965962792466	0.00012049346795469757
51919	ENSG00000226888	-3.9394893508206	0.00012057157553132505
35060	ENSG00000184933	-3.9382211375270284	0.0001211548332904845
17709	ENSG00000259208	-3.934636610491864	0.00012281792665024463
190	ENSG00000273913	-3.9290900591701203	0.00012543420011726638
32326	ENSG00000221783	-3.92865568898387	0.00012564131287459712



```
top_20_LGG.head(20)
```

	Gene	T-Score	P-value
25222	ENSG00000206726	-3.920374189943714	0.00010029409484940663
24770	ENSG00000218813	-3.9159363505516427	0.00010210284160130009
1022	ENSG00000252848	-3.9128579424811223	0.00010337561152366893
52016	ENSG00000229486	-3.912656073183661	0.00010345959705773733
24437	ENSG00000146555	-4.11052172779428	0.00010350795088073287
29798	ENSG00000277893	-3.91013171152516	0.00010392119569912693
26233	ENSG00000265258	-3.909271703907663	0.00010487728090971388
54562	ENSG00000225118	-3.9082035275392935	0.00010532853492331843
33163	ENSG00000234402	-4.078347286975894	0.00010572244301867368
40708	ENSG00000220125	-3.9069005109424393	0.0001058814859015023
32475	ENSG00000271655	-3.9064413267874456	0.0001060770000993768
31791	ENSG00000256496	-3.906003816980592	0.00010626360347503165
13118	ENSG00000223450	-4.080487616848988	0.00010886929559924482
19889	ENSG00000274631	-3.8995165348349468	0.00010906722297924822
52791	ENSG00000238257	-3.8992621527834195	0.00010917857320701762
2096	ENSG00000264623	-3.8988832432381075	0.00010934463209980386
13410	ENSG00000258098	-3.8979387592159096	0.00010975959754908272
240	ENSG00000201143	-3.897795123483335	0.00010982283526029816
36196	ENSG00000207330	-3.895103721016289	0.00011101415787588213
26857	ENSG00000258984	-4.1886840623765655	0.0001111765897846209 4

When we compare p values in a heat map for GBM data set, we get the following heat map in figure 1.1

Figure 1.1

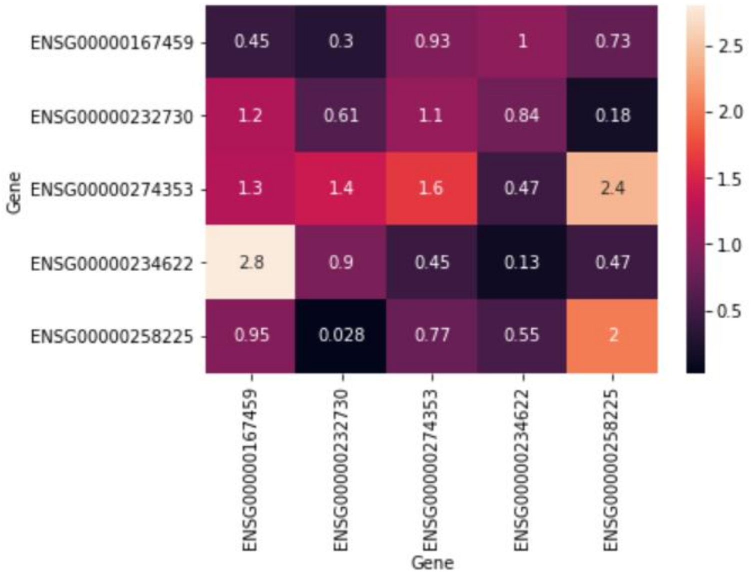


Figure 1.1

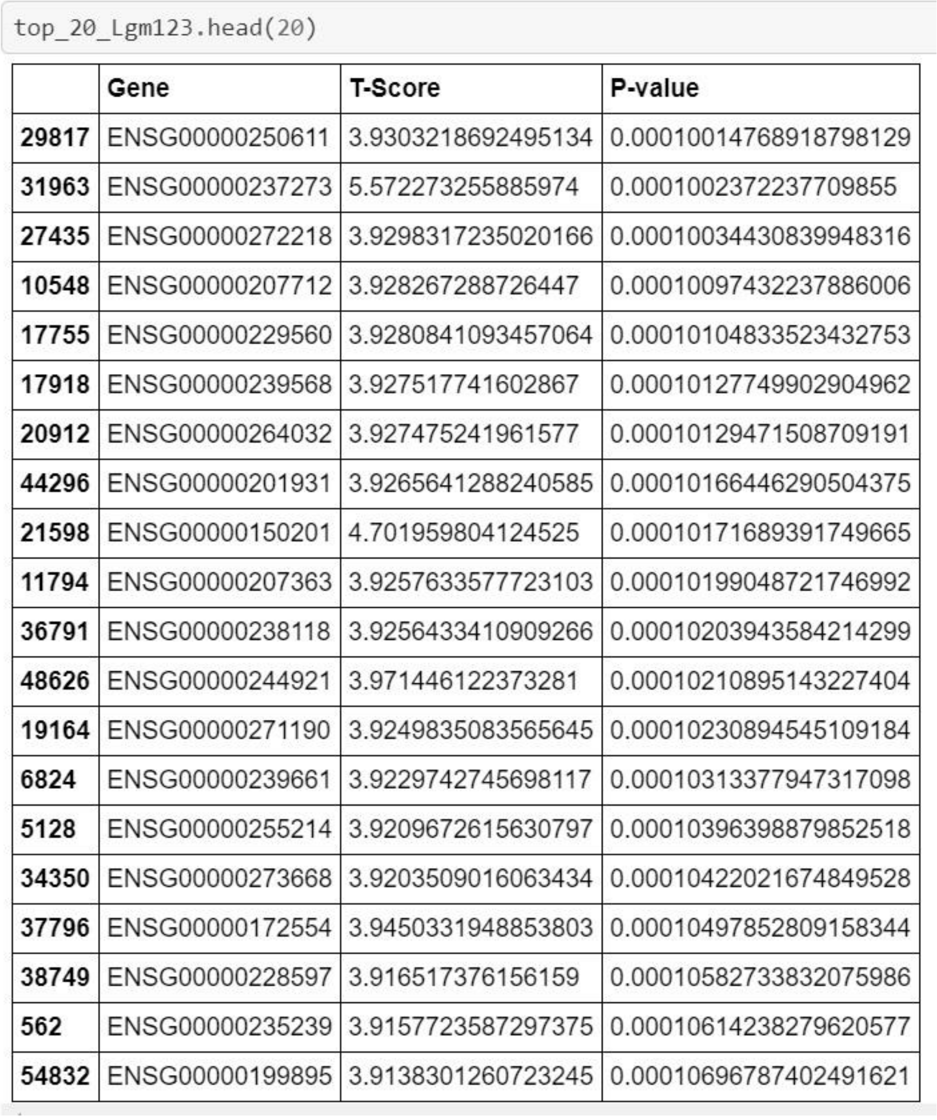
When we compare first 20 genes of LGG and GBM with lowest p- values for recurrent tumor and tumor cases, it is obviously seen that primary mutant genes of LGG tumor and recurrent tumor cases are totally different from primary mutant genes of LGG tumor and recurrent tumor cases. This indicates that causes of recurrent tumors are due to total new mutations rather than mutations responsible for first tumor cases.

When we scrutinize heat map for GBM cases we see that gene number 232730 is closely correlated with 258225. Besides that, 232730 is also closely correlated to 167459 and 234622 is closely correlated to 232730.

3.2 t- test for grade [2 and 3] and grade [4] tumor cells

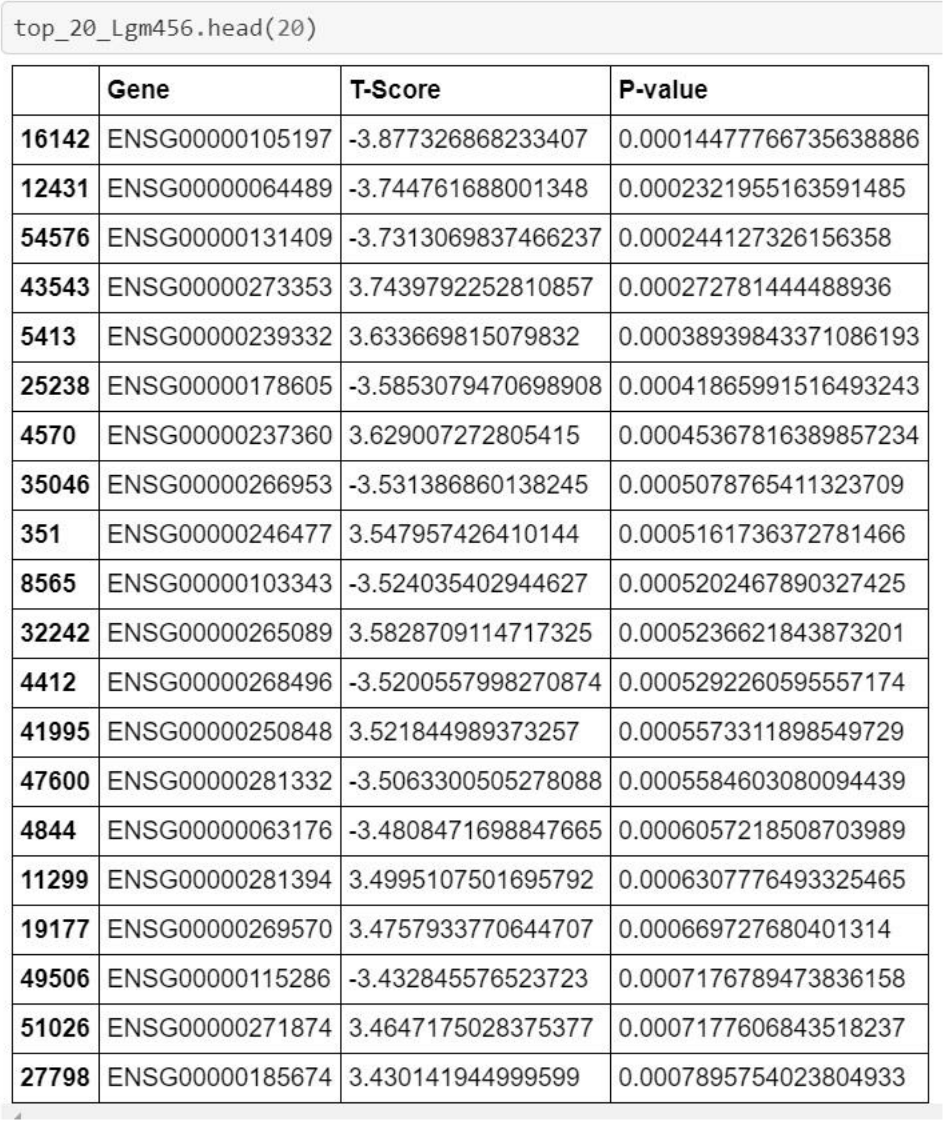
In the second part of our exploratory data analysis we divide glioma cases into two datasets LGM 123 and LGM 456 then we annotate samples by expression and methylation-based clusters.

For each gene expression value, we compared two groups (high grade, low grade) with t-test and we generated p-values to be able to find out which genes have higher expression values in other words to find out which genes would be responsible for deviation from normal case. Then we ordered genes in increasing order of p-values. First 20 genes with lowest p-values are shown in Table 1.2



```
top_20_Lgm123.head(20)
```

	Gene	T-Score	P-value
29817	ENSG00000250611	3.9303218692495134	0.00010014768918798129
31963	ENSG00000237273	5.572273255885974	0.0001002372237709855
27435	ENSG00000272218	3.9298317235020166	0.00010034430839948316
10548	ENSG00000207712	3.928267288726447	0.00010097432237886006
17755	ENSG00000229560	3.9280841093457064	0.00010104833523432753
17918	ENSG00000239568	3.927517741602867	0.00010127749902904962
20912	ENSG00000264032	3.927475241961577	0.00010129471508709191
44296	ENSG00000201931	3.9265641288240585	0.00010166446290504375
21598	ENSG00000150201	4.701959804124525	0.00010171689391749665
11794	ENSG00000207363	3.9257633577723103	0.00010199048721746992
36791	ENSG00000238118	3.9256433410909266	0.00010203943584214299
48626	ENSG00000244921	3.971446122373281	0.00010210895143227404
19164	ENSG00000271190	3.9249835083565645	0.00010230894545109184
6824	ENSG00000239661	3.9229742745698117	0.00010313377947317098
5128	ENSG00000255214	3.9209672615630797	0.00010396398879852518
34350	ENSG00000273668	3.9203509016063434	0.00010422021674849528
37796	ENSG00000172554	3.9450331948853803	0.00010497852809158344
38749	ENSG00000228597	3.916517376156159	0.00010582733832075986
562	ENSG00000235239	3.9157723587297375	0.00010614238279620577
54832	ENSG00000199895	3.9138301260723245	0.00010696787402491621



```
top_20_Lgm456.head(20)
```

	Gene	T-Score	P-value
16142	ENSG00000105197	-3.877326868233407	0.00014477766735638886
12431	ENSG00000064489	-3.744761688001348	0.0002321955163591485
54576	ENSG00000131409	-3.7313069837466237	0.000244127326156358
43543	ENSG00000273353	3.7439792252810857	0.000272781444488936
5413	ENSG00000239332	3.633669815079832	0.00038939843371086193
25238	ENSG00000178605	-3.5853079470698908	0.00041865991516493243
4570	ENSG00000237360	3.629007272805415	0.00045367816389857234
35046	ENSG00000266953	-3.531386860138245	0.0005078765411323709
351	ENSG00000246477	3.547957426410144	0.0005161736372781466
8565	ENSG00000103343	-3.524035402944627	0.0005202467890327425
32242	ENSG00000265089	3.5828709114717325	0.0005236621843873201
4412	ENSG00000268496	-3.5200557998270874	0.0005292260595557174
41995	ENSG00000250848	3.521844989373257	0.0005573311898549729
47600	ENSG00000281332	-3.5063300505278088	0.0005584603080094439
4844	ENSG00000063176	-3.4808471698847665	0.0006057218508703989
11299	ENSG00000281394	3.4995107501695792	0.0006307776493325465
19177	ENSG00000269570	3.4757933770644707	0.000669727680401314
49506	ENSG00000115286	-3.432845576523723	0.0007176789473836158
51026	ENSG00000271874	3.4647175028375377	0.0007177606843518237
27798	ENSG00000185674	3.430141944999599	0.0007895754023804933

Table 1.2.

When we compare p values in a heat map for the highgrade data set, we get the following heat map in figure 1.2

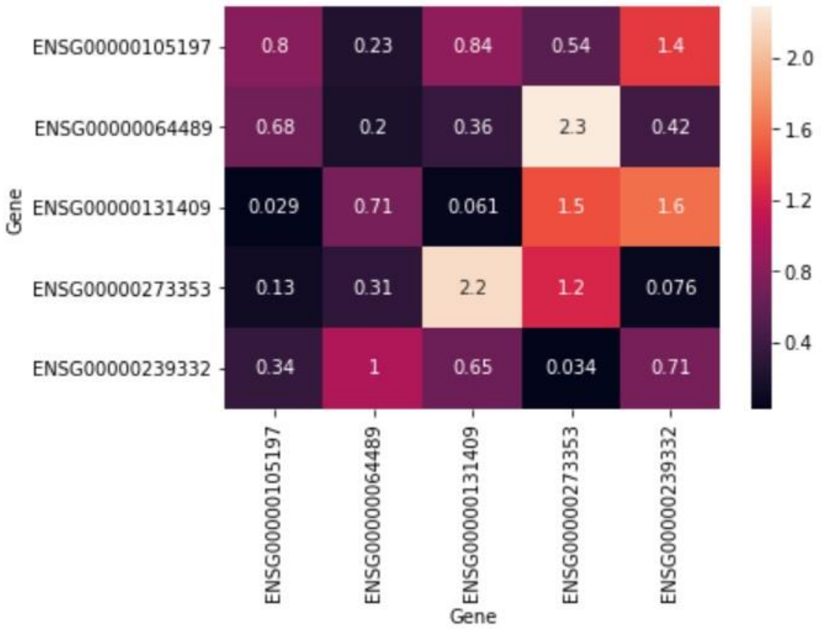


Figure 1.2

First 20 genes responsible for Lgg 123 doesn't match with first 20 genes of Lgg456. It seems that genes responsible for the two cases are totally independent of each other.

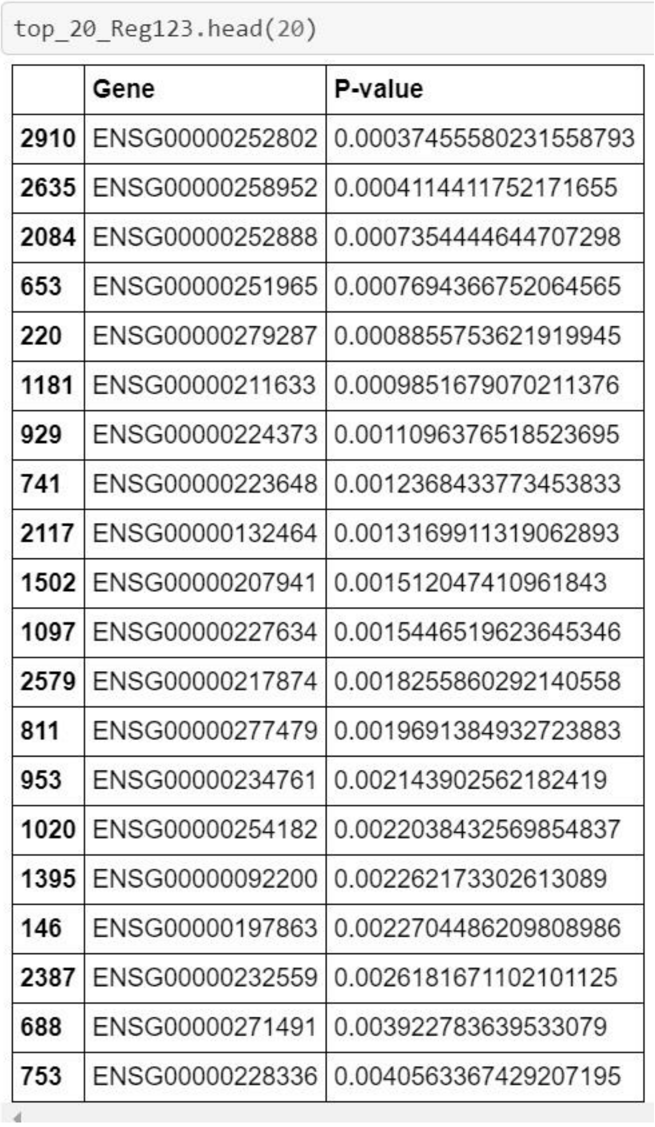
When we check out heat map it seems that some genes tend to work together such as gene 131409 tends to have some sort of correlation with gene 105197.

3.3 Regression analysis of age and gene expression values

In the third part of our study we annotated samples by expression and methylation based clusters. Each cluster had both LGG and GBM samples. Then we compared LGG and GBM samples within IDH-mut (LGm1-2-3) samples. After that we used linear to find genes that correlate with Patient Age for two subtypes.

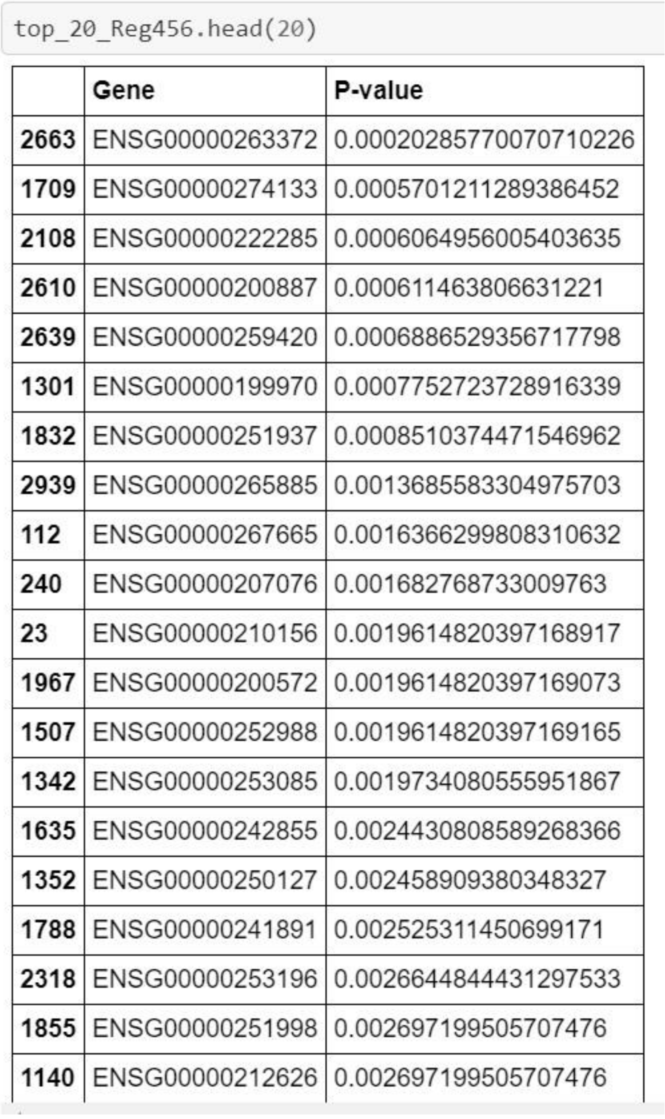
For each gene we carried out regression analysis while keeping age as independent variable and gene expression or methylation as dependent variable to find out whether there exist some sort of relationship between gene expression and age. We examined p-values in ascending order and took the first 20 genes with the lowest p-values. Which can be found in table 1.3.

Table 1.3



```
top_20_Reg123.head(20)
```

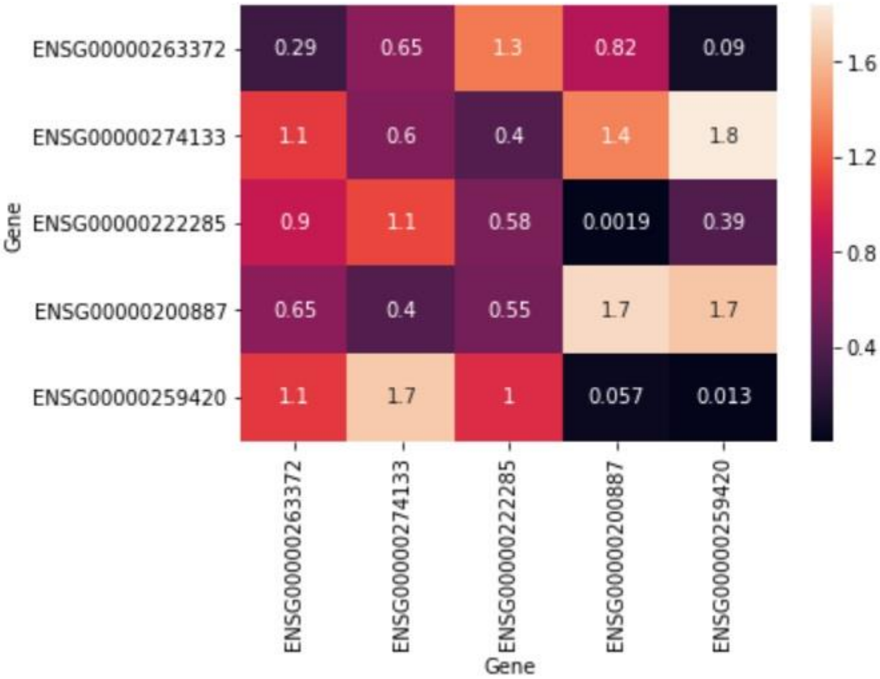
	Gene	P-value
2910	ENSG00000252802	0.00037455580231558793
2635	ENSG00000258952	0.0004114411752171655
2084	ENSG00000252888	0.0007354444644707298
653	ENSG00000251965	0.0007694366752064565
220	ENSG00000279287	0.0008855753621919945
1181	ENSG00000211633	0.0009851679070211376
929	ENSG00000224373	0.0011096376518523695
741	ENSG00000223648	0.0012368433773453833
2117	ENSG00000132464	0.0013169911319062893
1502	ENSG00000207941	0.001512047410961843
1097	ENSG00000227634	0.0015446519623645346
2579	ENSG00000217874	0.0018255860292140558
811	ENSG00000277479	0.0019691384932723883
953	ENSG00000234761	0.002143902562182419
1020	ENSG00000254182	0.0022038432569854837
1395	ENSG00000092200	0.002262173302613089
146	ENSG00000197863	0.0022704486209808986
2387	ENSG00000232559	0.0026181671102101125
688	ENSG00000271491	0.003922783639533079
753	ENSG00000228336	0.0040563367429207195



```
top_20_Reg456.head(20)
```

	Gene	P-value
2663	ENSG00000263372	0.00020285770070710226
1709	ENSG00000274133	0.0005701211289386452
2108	ENSG00000222285	0.0006064956005403635
2610	ENSG00000200887	0.000611463806631221
2639	ENSG00000259420	0.0006886529356717798
1301	ENSG00000199970	0.0007752723728916339
1832	ENSG00000251937	0.0008510374471546962
2939	ENSG00000265885	0.0013685583304975703
112	ENSG00000267665	0.0016366299808310632
240	ENSG00000207076	0.001682768733009763
23	ENSG00000210156	0.0019614820397168917
1967	ENSG00000200572	0.0019614820397169073
1507	ENSG00000252988	0.0019614820397169165
1342	ENSG00000253085	0.0019734080555951867
1635	ENSG00000242855	0.0024430808589268366
1352	ENSG00000250127	0.002458909380348327
1788	ENSG00000241891	0.002525311450699171
2318	ENSG00000253196	0.0026644844431297533
1855	ENSG00000251998	0.002697199505707476
1140	ENSG00000212626	0.002697199505707476

When we compare two tables there are various genes responsible for age related glioma cases.



After analysis of the we noticed that there are some genes which seems to work together such as 200887 seems to work together with 259420 and 222285. Besides that 259420 has close correlation with 200887 and 259420.

4. Discussion

We studied glioma case for primary tumor cases and recurrent tumor cases and found out that there are close relationships between certain genes and primary tumor cases and there are close relationships between recurrent tumor cases and certain genes however there is no correlation between genes responsible for recurrent tumors and primary tumors. Mechanisms behind new mutations would be studied.

In regression analysis we found out that there are more than 20 genes which increased their expression as people get older. Aging mechanism would be responsible for certain mutations as well.

Sources:

1. https://www.medicalnewstoday.com/articles/282929.php
2. https://www.cdc.gov/nchs/index.htm
3. Postgraduate School of Molecular Medicine, Medical University of Warsaw, Warsaw, Poland
4. Laboratory of Gene Therapy, Department of Cancer Immunology, The Greater Poland Cancer Centre, Poznan, Poland
5. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009; 10: 57-63.