Int J Clin Exp Med 2017;10(11):15632-15643
www.ijcem.com /ISSN:1940-5901/IJCEM0044755

Original Article

Identification of the glioma's related genes with a biological feature-based classification

Yitian Chen[1*], Lilei Peng[2*], Yang Ming[2], Ligang Chen[2]

[1]*Department of Neurosurgery, Medical College of Soochow University, Suzhou 215123, China;* [2]*Department of Neurosurgery, The Affiliated Hospital of Southwest Medical University, Luzhou 646000, China.* **Equal contributors.*

Received March 22, 2016; Accepted December 23, 2016; Epub November 15, 2017; Published November 30, 2017

Abstract: Gliomas are the most common primary intracranial tumors and the most aggressive type of brain malignancy of adults. This study aimed to employ an effective computational method to identify glioma-related biological mechanisms and predict new glioma genes through the fully utilization of the information of glioma genes. In terms of molecular features, the known glioma genes were fetched from database and literature mining, and encoded by the enrichment scores of gene ontology and pathways. With the Random Forest classification and incremental feature selection, the optimal features of the selection of the glioma's related genes were found. Random Forest classification was also used to predict novel glioma genes. The shortest path analysis, based on the gene interaction networks, was performed to identify the genes that have links with known genes. For the depiction of the glioma genes, 3318 gene ontology terms and 127 pathway terms were identified as the optimal features. 860 novel related genes were predicted based on those terms. 87 genes were identified, which reside in the hub of known genes interaction network. There were an intersection of 34 genes between predicted genes and shortest path genes. 25 out of 34 genes showed significant different expression between glioma and normal tissues, and highly possibility of being the candidate of glioma's related genes showed in most of them. Our proposed algorithm has a distinguished power to predict genes that are closely related to the glioma and provides the gene list to help achieve early detection.

Keywords: Glioma, biomarker prediction, genes interaction network, geneontology

Introduction

Gliomas are central nervous system neoplasms derived from glial cells which act as supportive cells in the central nervous system. Glioma makes up about 30% of all brain and central nervous system tumors and 80% of all malignant brain tumors [1]. Glioma is the most common and aggressive brain malignancy threating adults [2]. At any one of its stages of development, new abnormal neuroglial tissue grows through excessive cellular division and more rapidly proliferation, and then continues to grow after the stimuli initiated the new growth cease. The average lifespan of patients who suffer glioma is less than one year from the time of the diagnosis, even though the improvement in therapeutic interventions is significant, minimal improvement of it over the past 25 years [3]. Dismal has remined in the prognosis for most glioma patients regardless of the advance in clinical techniques [4]. Thus, it is crucial to elucidate the genetic factors of glioma, and contributes to the diagnosis and prognosis of glioma patients.

The incidence of glioma is not significantly affected by environmental factors such as UV light and carcinogen exposure because of the protective influence of the thick skull and the blood-brain barrier. In addition, there are unknown heritable factors that may cause glioma. Among the people, these tumors are seemed to have idiopathic occurrence in a random manner [5]. Therefore, the cellular mechanisms giving rise to glioma are not very clear yet. LOH 10q (over 70%), EGFR amplification (about 40%), MDM2 amplification, LOH 10p, 10q, and p16INK4a and PTEN mutation are the most common molecular alterations in the primary period of glioma [6]. Besides, the mutation of IDH1, TP53, and LOH on 17p, 10q, and

19q [3] as the first common molecular event in multistep carcinogenesis in the secondary period of glioma. IDH1/2 (isocitrate dehydrogenase 1/2) mutation and MGMT (O6-methylguanine-DNA methyltransferase) promoter methylation included in current molecular prognostic markers, which provide the improved prognosis and relative sensitivity to temozolomide treatment respectively [7].

The Gene Ontology (GO) is a database aim to unify the representation of genes and gene product features in all species [8]. In terms of encoding genes and updating continuously under explorations, it is an effective and efficient tool. GO annotations have been demonstrated to be excellent predictors of cancer genes. The Kyoto Encyclopedia of Genes and Genomes (KEGG) database used widely and created from published materials manually [9]. KEGG pathways elucidate in vivo comprehensive inferences of reactions. It provides pathway maps for metabolism and other cellular processes, human diseases also included.

It is desirable of system biology approaches for analysis of diseases t mechanisms. For the identification of glioma-related biological mechanisms and predict new glioma genes, a system biological measure was developed, that encode glioma genes through integrating gene ontology (GO) and KEGG annotations as features. Genes involved in glioma were well characterized and predicted with optimal features we analyzed. Predicted glioma's related gene as the candidates would help to promote the research progress of potential prognostic biomarkers, and new molecular drug targets aims to treat this devastating disease.

Material and methods

Datasets

Glioma's related genes searched from OMIM (Online Mendelian Inheritance in Man), GAD (The Genetic Association Database) and DisGeNET. OMIM [10] is a comprehensive database concerning human beings' genes and hereditary diseases. With the searching key word of "glioma", seven genes had been found in that database. COSMIC [11] fetches cancer's related genes in the reported references and the high flux experimental data of Sanger laboratory's cancer genome project, and 19 genes had been found with the searching key word of "glioma".

GAD is a comprehensive database collecting human beings' complex diseases, complex diseases' pathogenic genes in reported references and GWAS experimental data included in its collection. 12 genes were obtained with the searching key word of "glioma" in GAD. DisGeNET [12] annotates pathogenic genes by integrating public databases and gene-disease relation in reported references. Currently, there are 381056 gene-disease relations DisGeNET, including 16666 genes and 13172 diseases, and 39 genes had been found with the searching key word of "glioma". 62 different genes had been found in those databases, and the specific names and sources are shown in Supplementary Table 1. We had 143 non-repetitive genes related to glioma in databases and references in total.

Many databases are developed to collect pathogenic gene as mentioned above, however, the collection might not be very comprehensive owing to different kinds of data in different databases that covers various data of diseases phenotypes and genotypes, and lag of database maintenance. Therefore, it is necessary of manual screen for further analyze the pathogenic gene of specific disease phenotype. Pubmed's searching tool was used to examine the pathogenic relation between gene and glioma. "Gene symbol" or "gene" and "glioma" was used as searching key words, and if those two words both showed in the title and abstract of an article, the article would be recorded as the evidence to verify that gene is related to glioma. 81 pathogenic genes found in references in total. The specific names of those genes are also shown in Supplementary Table 1.

Encoding glioma's related genes

GO analysis, a well-known biological information analysis tool based on definition of GO terms to label the features of all species' gene products. KEGG is a comprehensive database based on known molecule interaction network, the analyses of biological pathway and systematic information was included [13]. So, we used GO terms and KEGG pathways to encode gene. The relation between gene and its feature terms can be reflected by the enrichment information of GO and pathways analysis.

Considering one gene and its directly interacting partners [14] in STRING network, gene's gene ontology enrichment score is defined as its -log10 of *P* value for examining hyper geometric test. Higher enrichment score indicates higher degree of enrichment. Thus, a gene encoded as a one-dimensional vector containing 6242 GO terms and 214 KEGG pathways.

Removing irrelevant features

The association between two variables measured by Cramer's coefficient [15]. The coefficient value of Cramer's is 0-1. Higher coefficient value of two variables' Cramer's means higher correlation of them, vice versa. 0.1 was taken as threshold value in this essay, the feature that Cramer's coefficient is lower than 0.1 was excluded.

Screening the optimal feature

This research used some feature selection approaches to identify key GO terms and KEGG pathways, including minimum redundancy maximum relevance (mRMR), incremental feature selection (IFS) [15], Random Forest (RF) algorithm [16].

Specifically, the screened features, sorted through mRMR the key features in the feature list fetched through IFS and RF with the help of mRMR. The mRMR, created by Peng, et al. [15], has two criteria: Max-Relevance and Min-Redundancy.

The IFS based on the order made by mRMR feature list. The feature was added one by one in the process of analysis. Every time a new feature added, a new sample subset of positive samples and negative samples, which based on the feature we selected, it would be built and we would examine and evaluate every subset of data.

Weka 3.6.4 [17] with the default parameter was used to carry out the classification analysis of Random Forest algorithm. And Ten-fold cross validation is used to study the performance of classification model. The testing performance evaluation is based on the Matthews's correlation coefficient (MCC).

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Where the TP indicates the rate of true positive, the FP indicates the false positive, the TN indicates the true negative and the FN indicates the false negative.

Thus, we could get the optimal feature subset which is the subset with the maximum MCC value firstly appearing. Apart from that, the IFS curve of MCC value was drawn.

Gene network and shortest path trace

The weighing gene interaction network could get from STRING [18]. STRING has a grading mechanism to weigh the results of different methods and provides a comprehensive grade. The score of its dependency which is the possibility of interaction represented by every straight line in this network.

Therefore, we can obtain the relevance network of a functional protein and the connection weight. We searched for the shortest path of every pair of known glioma genes in the graph with the calculate the shortest path of one node to all the other nodes accord to Dijkstra's algorithm. Selection of all existed shortest path genes, and arranged these of them in accordance with their Betweenness value. The Betweenness value indicates the amount of shortest path of these genes as inner nodes in the known gliomas related genes links.

Identifying significant shortest path genes based on betweenness and permutation

Those genes with high Betweennes value have higher dependency with glioma genes compared with those genes of low Betweennes value. If the Betweennes value of protein is 0, it would be treated as a gene without dependency with glioma genes.

We used Permutation test to further screen shortest path genes to get rid of effects of the network structure. To calculate the shortest path of these genes, the same amount of genes were selected randomly as the related genes for 500 times. When the real Betweennes value of shortest path gene is less than the Betweennes value after substitution, we calculated once. After 500 random tests, we got a frequency which was identified as the permutation FDR of shortest path gene. The shortest path gene who's FDR less than 0.05 can be the glioma related gene.

Table 1. The sizes of 10 datasets and the corresponding number of optimal features for predicting glioma-related genes

Dataset	Rest features number	Optimal feature number	Sn	Sp	Acc	Mcc
1	4124	1041	0.930769	0.842105	0.903743	0.772875
2	5118	1379	0.976923	0.921569	0.961326	0.903962
3	4342	1193	0.984615	0.684211	0.893048	0.744139
4	4209	1451	0.953846	0.736842	0.887701	0.727886
5	4291	288	0.923077	0.842105	0.898396	0.761492
6	4616	820	0.969231	0.842105	0.930481	0.833569
7	4400	1787	0.976923	0.736842	0.903743	0.768624
8	4603	426	0.961539	0.877193	0.935829	0.847352
9	4665	385	0.961539	0.824561	0.919786	0.807641
10	4869	26	0.976923	0.894737	0.951872	0.885388

Note. Sn: sensitivity; Sp: specificity; Acc: accuracy; MCC: Matthews's correlation coefficient.

Tissue total RNA extraction and quantitative RT-PCR

Glioma tissue (12 cases) were obtained from surgical resection and confirmed by pathology, and normal brain tissue were harvest from patient with decompression in traumatic brain injury. Total RNA were extracted with TRIzol reagent (Life technology), and cDNA were obtained by a reverse transcription of RNA. For all the predicted genes quantitative RT-PCR (Q-PCR) was performed using cDNA using the ABI PRISM 7900 system (Applied Biosystems) with the SYBR Green Realtime PCR Master Mix plus (TOYOBO). The detailed primer sequences were available in Supplementary Table 2.

Ethics statement

All the patients who participated in this study provided the informed consent, and the research was approved by the ethics committee at The Affiliated Hospital of Southwest Medical University, Sichuan, China.

Statistical analysis

The packages and functions in R software were used to do the statistical analysis. The function "phyper" was used to obtain *P* value for examining hyper geometric test in calculating GO and KEGG enrichment scores. The function "shortest path" in the igraph package was used to achieve the Dijkstra's algorithm to calculate the shortest path. The significantly different expression genes were identified with t-test and the significance level was set at P < 0.05.

Results

Describing the key features of glioma related gene

As mentioned in the chapter of "Dataset", these 143 genes were regarded as positive samples (glioma-related genes, Supplementary Table 1) in this study, while 143 × 40 = 5720 background genes in the Ensemble database were randomly selected as the negative samples (non-glioma-related genes, data not shown). To release the imbalance, the negative samples were split into 10 groups each of which were mixed with the positive sample, constructing 10 datasets (S1 to S10) with sample's classified labels and separately calculating the Cramer's value of them, Cramer's value less than 0.1 were excluded and other features were retained to be further selected. The amount of every dataset's rest features is presented in **Table 1**.

IFS, mRMR and Random Forest algorithm were used select the optimal feature of every dataset's remaining features. The significance of every dataset's feature was performed in mRMR analysis, and then every dataset would return MaxRel feature list and mRMR feature list in the light of the chapter "Screening the Optimal Feature".

IFS and RF structure dataset and classify were based on the feature order of mRMR feature list, and we evaluated classification result by Ten-fold cross validation. SNs, SPs, ACCs and MCCs of 10 datasets are presented in **Table 1**. We draw the IFS curve of every dataset to bet-

Table 2. The top 20 optimal features of the union of 10 datasets

Order	Features	Name
1	GO: 0016035	Zeta DNA polymerase complex
2	GO: 0009314	Response to radiation
3	GO: 0008329	Signaling pattern recognition receptor activity
4	GO: 0016829	Lyase activity
5	GO: 0010243	Response to organonitrogen compound
6	GO: 1900029	Positive regulation of ruffle assembly
7	GO: 0009636	Response to toxic substance
8	GO: 0031424	Keratinization
9	GO: 0048147	Negative regulation of fibroblast proliferation
10	GO: 0090399	Replicative senescence
11	GO: 0042743	Hydrogen peroxide metabolic process
12	GO: 0071158	Positive regulation of cell cycle arrest
13	GO: 0006886	Intracellular protein transport
14	GO: 0030868	Smooth endoplasmic reticulum membrane
15	GO: 0009374	Biotin binding
16	GO: 0048702	Embryonic neurocranium morphogenesis
17	GO: 0050658	RNA transport
18	GO: 0055038	Recycling endosome membrane
19	GO: 0043120	Tumor necrosis factor binding
20	GO: 0008340	Determination of adult lifespan

ter observe (the entire IFS curves are showed in Supplementary Figures 1, 2, 3, 4, 5, 6, 7, 8, 9, 10). Therefore, these 10 optimal feature sets (OS1, OS2 ...... OS10) could get from the first 1041, 1379, 1193, 1451, 288, 820, 1787, 426, 385 and 26 features in mRMR feature list of corresponding datasets. We calculated the union of these sets, and got a new dataset OS that is called the optimal feature list, including 3318 GO terms and 127 KEGG pathway (the top 20 optimal features are presented in **Table 2** and the entire optimal features are presented in Supplementary Table 3).

GO terms include three main types: biological process, cell component and molecule function. To generally illustrate these optimal features in GO terms, we divided these features into biological process, cell constituent and molecule function to describe the optimal GO terms, thus demonstrating the feature of glioma's related gene (**Figure 1**).

Predicting glioma's related gene

Glioma's related gene prediction depended on the optimal feature that can be defined as the key features of glioma genes. Those genes that are almost the same with the known genes in the terms of screened optimal features can be candidate gene of glioma, for these genes may have the same function with known glioma genes. We predicted 988 glioma related genes from the annotated genes in Ensemble database by Random Forest algorithm, including 860 novel glioma related genes in addition to the achieved glioma genes (Supplementary Table 4). We held the view that these genes may also affect the growth of glioma or relate to its development.

Genes that interact with known glioma gene and short path gene

Relationship of gene interaction in the STRING database was investigated to find out the hub genes related to glioma. As "guilt by association" rule, two interactional genes have the same or similar function in organism and take part in common pathways. In gene interaction network, we could forecast glioma gene on the basis of interaction relation in STRING database, study those genes that interacted with known glioma gene, structure interactional sub-networks to look for the core gene, and these genes could also be glioma's candidate and hub genes.

We searched genes that are connected with the 143 known glioma related genes in the shortest path and calculated this inner node's Betweenness value in the path, and then we got 345 short path genes whose Betweenness values are bigger than 0, which are listed in Supplementary Table 5. To further screen these genes, we used Permutation test and calculated their permutation FDR, and the results are also listed in Supplementary Table 5. The 87 genes whose permutations FDRs are less than 0.05 have high dependency with glioma (**Table 3**).

These 87 short path genes coincide with 34 genes that found through optimal feature (**Table 3**). Therefore, we used two approaches to identify 34 genes that not only had similar

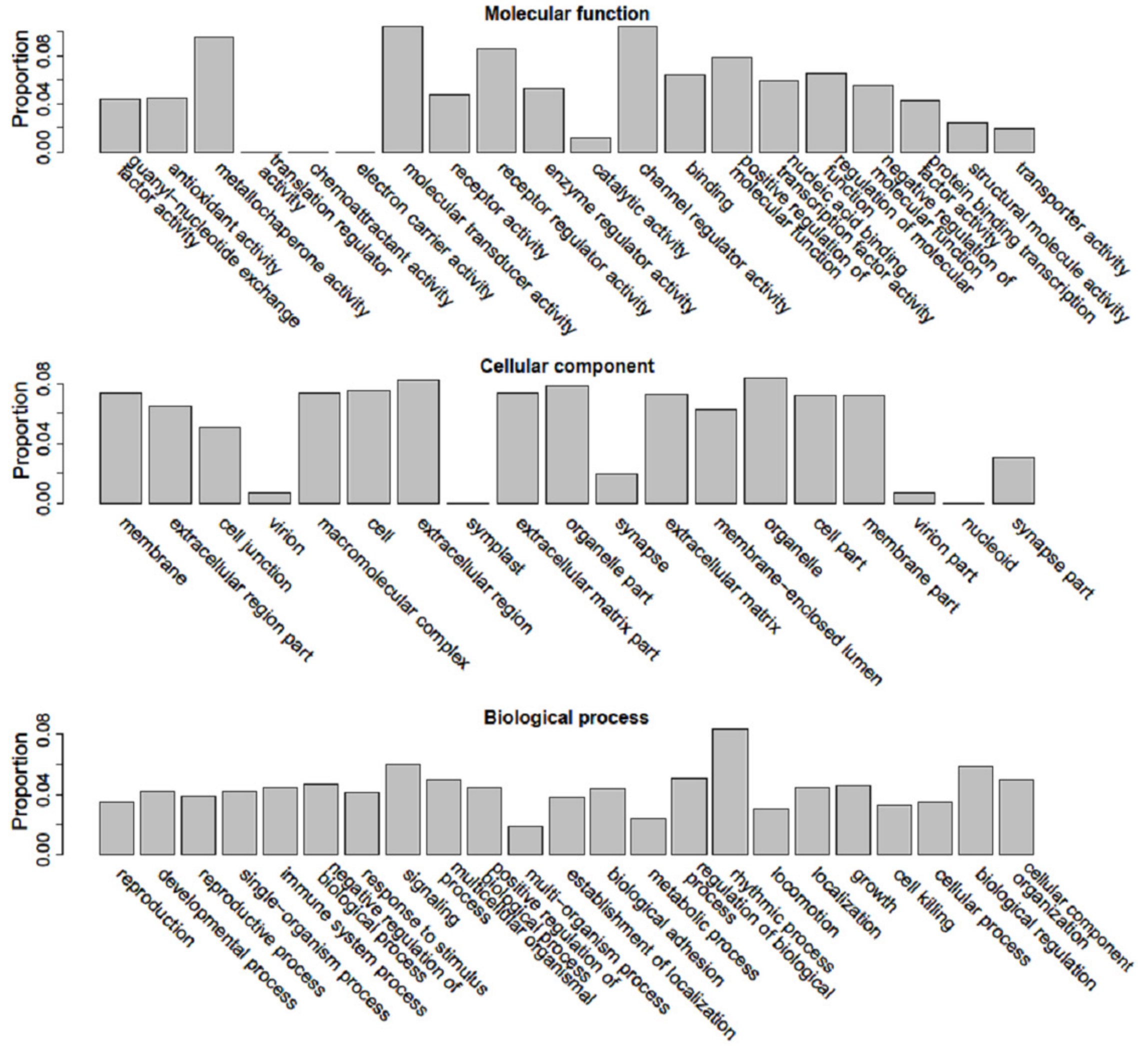


Figure 1. The profile of optimal GO terms for predicting glioma-related genes.

genetic annotation with known glioma genes but also resided in the core of gene interaction in the network of known glioma genes. These genes seemed to be the more reliable glioma related genes that are worthy of further research.

Comparing the expression of the identified genes in glioma and normal brain tissues

Derived from data matrix dataset in TCGA (The Cancer Genome Atlas) by selecting Expression-Genes at data type, level 3 at data level and Agilent 244K Custom Gene Expression G4502A-07 at Center/Platform, the publicly available expression data were employed to find the significantly expressed genes in the glioblastoma multiform. There were 20 genes showing significantly different expression in the TCGA expression data among the 34 genes found by two approaches.

With quantitative RT-PCR, we compared the expression of these genes between glioma tissues and normal brain tissues to further verify whether these overlapped genes are really glioma related genes. 25 genes out of 34 genes, 5 significantly down-regulating genes and 20 significantly up-regulating genes identified through two approaches in glioma tissues (**Figure 2**), had expression changes and would indeed participate in the occurrence and development of glioma, suggesting effective and reliable system biological measure for identifying glioma-related genes. 3 genes in the down-regulating genes are associated with tricarboxylic acid

Table 3. The 87 significant shortest path genes and the 34 overlaps with predicted genes through optimal features

Symbol	Overlap	Betweenness	*P* value	Symbol	Overlap	Betweenness	*P* value
TP53		1945	0.014925373	NA		129	0.004975
PCNA	YES	986	0.039800995	SDHA		129	0.00995
NCOA1	YES	629	0.004975124	SUFU		129	0.00995
BRCA1		512	0.019900498	SPI1		129	0.024876
RAD51		405	0.004975124	TIMP1	YES	129	0.024876
STAT6		379	0.004975124	IL10RA		129	0.00995
UQCRFS1		372	0.004975124	ATP1A1		129	0.00995
PRKDC	YES	372	0.004975124	RET	YES	129	0.034826
RIPK1	YES	270	0.034825871	RPA3	YES	129	0.004975
XPA	YES	266	0.004975124	RAD51C	YES	129	0.004975
RAD23B		265	0.004975124	CXCR4	YES	129	0.014925
GJA1	YES	255	0.009950249	UBQLN4		129	0.00995
IL4R		255	0.004975124	TNFRSF6B		128	0.034826
SOD1		254	0.009950249	MTHFR		128	0.024876
SDHB		254	0.019900498	SFTPD		128	0.024876
MAP3K5	YES	250	0.024875622	PLEK	YES	128	0.039801
MSH2		248	0.004975124	ZBTB33		128	0.00995
PARP1		240	0.039800995	ADAM22		128	0.029851
CDKN1A	YES	160	0.049751244	TOP2A		128	0.014925
CDK7		155	0.019900498	EDNRB	YES	128	0.0199
ERCC1		144	0.004975124	GLI1		127	0.00995
MDH2		136	0.004975124	GLTSCR2		126	0.0199
CASP8		134	0.024875622	ERCC8		125	0.034826
ACO2	YES	133	0.004975124	TG		124	0.00995
CS	YES	133	0.004975124	POLB	YES	121	0.029851
CDK6	YES	131	0.019900498	ERCC3	YES	113	0.0199
PDGFRB	YES	131	0.029850746	CFLAR		85	0.024876
GSTM2	YES	130	0.004975124	CCNE1	YES	82	0.014925
TERF2		129	0.029850746	NFE2L2	YES	75	0.034826
GPC1		129	0.004975124	TBXA2R		43	0.024876
PUF60		129	0.009950249	MRE11A	YES	41	0.014925
ATXN1		129	0.014925373	NR3C1	YES	39	0.039801
CSF1R		129	0.014925373	FH		9	0.004975
CD44	YES	129	0.009950249	GOT2	YES	7	0.039801
ALAS1		129	0.004975124	ERCC5		5	0.004975
APOA2		129	0.004975124	TPO		5	0.014925
IL23A	YES	129	0.004975124	MKI67		4	0.024876
RAD51D		129	0.004975124	GSTA2	YES	3	0.044776
MUS81	YES	129	0.004975124	LDHB		2	0.004975
PLA2G1B	YES	129	0.024875622	ANAPC5		1	0.00995
CAT	YES	129	0.024875622	NPSR1		1	0.014925
MLH1	YES	129	0.014925373	IDH3G		1	0.0199
SDC4		129	0.004975124	IDH3A		1	0.029851
TPT1		129	0.009950249				

Note. The overlap column indicates whether the shortest genes have overlaps with the gene list derived from the prediction based on the optimal features.

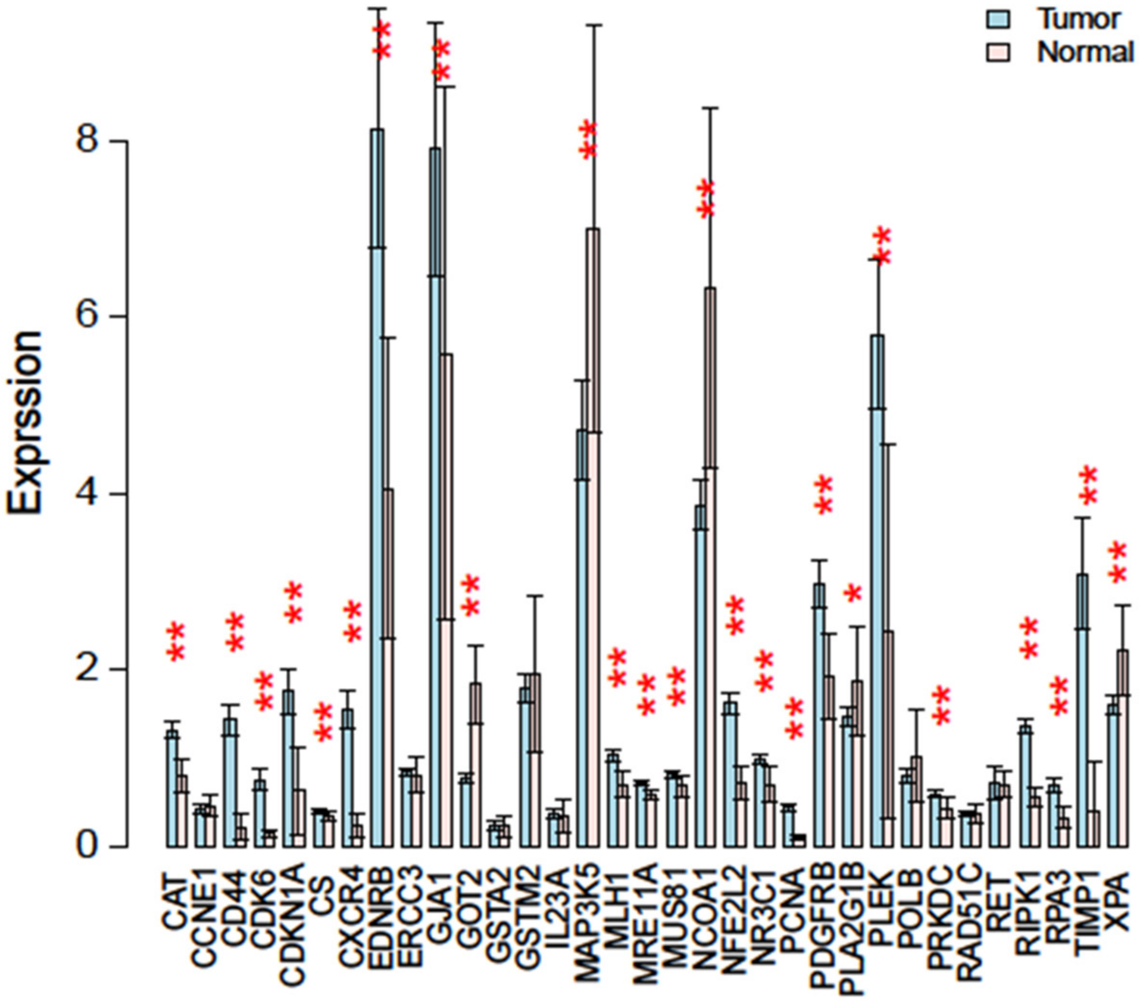


Figure 2. Gene expression verification of the 34 identified genes with Q-PCR. Data are presented as mean ± SD, n = 12, *P ≤ 0.05, ***P ≤ 0.01.

cycle while 13 genes among the predicted up-regulating genes are related with immune, cell cycle and proliferation, and most of the others are receptor-dependent protein kinase coding genes which are also related with cell growth and proliferation.

Discussion

In view of the GO structure, we first classified GO terms in the optimal list into three types: biological process (BP), cellular component (CC), and molecular function (MF) so as to illustrate the biological meanings of the selected optimal feature subset. The GO terms in the mRMR feature list were mapped to the children of the three root GO terms.

Biological process GO terms

As for the percentage of BP terms, the top five GO biological processes are GO: 0048511: rhythmic process (31.8%), GO: 0023052: signaling (22.9%), GO: 0065007: biological regulation (22.4%), GO: 0050789: regulation of biological process (19.2%) and GO: 0071840: cellular component organization or biogenesis (19.0%). Many cancer cells are rhythmic, i.e., key gene products serving in the circadian clock are important targets for manipulating cancer growth [19]. The relationship between the circadian timing system and cancer is very obvious from studies linking disruption of circadian rhythms with higher cancer risk and greater malignancy. Specifically, circadian genes are proved to be very important in regulating glioma proliferation, migration and apoptosis [20, 21]. The terms signaling, biological regulation and regulation of biological process are also consistent with the common knowledge that glioma-driving mutations affect pathways regulating cellular parameters such as cell growth, apoptosis, migration, and angiogenesis [22]. For instance, master regulators of several core biological processes frequently mutate in glioma are TP53, PTEN, NF1, or EGFR and Notch signaling, and are deregulated in malignant brain tumors as well [23].

Cellular component GO terms

Top eight GO cellular component terms of CC terms percentage are GO: 0043226: organelle (17.9%), GO: 0005576: extracellular region (17.6%), GO: 0044422: organelle part (16.7%), GO: 0005623: cell (16.1%), GO: 0016020: membrane (15.8%), GO: 0044420: extracellular matrix component (15.8%), GO: 0032991: macromolecular complex (15.7%) and GO: 0031012: extracellular matrix (15.6%). Many behavioral patterns of cells could be linked to the extracellular matrix including cancerous processes [24]. Extracellular region, matrix and matrix component provide the structural environment supporting cell adhesion and migration [25]. For example, extracellular matrix glycoprotein-derived synthetic peptides could differentially modulate glioma cell migration [26]. Similarly, high levels of the extracellular matrix glycoprotein TN-C were found in cancerous tissues and were directly linked to enhanced cell migration [27].

Molecular function G0 terms

The top five GO molecular function terms of percentage are GO: 0016247: channel regulator activity (36.4%), GO: 0060089: molecular transducer activity (36.2%), GO: 0016530: metallochaperone activity (33.3%), GO: 0030545: receptor regulator activity (30.0%) and GO: 0044093: positive regulation of molecular function (27.4%). The highlight of channel regulator activity may be attributed to the surprising fact that ion channel mutations are frequent with 90% of human glioma samples presenting with ion channel and transporter mutations [28]. Hence ion channels are emerging as potential genes involved in the aetiology of gliomas, and also as potential future therapeutic targets. Molecular receptor activity, metallochaperone activity and receptor regulator activity are all interrelated with these in BP percentage and CC percentage. For example, the exploratory cancer drug zinc metallochaperone-1 (ZMC1) was designed as p53 is a Zn^{2+}-dependent tumor suppressor inactivated in > 50% of human cancers [29]. Thiosemicarbazones like ZMC1 are known to interact with a number of metals involved in a variety of biologic processes. Source of Zn^{2+} is extracellular and that ZMC1 transports the metal across the plasma membrane as a transition metal-specific ionophore.

The KEGG pathways in the optimal set

Several showed certain connections with glioma among the KEGG pathway terms in the optimal set of features. Base excision repair (hsa03410) is one of the main DNA repair pathways in human that is direct reversal. DNA damage is considered to be an important mechanism in the development of glioma and it has been indicated that polymorphisms of DNA repair-related genes play important roles in the occurrence of glioma as well [30]. SNARE interactions in vesicular transport (hsa04130) are cellular mechanisms involved in glioma pathology [31]. For instance, TI-VAMP/VAMP7, a member of the vesicular SNARE proteins down-regulation, has been reported to significantly reduce secretion of cathepsin B from glioma [32]. MAPK signaling pathway (hsa04010) is known to make greate contributions to the initiation and maintenance of glioma and other brain tumors as well as normal development [33, 34]. So the up- and down-stream genes in such pathway should be very important in searching new potent antitumor target for glioma treatment.

Prediction of glioma-related genes based on optimal features

The prediction list was submitted to the functional annotation clustering tool provided by the Database for Annotation, Visualization, and Integrated Discovery (DAVID) and the result demonstrated that many genes predicted were highly associated with glioma. For example, one significant function cluster was annotated to enrich in terms like regulation of apoptosis and program cell death. In the corresponding gene list of this cluster, ENSG00000087088 (BAX) was reported to have association with glioma previously that tubeimoside-1 induced glioma cell apoptosis in a concentration-dependent manner by increasing the expression of BAX [35]. Another function cluster that intrigued us was related to DNA damage and repair. ENSG00000137337 (MDC1) the mediator of DNA damage checkpoint protein 1 in nuclear accumulation and its implication in further signal transduction, regulation of DNA damage checkpoints was reported in a research where Survivin regulated DNA-double-strand break repair machinery that led to a significant improvement of survival of glioma cells [36]. Another enriched function cluster was associated with human body response. In response to body's defense system, tumor cells often change gene expression to facilitate their survival [37]. Additionally, documented evidences showed that ENSG00000115009 (CCL20) and ENSG00000112486 (CCR6) might play an important role in the regulation of aggressiveness in human gliomas [38].

Further selection of predicted glioma-related genes using gene interaction network

An intersection analysis between the prediction genes based on optimal features found that genes had link with known genes in interaction network, a list of 34 genes received were a high possibility of being glioma related candidate genes. Function annotation clustering of these genes was performed as well. A remarkable enrichment of terms was related to DNA replication, DNA damage and repair. Another significant enrichment was concerning regulation of apoptosis and cell death. The result was consistent with the preceding prediction list enrichment and it suggested the robustness of our method. ENSG00000076242 (MLH1), as for specific gene example, was cleared of operating on temozolomide-induced autophagy via ataxia-telangiectasia mutated in glioma in a recent research [39]. ENSG00000213366 (GSTM2) Glutathione transferase mu 2 could protect glioblastoma cells against aminochrome toxicity by preventing autophagy and lysosome dysfunction [40]. A study supported a major role for ENSG00000137275 (RIPK1) in the induction of necrotic cell death based on their finding that necroptosis is associated with low procaspase-8 and active RIPK1 and -3 in human glioma cells [41]. Senescence is a state of irreversible cell growth arrest and metabolic activity maintenance that acts as an endogenous antitumor mechanism by avoiding the proliferation of transformed, pretumor cells. Senescence establishment is driven by proteins that control the cell cycle and the stress response, such as ENSG00000124762 (CDKN1A) [42]. Currently study identified ENSG00000102265 (TIMP1) as a key molecule that was acting on human neural stem cell (hNSC) adhesion and migration [43]. TIMP1, as a new chemo attractant molecule, could be utilized for the future clinical development of an hNSC-based cell-therapeutic strategy for targeting human glioma.

Conclusion

The literature review above and expression analysis by RT-PCR show that our proposed algorithm has a distinguished power to predict genes with close impact on the glioma. Other advantages of our means are short time consuming and are of little cost. We identified 3318 gene ontology terms and 127 pathway terms as the optimal features to depict the glioma genes. B860 novel related genes were predicted under those terms and 87 genes were identified that reside in the hub of known genes interaction network. There were an intersection of 34 genes between predicted genes and shortest path genes. The ultimate goal of this research is to create glioma related gene list to help achieve early detection, correct diagnosis and proper treatment strategy, finally to save the lives of patients.

Acknowledgements

The authors would like to thank the editors and reviewers for their administrative assistance in formatting the paper.

Disclosure of conflict of interest

None.

Address correspondence to: Ligang Chen, Department of Neurosurgery, The Affiliated Hospital of Southwest Medical University, No. 25, Taiping Street, Jiangyang District, Luzhou 646000, China. E-mail: ligang_chen2016@163.com

References

[1] Goodenberger ML and Jenkins RB. Genetics of adult glioma. Cancer Genet 2012; 205: 613-621.

[2] Ajaz M, Jefferies S, Brazil L, Watts C and Chalmers A. Current and investigational drug strategies for glioblastoma. Clin Oncol (R Coll Radiol) 2014; 26: 419-430.

[3] Ohgaki H and Kleihues P. Population-based studies on incidence, survival rates, and genetic alterations in astrocytic and oligodendroglial gliomas. J Neuropathol Exp Neurol 2005; 64: 479-489.

[4] Kogiku M, Ohsawa I, Matsumoto K, Sugisaki Y, Takahashi H, Teramoto A and Ohta S. Prognosis of glioma patients by combined immunostaining for survivin, Ki-67 and epidermal growth factor receptor. J Clin Neurosci 2008; 15: 1198-1203.

[5] Ostrom QT and Barnholtz-Sloan JS. Current state of our knowledge on brain tumor epidemiology. Curr Neurol Neurosci Rep 2011; 11: 329-335.

[6] Muracciole X, Romain S, Dufour H, Palmari J, Chinot O, Ouafik L, Grisoli F, Branger DF and Martin PM. PAI-1 and EGFR expression in adult glioma tumors: toward a molecular prognostic classification. Int J Radiat Oncol Biol Phys 2002; 52: 592-598.

[7] Riemenschneider MJ, Hegi ME and Reifenberger G. MGMT promoter methylation in malignant gliomas. Target Oncol 2010; 5: 161-165.

[8] Glass K and Girvan M. Finding new order in biological functions from the network structure of gene annotations. PLoS Comput Biol 2015; 11: e1004565.

[9] Kanehisa M, Goto S, Kawashima S, Okuno Y and Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res 2004; 32: D277-280.

[10] Hamosh A, Scott AF, Amberger JS, Bocchini CA and McKusick VA. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 2005; 33: D514-517.

[11] Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague JW, Campbell PJ, Stratton MR and Futreal PA. COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. Nucleic Acids Res 2011; 39: D945-950.

[12] Bauer-Mehren A, Rautschka M, Sanz F and Furlong LI. DisGeNET: a cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. Bioinformatics 2010; 26: 2924-2926.

[13] Kanehisa M, Goto S, Sato Y, Furumichi M and Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 2012; 40: D109-114.

[14] Yang J, Chen L, Kong X, Huang T and Cai YD. Analysis of tumor suppressor genes based on gene ontology and the KEGG pathway. PLoS One 2014; 9: e107202.

[15] Peng H, Long F and Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 2005; 27: 1226-1238.

[16] Ge H and Zhang G. Retracted: identifying halophilic proteins based on random forests with preprocessing of the pseudo-amino acid composition. J Theor Biol 2014; 361: 175-181.

[17] Chu A, Cui J and Dinov ID. SOCR analyses-an instructional Java web-based statistical analysis toolkit. J Online Learn Teach 2009; 5: 1-18.

[18] Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ and von Mering C. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 2011; 39: D561-568.

[19] Sharma VP, Anderson NT and Geusz ME. Circadian properties of cancer stem cells in glioma cell cultures and tumorspheres. Cancer Lett 2014; 345: 65-74.

[20] Li A, Lin X, Tan X, Yin B, Han W, Zhao J, Yuan J, Qiang B and Peng X. Circadian gene clock contributes to cell proliferation and migration of glioma and is directly regulated by tumor-suppressive miR-124. FEBS Lett 2013; 587: 2455-2460.

[21] Zhanfeng N, Yanhui L, Zhou F, Shaocai H, Guangxing L and Hechun X. Circadian genes Per1 and Per2 increase radiosensitivity of glioma in vivo. Oncotarget 2015; 6: 9951-9958.

[22] Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008; 455: 1061-1068.

[23] Teodorczyk M and Schmidt MH. Notching on cancer's door: notch signaling in brain tumors. Front Oncol 2014; 4: 341.

[24] Rozario T and DeSimone DW. The extracellular matrix in development and morphogenesis: a dynamic view. Dev Biol 2010; 341: 126-140.

[25] Buckingham SC and Robel S. Glutamate and tumor-associated epilepsy: glial cell dysfunction in the peritumoral environment. Neurochem Int 2013; 63: 696-701.

[26] Bhadriraju K, Chung KH, Spurlin TA, Haynes RJ, Elliott JT and Plant AL. The relative roles of collagen adhesive receptor DDR2 activation and matrix stiffness on the downregulation of focal adhesion kinase in vascular smooth muscle cells. Biomaterials 2009; 30: 6687-6694.

[27] Brosicke N, Sallouh M, Prior LM, Job A, Weberskirch R and Faissner A. Extracellular matrix glycoprotein-derived synthetic peptides differentially modulate glioma and sarcoma

cell migration. Cell Mol Neurobiol 2015; 35: 741-753.

[28] Sforna L, Cenciarini M, Belia S, D'Adamo MC, Pessia M, Franciolini F and Catacuzzeno L. The role of ion channels in the hypoxia-induced aggressiveness of glioblastoma. Front Cell Neurosci 2014; 8: 467.

[29] Blanden AR, Yu X, Wolfe AJ, Gilleran JA, Augeri DJ, O'Dell RS, Olson EC, Kimball SD, Emge TJ, Movileanu L, Carpizo DR and Loh SN. Synthetic metallochaperone ZMC1 rescues mutant p53 conformation by transporting zinc into cells as an ionophore. Mol Pharmacol 2015; 87: 825-831.

[30] Rajaraman P, Hutchinson A, Wichner S, Black PM, Fine HA, Loeffler JS, Selker RG, Shapiro WR, Rothman N, Linet MS and Inskip PD. DNA repair gene polymorphisms and risk of adult meningioma, glioma, and acoustic neuroma. Neuro Oncol 2010; 12: 37-48.

[31] Christmann M, Tomicic MT, Roos WP and Kaina B. Mechanisms of human DNA repair: an update. Toxicology 2003; 193: 3-34.

[32] Verderio C, Cagnoli C, Bergami M, Francolini M, Schenk U, Colombo A, Riganti L, Frassoni C, Zuccaro E, Danglot L, Wilhelm C, Galli T, Canossa M and Matteoli M. TI-VAMP/VAMP7 is the SNARE of secretory lysosomes contributing to ATP secretion from astrocytes. Biol Cell 2012; 104: 213-228.

[33] Penman CL, Faulkner C, Lowis SP and Kurian KM. Current understanding of BRAF alterations in diagnosis, prognosis, and therapeutic targeting in pediatric low-grade gliomas. Front Oncol 2015; 5: 54.

[34] Jones DT, Gronych J, Lichter P, Witt O and Pfister SM. MAPK pathway activation in pilocytic astrocytoma. Cell Mol Life Sci 2012; 69: 1799-1811.

[35] Jia G, Wang Q, Wang R, Deng D, Xue L, Shao N, Zhang Y, Xia X, Zhi F and Yang Y. Tubeimoside-1 induces glioma apoptosis through regulation of Bax/Bcl-2 and the ROS/Cytochrome C/Caspase-3 pathway. Onco Targets Ther 2015; 8: 303-311.

[36] Reichert S, Rodel C, Mirsch J, Harter PN, Tomicic MT, Mittelbronn M, Kaina B and Rodel F. Survivin inhibition and DNA double-strand break repair: a molecular mechanism to overcome radioresistance in glioblastoma. Radiother Oncol 2011; 101: 51-58.

[37] Musumeci M, Coppola V, Addario A, Patrizii M, Maugeri-Sacca M, Memeo L, Colarossi C, Francescangeli F, Biffoni M, Collura D, Giacobbe A, D'Urso L, Falchi M, Venneri MA, Muto G, De Maria R and Bonci D. Control of tumor and microenvironment cross-talk by miR-15a and miR-16 in prostate cancer. Oncogene 2011; 30: 4231-4242.

[38] Wang L, Qin H, Li L, Zhang Y, Tu Y, Feng F, Ji P, Zhang J, Li G, Zhao Z and Gao G. Overexpression of CCL20 and its receptor CCR6 predicts poor clinical prognosis in human gliomas. Med Oncol 2012; 29: 3491-3497.

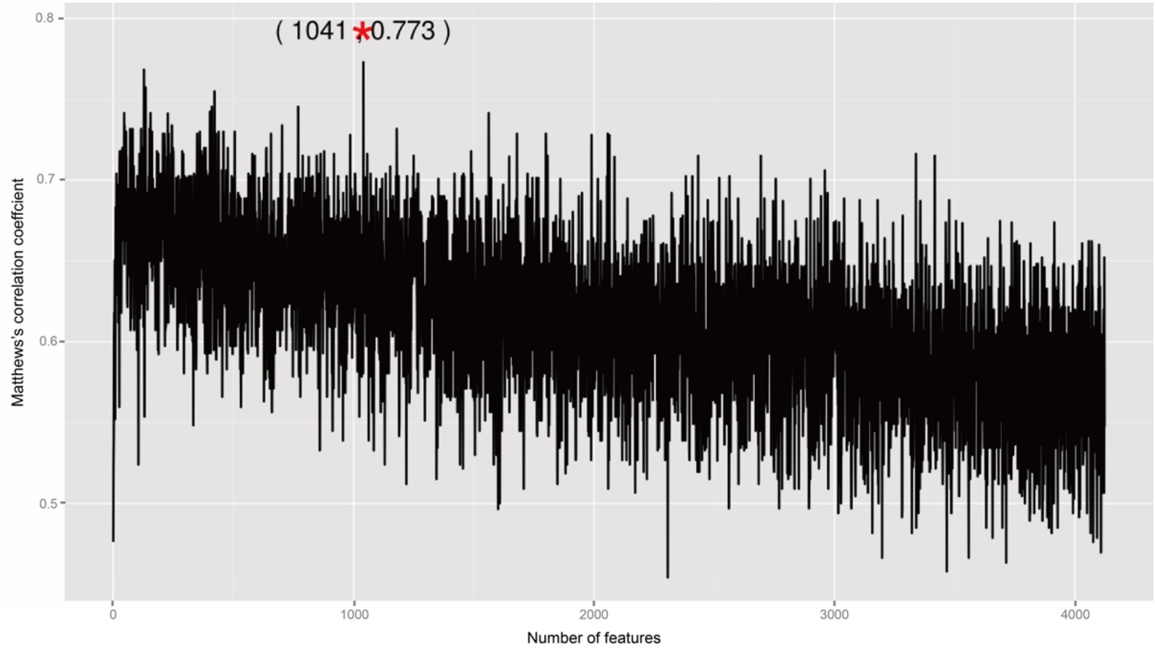
[39] Zou Y, Wang Q and Wang W. MutL homolog 1 contributes to temozolomide-induced autophagy via ataxia-telangiectasia mutated in glioma. Mol Med Rep 2015; 11: 4591-4596.

[40] Huenchuguala S, Munoz P, Zavala P, Villa M, Cuevas C, Ahumada U, Graumann R, Nore BF, Couve E, Mannervik B, Paris I and Segura-Aguilar J. Glutathione transferase mu 2 protects glioblastoma cells against aminochrome toxicity by preventing autophagy and lysosome dysfunction. Autophagy 2014; 10: 618-630.

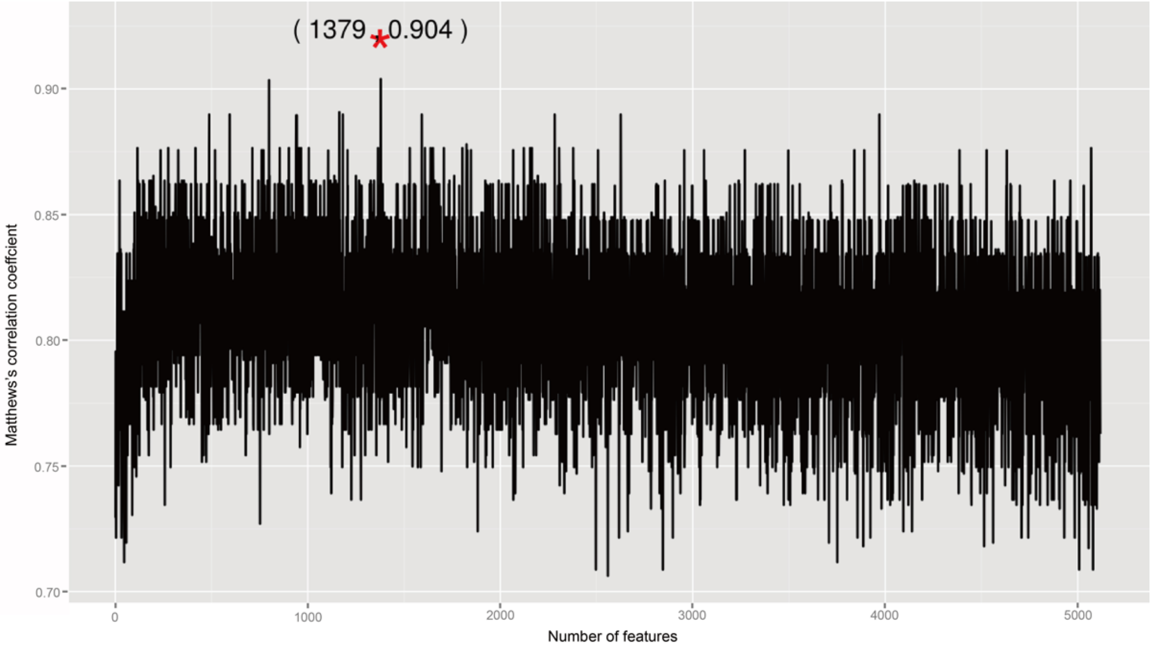
[41] Melo-Lima S, Celeste Lopes M and Mollinedo F. Necroptosis is associated with low procaspase-8 and active RIPK1 and -3 in human glioma cells. Oncoscience 2014; 1: 649-664.

[42] Filippi-Chiela EC, Bueno e Silva MM, Thome MP and Lenz G. Single-cell analysis challenges the connection between autophagy and senescence induced by DNA damage. Autophagy 2015; 11: 1099-1113.

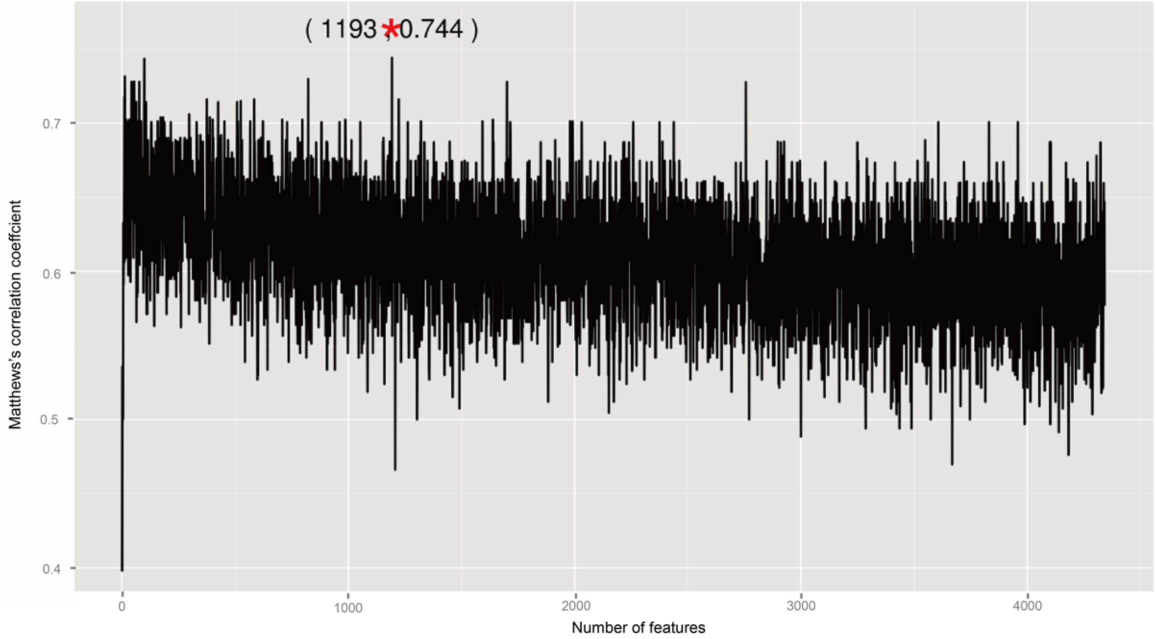
[43] Lee SY, Kim JM, Cho SY, Kim HS, Shin HS, Jeon JY, Kausar R, Jeong SY, Lee YS and Lee MA. TIMP-1 modulates chemotaxis of human neural stem cells through CD63 and integrin signalling. Biochem J 2014; 459: 565-576.



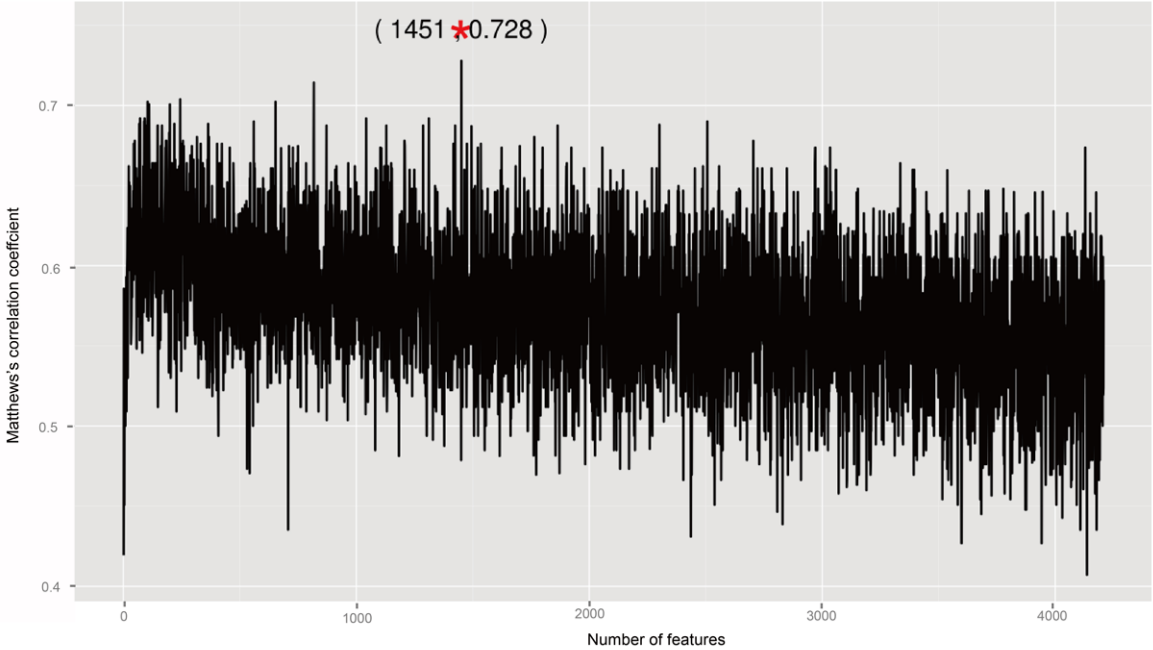
Supplementary Figure 1. The IFS curve for the dataset. The red asterisk indicates where the maximal MCC value appears.



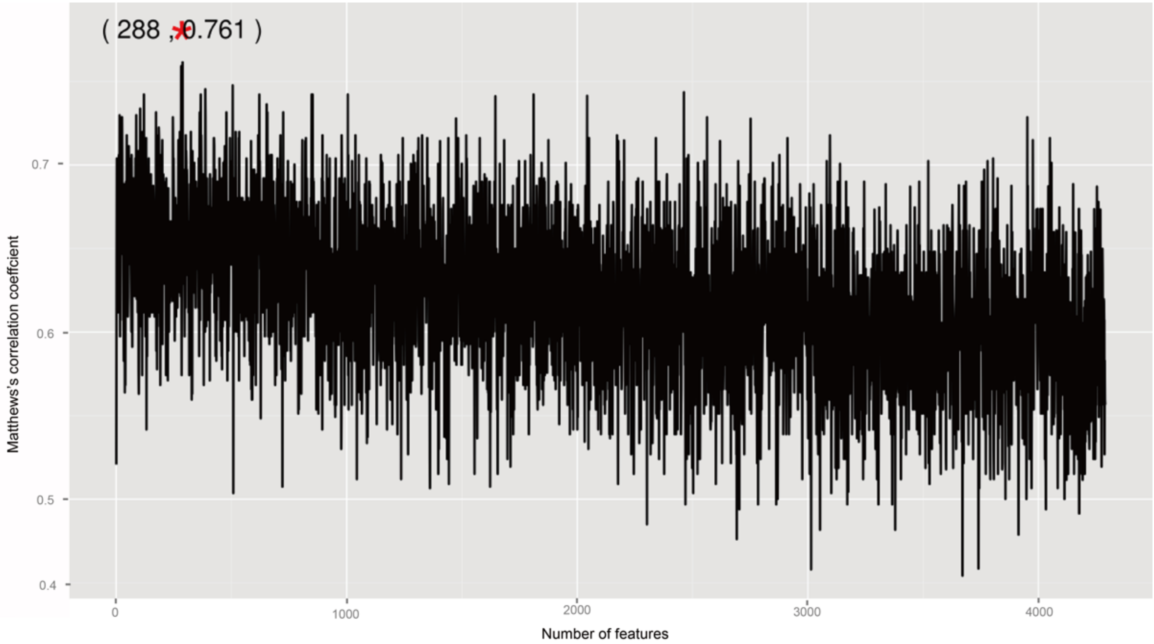
Supplementary Figure 2. The IFS curve for the dataset. The red asterisk indicates where the maximal MCC value appears.



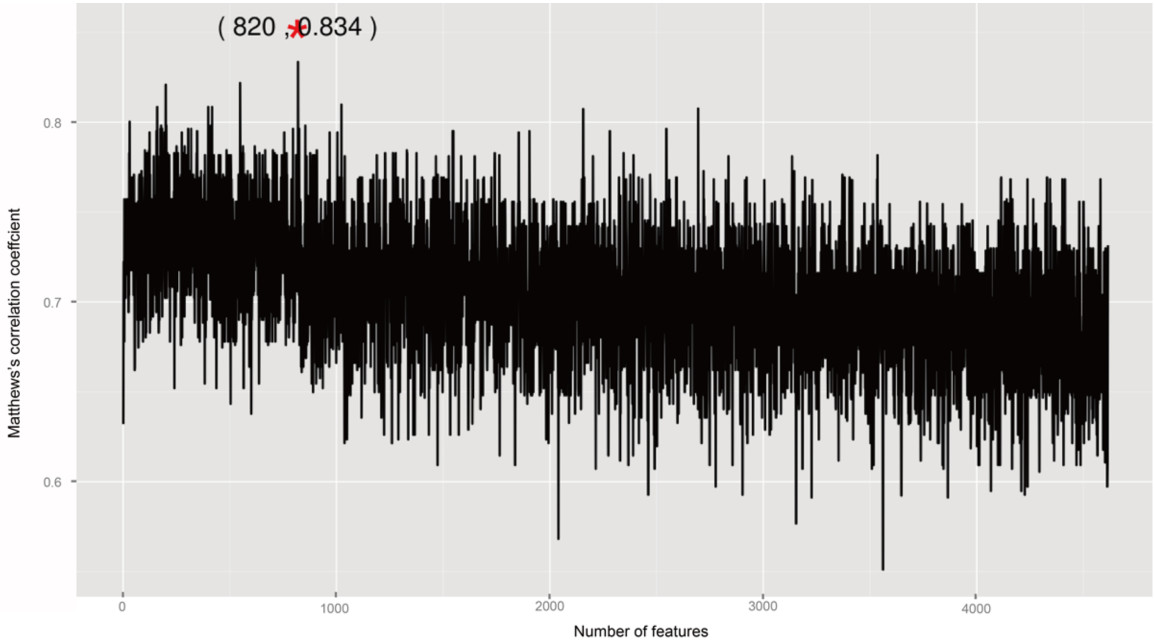
Supplementary Figure 3. The IFS curve for the dataset. The red asterisk indicates where the maximal MCC value appears.



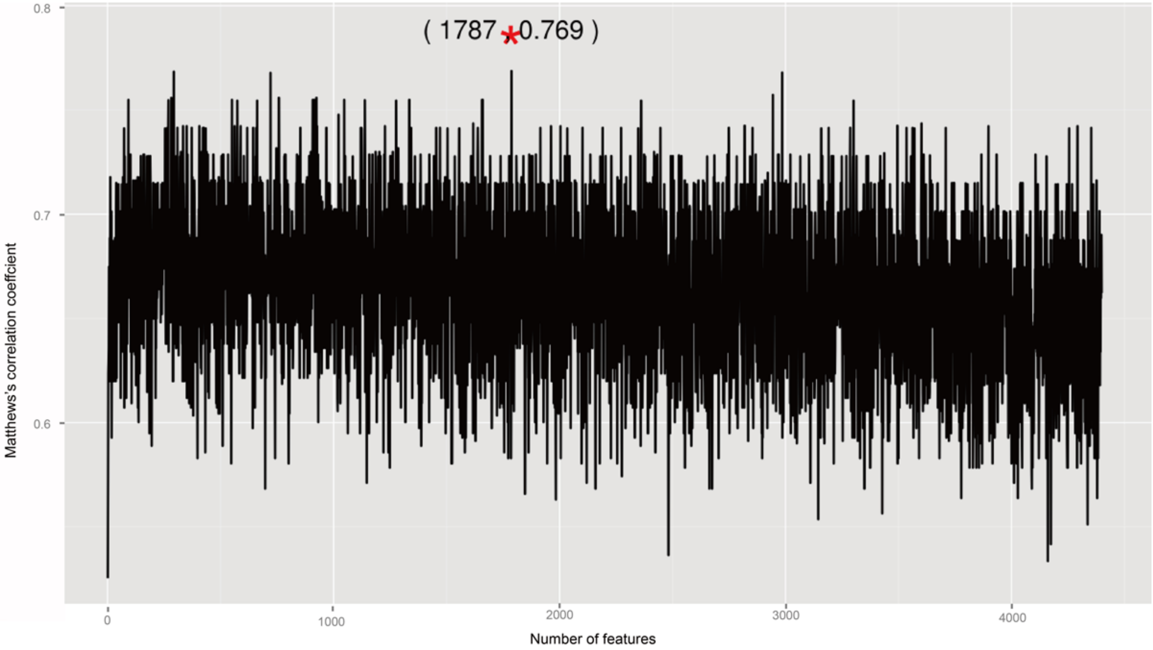
Supplementary Figure 4. The IFS curve for the dataset. The red asterisk indicates where the maximal MCC value appears.



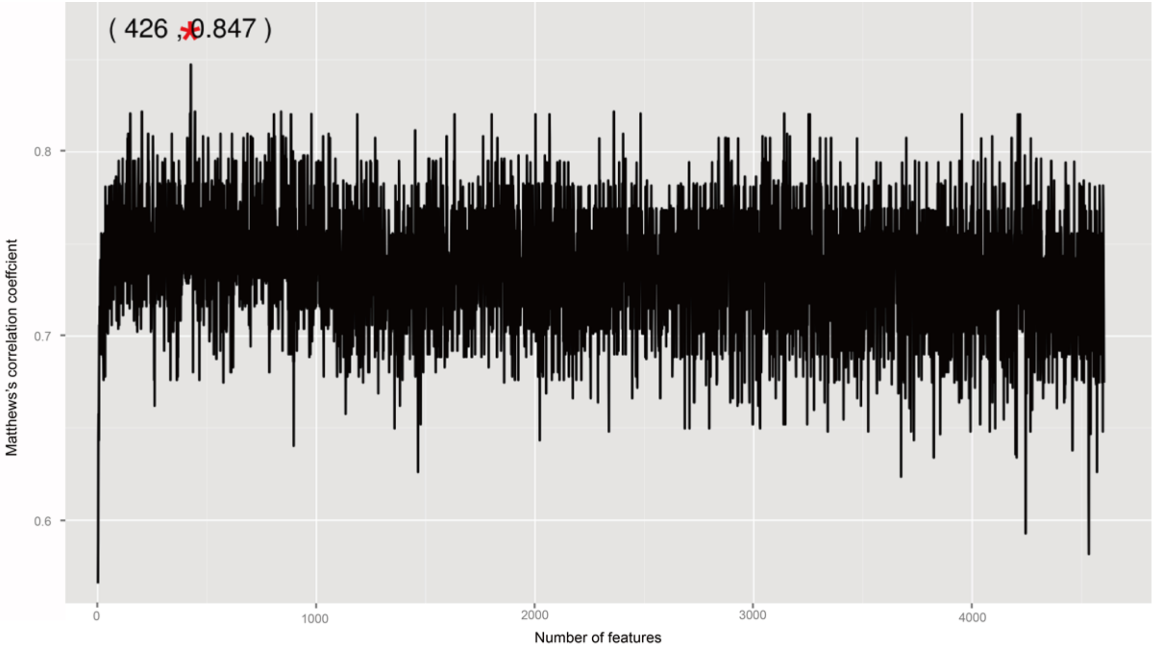
Supplementary Figure 5. The IFS curve for the dataset. The red asterisk indicates where the maximal MCC value appears.



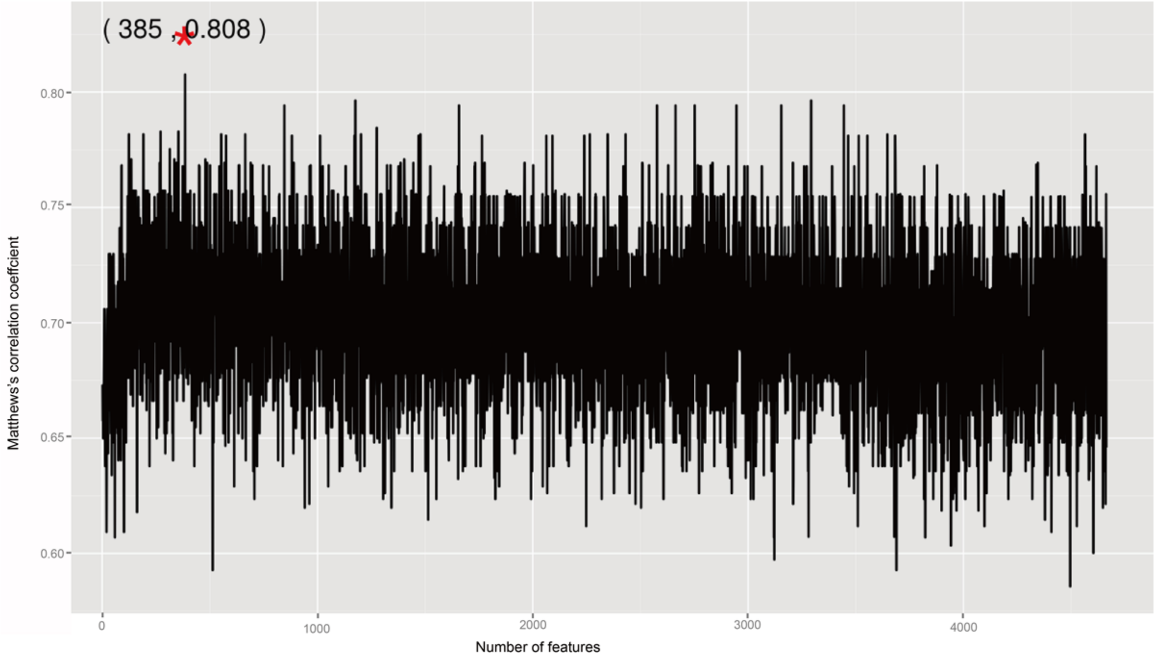
Supplementary Figure 6. The IFS curve for the dataset. The red asterisk indicates where the maximal MCC value appears.



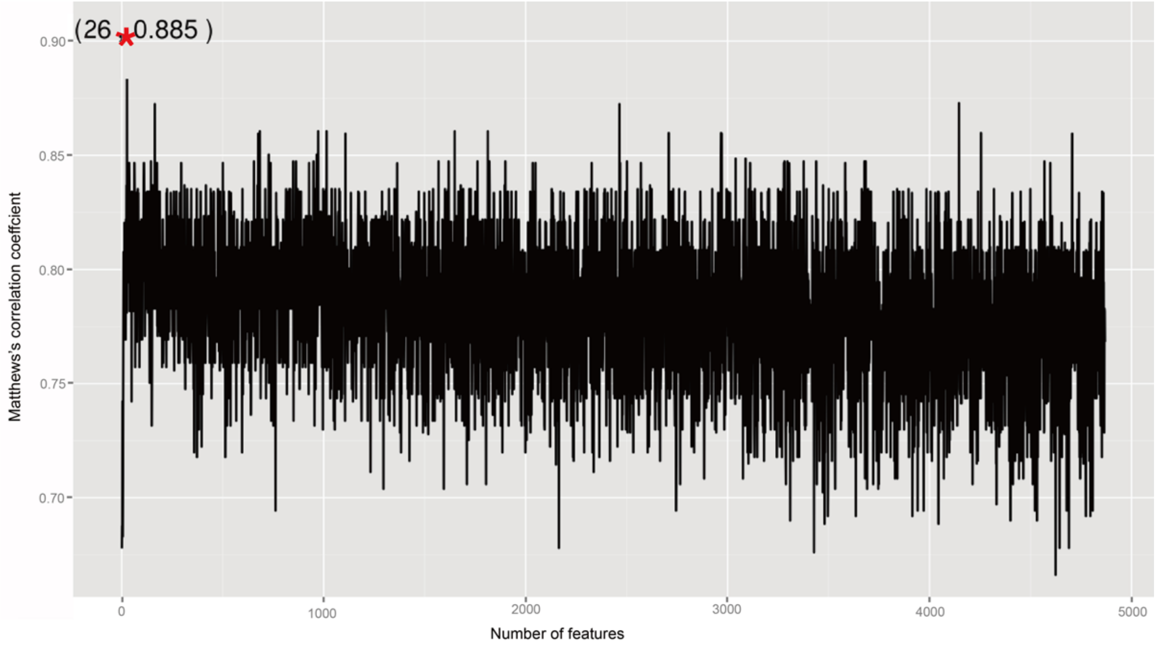
Supplementary Figure 7. The IFS curve for the dataset. The red asterisk indicates where the maximal MCC value appears.



Supplementary Figure 8. The IFS curve for the dataset. The red asterisk indicates where the maximal MCC value appears.



Supplementary Figure 9. The IFS curve for the dataset. The red asterisk indicates where the maximal MCC value appears.



Supplementary Figure 10. The IFS curve for the dataset. The red asterisk indicates where the maximal MCC value appears.