frontiers in Genetics

ORIGINAL RESEARCH
published: 24 November 2020
doi: 10.3389/fgene.2020.604336

Distinguishing Glioblastoma Subtypes by Methylation Signatures

Yu-Hang Zhang[1,2†], Zhandong Li[3†], Tao Zeng[4], Xiaoyong Pan[5], Lei Chen[6], Dejing Liu[7], Hao Li[3], Tao Huang[7*] and Yu-Dong Cai[1*]

[1] School of Life Sciences, Shanghai University, Shanghai, China, [2] Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States, [3] College of Food Engineering, Jilin Engineering Normal University, Changchun, China, [4] Shanghai Research Center for Brain Science and Brain-Inspired Intelligence, Shanghai, China, [5] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, China, [6] College of Information Engineering, Shanghai Maritime University, Shanghai, China, [7] Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China

OPEN ACCESS

Edited by:
Min Tang,
Jiangsu University, China

Reviewed by:
Yanding Zhao,
Baylor College of Medicine,
United States
Xuefeng Gu,
Shanghai University of Medicine
and Health Sciences, China

****Correspondence:***
Tao Huang
tohuangtao@126.com
Yu-Dong Cai
cai_yud@126.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: *09 September 2020*
Accepted: *02 November 2020*
Published: *24 November 2020*

Citation:
Zhang Y-H, Li Z, Zeng T, Pan X, Chen L, Liu D, Li H, Huang T and Cai Y-D (2020) Distinguishing Glioblastoma Subtypes by Methylation Signatures. Front. Genet. 11:604336. doi: 10.3389/fgene.2020.604336

Glioblastoma, also called glioblastoma multiform (GBM), is the most aggressive cancer that initiates within the brain. GBM is produced in the central nervous system. Cancer cells in GBM are similar to stem cells. Several different schemes for GBM stratification exist. These schemes are based on intertumoral molecular heterogeneity, preoperative images, and integrated tumor characteristics. Although the formation of glioblastoma is remarkably related to gene methylation, GBM has been poorly classified by epigenetics. To classify glioblastoma subtypes on the basis of different degrees of genes' methylation, we adopted several powerful machine learning algorithms to identify numerous methylation features (sites) associated with the classification of GBM. The features were first analyzed by an excellent feature selection method, Monte Carlo feature selection (MCFS), resulting in a feature list. Then, such list was fed into the incremental feature selection (IFS), incorporating one classification algorithm, to extract essential sites. These sites can be annotated onto coding genes, such as *CXCR4, TBX18, SP5*, and *TMEM22*, and enriched in relevant biological functions related to GBM classification (e.g., subtype-specific functions). Representative functions, such as nervous system development, intrinsic plasma membrane component, calcium ion binding, systemic lupus erythematosus, and alcoholism, are potential pathogenic functions that participate in the initiation and progression of glioblastoma and its subtypes. With these sites, an efficient model can be built to classify the subtypes of glioblastoma.

Keywords: glioblastoma, methylation, signature, subtype, classification

INTRODUCTION

Glioblastoma, also called as glioblastoma multiform (GBM), is the most aggressive cancer that initiates within the brain. The cause of this disease is unclear. The risk factors of GBM include genetic factors and environmental factors, such as smoking and exposure to pesticides. Similar to other brain cancers, GBM can cause epilepsy, nausea, vomiting, headaches, and mild hemiplegia. The typical symptoms of glioblastoma are deteriorating memory and personality or decline in neurological function. Most symptoms are caused by the destruction of the temporal lobes and the frontal lobes. Different subspecies of glioblastomas are produced in the central nervous system, and cancer cells in GBM are similar to stem cells.

Several different schemes for glioblastoma stratification exist. One is based on intertumoral molecular heterogeneity in GBM. This scheme identities the subtypes of procedural and mesenchymal glioblastoma on the basis of the biomarker genes *VEGF-A*, *VEGF-B*, *ANG1*, and *ANG2* (Sharma et al., 2017). The second technique involves the use of preoperative images as predictive markers of GBM subtypes; in this approach, the distinctive imaging phenotypes and imaging patterns of glioblastoma subtypes are detected by employing machine-learning techniques (Macyszyn et al., 2016). The third technique is based on integrated tumor subtypes, which have been discovered through an integrative subtype analysis of the GBM dataset from the cancer genome atlas (TCGA) (Shen et al., 2012).

The promoter region is a functional part of the genome that is regulated by methylation and contributes to the regulation of gene expression during the pathogenesis of glioblastoma. Such genomic modification affects the expression of a group of important proteins, including *MGMT*, *GATA6*, and *CASP8*; the dysmethylation of these genes is remarkable in glioblastoma (Skiriute et al., 2012). For example, through whole-genome wide methylation screening, a study found that 5 m-dC level is the best discriminant among methylation classes, and the upregulation of *LINE1* methylation is an independent prognostic factor in GBM diseases (Lai et al., 2014). Although the formation of glioblastoma is related to gene methylation, glioblastoma has been poorly classified on the basis of epigenetics.

Preliminary attempts on clustering GBMs using epigenetic biomarkers have already started. According to a systematic analysis on the DNA methylation-based classification of central nervous system tumors (Guardiola Bagán et al., 2017; Capper et al., 2018), central nerve system (CNS) tumors can be further classified into multiple subgroups based on the whole-genome wide methylation status. As one important part of the CNS tumors, GBM can be further classified into eight classes, which is DMG K27, GBM G34, GBM MES, GBM RTK I, GBM RTK II, GBM RTK III, GBM MID, and GBM MYCN. Researchers tried to use unsupervised clustering of reference samples using t-SNE dimensionality reduction. According to the original publications, group DMG K27 can be easily distinguished from other seven groups based on the results of t-SNE based separation. However, the differences between the other seven subgroups cannot be clarified clearly and the specific methylation locus that contribute to the separation have not been identified. Therefore, in this study, we used methylation datasets downloaded from Gene Expression Omnibus (GEO) database to identify specific methylation locus/biomarkers that contribute to the classification and annotation of different GBM subgroups (Capper et al., 2018).

We aimed to identify essential methylation sites (features) in this study, on which the subtypes of glioblastoma can be efficiently classified. To this end, we employed two datasets collected in GEO. One dataset was termed as the training dataset, whereas the other was treated as the independent test dataset. A powerful feature selection method, Monte Carlo feature selection (MCFS) (Dramiński et al., 2007), was applied on the training dataset. A feature list, indicating the importance of features, was produced. After that, incremental feature selection (IFS) (Liu and Setiono, 1998) was executed on this list, which incorporated one classification algorithm, to extract essential methylation sites. As a result, we found 4100 methylation sites (features) associated with the classification of GBM. These sites can be annotated onto coding genes, such as *CXCR4, TBX18, SP5*, and *TMEM22*. Through the further functional enrichment analysis of these dysmethylated genes using GO and KEGG databases, we identified several biological functions related to GBM classification (e.g., subtype-specific functions). Also, with these methylation sites, an efficient model with support vector machine (SVM) (Cortes and Vapnik, 1995) as the prediction engine can be built to classify subtypes of glioblastoma. In summary, on the basis of the powerful computational approaches, we identified various novel potential pathogenic genes at the epigenetics level and revealed several potential pathogenic functions that participate in the initiation and progression of glioblastoma and its subtypes with wide support from recent reports.

MATERIALS AND METHODS

Dataset

Two sets of methylation profiles of patients with GBM were downloaded from GEO with the accession numbers GSE90496 and GSE109379 (Capper et al., 2018). The first dataset included 347 GBM cases and the second dataset contained 324 GBM cases. These two datasets were used as the training dataset and independent test dataset, respectively. All GBM cases are classified into seven categories. The distribution of GBM cases on seven categories is listed in **Table 1**. The methylation levels of 42,383 probes were used to represent each patient. The goal was to identify discriminative methylation features (e.g., dysmethylated sites or genes) corresponding to different GBM subtypes.

Feature Selection

In this study, we first used MCFS (Dramiński et al., 2007) to identify the general interpretable information of features (methylation sites) in tumor samples from the central nervous system. Then, we applied IFS (Liu and Setiono, 1998) to

TABLE 1 | Breakdown of the GBM samples in the training and independent datasets.

Category	Training dataset	Independent dataset
G34	41	13
MES	56	104
MID	14	19
MYCN	16	17
RTK	64	44
RTK II	143	118
RTK III	13	9

improve classification performance by obtaining a group of optimal features with the strong recognition ability of central nervous system tumors.

MCFS

Monte Carlo feature selection is a classical and powerful feature selection method wherein decision trees are used to find distinguishable features for classification (Dramiński et al., 2007). It is quite suitable to analyze datasets with features much more than samples. The datasets described in section "Dataset" are in such type. Thus, we adopted MCFS to analyze the training dataset, aiming to extract essential features. Furthermore, such feature selection method can deeply investigate complicated relationship between features or class labels, extracting essential features in deep levels.

The MCFS method evaluates the importance of features by constructing lots of decision trees. Given a dataset with M features, randomly construct s feature subsets consisting of m features, where m is much smaller than M. For each feature dataset, t bootstrap sample sets are constructed from the original dataset, in which samples are represented by features in such feature subset. Accordingly, t decision trees are built. After all feature subsets are processed by the above procedures, $s{\cdot}t$ decision trees are constructed. Based on these trees, a feature g is assigned a relative importance (RI) value, which can be calculated by

$$RI_g = \sum_{\tau=1}^{st} (wAcc)^u \sum_{n_g(\tau)} IG(n_g(\tau)) \left(\frac{\text{no. in } n_g(\tau)}{\text{no. in } \tau} \right)^v, \qquad (1)$$

where $IG(n_g(\tau))$ stands for the gain information of node $n_g(\tau)$, (no. in $n_g(\tau)$) represents the number of samples in node $n_g(\tau)$, (no. in τ) denotes the number of samples in tree τ, $wAcc$ indicates the weighted accuracy of the tree. u and v are the regular factors, which were suggested to set to one (Dramiński et al., 2007). All investigated features are ranked in a list with the decreasing order of their RI values. Clearly, features with high ranks are more important than those with low ranks.

In present study, we used the MCFS program retrieved from http://www.ipipan.eu/staff/m.draminski/mcfs.html. Default parameters were adopted.

IFS

Incremental feature selection is a feature selection method used to distinguish between samples from different classes (e.g., normal and diseased) (Liu and Setiono, 1998). In this study, different classes of samples were discerned by a set of optimal features screened by IFS performed in a rank-descending feature list. We set candidate high-performance feature subsets as feature subsets with large interval sizes (e.g., 10 features) from the ranked feature list. Suppose N candidate feature subsets $\mathrm{F} = [F^1, F^2, \ldots, F^N]$ exist. The i-th feature subset includes $10 * i$ features yielding $F^i = [f_1, f_2, \ldots, f_{i*N}]$. We construct and evaluate the classifier on each candidate feature subset. The candidate feature subset with the maximal prediction performance is the optimal feature subset, and the classifier constructed from these optimal features is the optimal classifier.

Classification Algorithm

Support Vector Machine

The classifier acts as a classification model that maps data samples to a given category for data class prediction. We use support vector machine (SVM) (Cortes and Vapnik, 1995) based on statistical learning theory for supervised data classification. It has wide application for tackling different biological problems (Muthukrishnan et al., 2014; Chen et al., 2017, 2020; Liu et al., 2020; Sang et al., 2020; Zhou et al., 2020a,b). The basic principle is to use a given kernel function (e.g., Gaussian kernel) to transform data from a low-dimensional space to a high-dimensional space. The SVM model can separate the samples of each class/category by maximizing the data interval and also predicts (new) sample categories on the basis of the interval where this sample falls in. For two-class classification, the largest margin between the two categories of samples can be inferred by SVM, where large margins are associated with small generalization error. For multiclass classification, SVM uses the "One Versus the Rest" strategy. In this study, we solved the optimization problem of SVM by using the sequence minimization optimization (SMO) algorithm (Platt, 1998; Keerthi et al., 2001) implemented by the tool "SMO" in Weka software (Frank et al., 2004; Witten and Frank, 2005), which can be downloaded at https://www.cs.waikato.ac.nz/ml/weka/. For convenience, the default parameters were adopted, where the kernel was a polynomial function and the regularization parameter C was set to one.

Random Forest

A random forest (RF) (Breiman, 2001) is a metaclassifier that contains a large number of tree classifiers for establishing final joint classification, which determines the output categories/classes by summarizing votes from different decision trees (Breiman, 2001). The RF is a commonly used method in machine learning and is widely applied in computational biology (Pan et al., 2010; Zhao et al., 2018; Jia et al., 2020; Liang et al., 2020; Yuan et al., 2020). Notably, a slight difference exists between each decision tree and other decision trees in a RF. Thus, the predictions of all decision trees are averaged to obtain the final decision of RF. This approach can avoid over-fitting and improve the performance of the integrated model. However, it slightly increases the bias of the overall model and causes the loss of some interpretability. In this study, we used the tool "RandomForest" in Weka (Frank et al., 2004; Witten and Frank, 2005), which implemented the above RF. The number of decision trees was set to ten.

Rule Learning

In this study, we used the rule learner known as repeated incremental pruning to produce error reduction (RIPPER) to generate classification rules for classifying samples from different GBM subtypes (Cohen, 1995). RIPPER learns interpretable classification rules consisting of IF–ELSE rules. Briefly, RIPPER learns the rules of one class and then moves to learn the next class in a given order, e.g., it learns from the first minority class to the next until the dominant class. To quickly implement the RIPPER algorithm, we directly employed the tool "JRip" in Weka (Frank et al., 2004; Witten and Frank, 2005). Default parameters were used.

Functional Enrichment Analysis

The selected optimal methylation probes (features) were mapped onto genes on the basis of the annotation files of GPL13534 downloaded from GEO. The enrichments of these genes on GO terms and KEGG pathways were evaluated with hypergeometric tests measured by phyper function in R[1]. The cutoff of the adjusted hypergeometric test *p*-values, i.e., FDR (false discovery rate), was set to 0.05. In other words, only the GO terms and KEGG pathways with FDR $<$ 0.05 were considered to be statistically significant.

Performance Measurement

We employed Matthew Correlation Coefficients (MCC) (Matthews, 1975; Gorodkin, 2004) to evaluate the performance metrics of different kinds of classifiers. The MCC accounts for true and false positives and true and false negatives, and this measurement has values ranging from -1 and $+1$. It is a common method for calculating the correlation between target and prediction classes. Applying 10-fold cross-validation (Kohavi, 1995), we used MCC to evaluate the performance of different training models for glioblastoma classification.

RESULTS

In this study, we investigated the methylation profiles of GBM patients. The entire procedures are illustrated in **Figure 1**.

Results of MCFS Method on the Training Dataset

We first used MCFS to analyze the training dataset. Each feature was evaluated by a RI value. Accordingly, all features were ranked in the decreasing order of their RI values. Obtained feature list is provided in **Supplementary Table 1**.

IFS Results

Next, we generated a series of feature subsets from the MCFS feature list and then subjected them to IFS with SVM, RF, and RIPPER to obtain the best features for classifying different categories of GBM samples. The complete results of the three classifiers using different number of features are given in **Supplementary Table 2**. For an easy observation, an IFS curve was plotted with number of used features as X-axis and MCC as the Y-axis for each classification algorithm, as shown in **Figure 2**, in which the highest MCC of each classification is marked. It can be observed that the highest value of MCC generated by SVM was 0.939 when using the top-ranked 4100 features. Accordingly, we constructed the optimal SVM classifier with these 4100 features. For RF, when using the top-ranked 1690 features, the largest MCC value of 0.882 was achieved. These 1690 features were used to build the optimal RF classifier. When using the top-ranked 1180 features, the highest MCC value of 0.737 was obtained by RIPPER. The optimal RIPPER classifier was built based on these 1180 features. The overall accuracies of above-mentioned classifiers are listed in **Table 2** and the accuracies on seven categories are shown in **Figure 3**. As shown by these results, the optimal classifier was SVM, which was superior to RIPPER and RF although it used additional features.

Performance of Optimal Classifiers on the Test Dataset

To show the generalizability of our pipeline, we also evaluated above-constructed classifiers on a completely independent test dataset. The MCCs generated by the optimal SVM, RF, and RIPPER classifiers were 0.798, 0.832, and 0.937. These results are summarized in **Table 3**, in which the corresponding overall accuracies are also listed. The detailed performance on each category is shown in **Figure 4**. The results indicated that the RIPPER classifier had better generalizability than other two algorithms, and SVM shown the worst generalizability performance in this study.

Results of Enrichment Analysis

On the training dataset, the optimal SVM classifier gave the best performance, which adopted 4100 top-ranked features (methylation sites). These sites were mapped onto genes based on the annotation file of Illumina HumanMethylation450 BeadChip from GEO with platform number of GPL13534[2], resulting

[1] http://finzi.psych.upenn.edu/R/library/stats/html/Hypergeometric.html

[2] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL13534

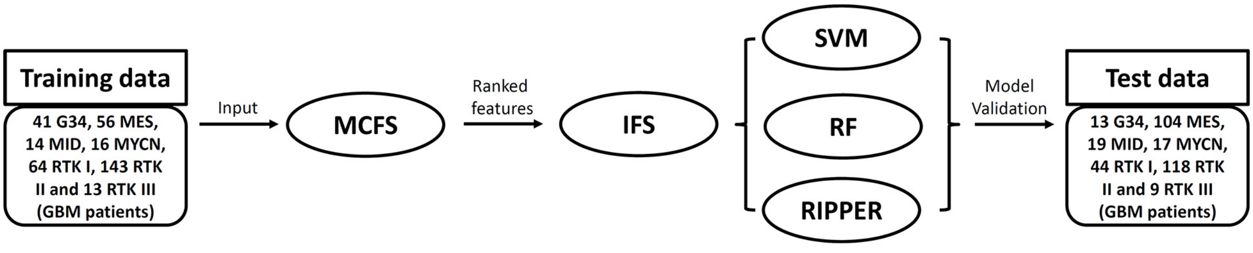
Training data
41 G34, 56 MES, 14 MID, 16 MYCN, 64 RTK I, 143 RTK II and 13 RTK III (GBM patients)
Input
MCFS
Ranked features
IFS
SVM
RF
RIPPER
Model Validation
Test data
13 G34, 104 MES, 19 MID, 17 MYCN, 44 RTK I, 118 RTK II and 9 RTK III (GBM patients)

FIGURE 1 | Flowchart of the analysis performed in this study. The training dataset is first analyzed by the Monte Carlo feature selection (MCFS) method. Features are ranked in a list, which is fed into the incremental feature selection (IFS) with one of three classification algorithms. The optimal classifiers based on different classification algorithms are built and further evaluated their performance on a test dataset.

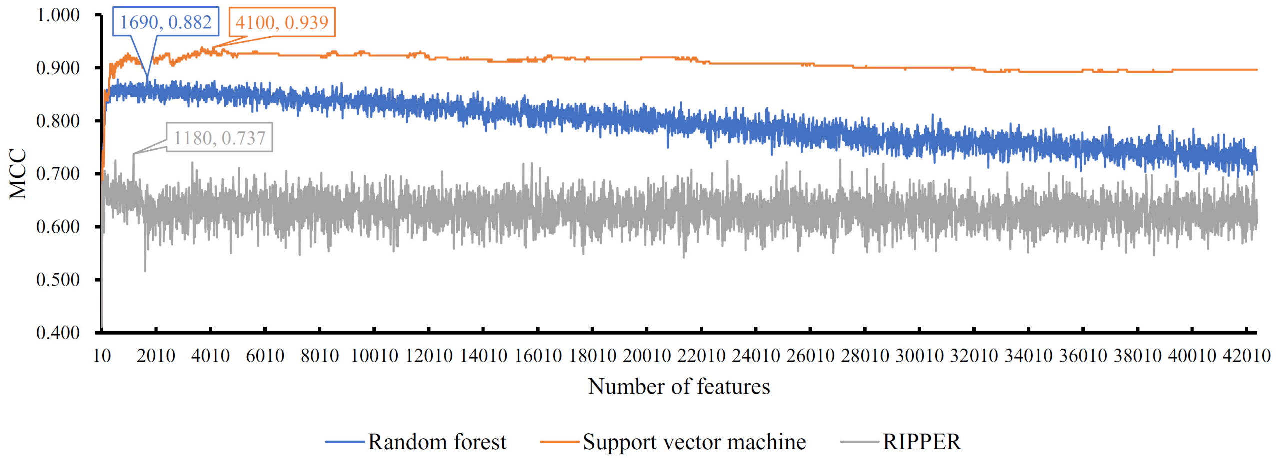

FIGURE 2 | IFS curves with support vector machine, random forest, and RIPPER on the training set. The support vector machine can yield the highest MCC (0.939) when top 4100 features are used, while the highest MCCs of random forest and RIPPER are 0.882 and 0.737, respectively, when top 1690 and 1180, features respectively, are adopted.

in 1813 coding genes, which are provided in **Supplementary Table 3**. For consistency, these genes were called optimal genes in the following text.

The enrichment analysis was done on the above 1813 genes. The results are listed in **Supplementary Table 4**. Several GO terms and KEGG pathways with FDR < 0.05 were obtained. In detail, we obtained 167 biological process (BP) GO terms, 28 cellular component (CC) GO terms, 26 molecular function (MF) GO terms and four KEGG pathways. Some of them would be analyzed in section "Biological Functions Relevant to GBM Based on Optimal Genes" and "Biological Pathways Relevant to GBM Based on Optimal Genes."

DISCUSSION

Optimal Genes Relevant to GBM

As mentioned in section "Results of Enrichment Analysis," 1813 optimal genes were obtained. We selected some of them for analysis in this section. These genes are targeted by probes with high RI values.

The first gene is ***CXCR4*** (targeted by probe **cg02902079** and **cg10824187**), which is a lymphocyte activity regulation molecule and acts as an alpha-chemokine receptor specific for stromal-derived-factor-1. Chemokines play important autocrine and paracrine roles during tumor initiation and progression. Generally, the *in vivo* secretion of chemokines regulates the biological effects of various components in the microenvironment of *CXCR4* (Würth et al., 2014). In cancer stem cells, *CXCR4* is upregulated and plays an irreplaceable role in perivascular invasion, a specific tumor behavior in GBM (Yadav et al., 2016). In addition, *CXCR4* is an effective target for improving tumor sensitivity in GBM in conjunction with radiation therapy (Yadav et al., 2016). Moreover, *CXCR4* is suppressed by *PATZ1*, which is enriched in the proneural subtype and colocalizes with stemness markers of GBMs (Guadagno et al., 2017).

TABLE 2 | 10-fold cross-validation performance of the optimal SVM, RF, and RIPPER classifiers on the training set.

Classification algorithm	Number of features	Overall accuracy	MCC
SVM	4100	0.954	0.939
RF	1690	0.911	0.882
RIPPER	1180	0.804	0.737

The next identified probe turns out to be **cg26558485**, targeting the 5′UTR of CYP4X1. As a member of the cytochrome P450 superfamily of enzyme, such gene has been generally reported to participate in neurovascular function in the brain (Bylund et al., 2002). As for its correlations with GBM, recently, two successive related publications (Wang et al., 2018, 2019) confirmed that CYP4X1 contributes to the inhibition of glioma angiogenesis. Glioma vasculature is quite significant for the initiation and progression of such disease (Hardee and Zagzag, 2012). The methylation of related functional regions of such gene definitely affect its biological functions, which further plays an irreplaceable role for GBM pathogenesis. Therefore, such target gene can be an effective GBM associated gene.

Apart from probe **cg26558485**, another probe named as **cg07028914** targets a transcription factor named as ***TBX18***. According to recent publications, an independent study in 2015 confirmed that microRNA miR-205 prevent the invasion of glioma by targeting *TBX18* (Zheng et al., 2015), reflecting the potential regulatory role of *TBX18* during glioma pathogenesis. Most of microRNAs' biological effects on glioma pathogenesis relied on the regulation on gene expression, which is similar with methylation mediated biological processes. Therefore, the methylation status of such gene may also play a potential regulatory role for the invasion of glioma.

The next gene, ***SP5*** is targeted by multiple probes including **cg26766005** and **cg14768335**. According to recent publications,

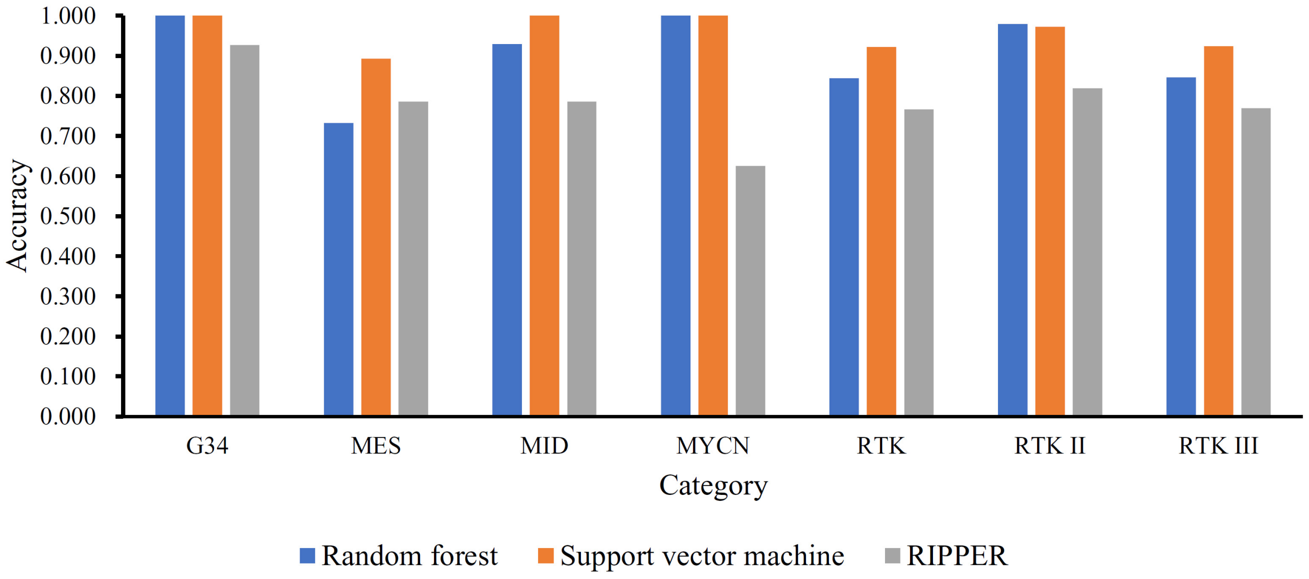

FIGURE 3 | Performance of the optimal SVM, RF, and RIPPER classifiers on different categories in the training dataset. The optimal SVM and RF classifiers are much superior to the optimal RIPPER classifier, and the optimal SVM classifier is slightly superior to the optimal RF classifier.

SP5 has been shown to be therapeutic target and a prognostic biomarker for multiple cancer subtypes, including glioma (Safe and Abdelrahim, 2005; Safe et al., 2014). Considering that methylation can regulate the expression level and biological effects of a target gene, the methylation status of the regulatory region of such gene may also probably affect the pathogenesis of glioma and have different pathological effects in different glioma subgroups indirectly.

As for ***TMEM22***, also known as *SLC35G2*, which is targeted by the optimal features **cg25836094, cg13383019**, and **cg22304507**, it has been generally reported to participate in cell proliferation and tumorigenesis with few publications (Dobashi et al., 2009). Although such gene has not been directly reported to be functionally correlated with glioma, it has been widely reported to be associated with renal cell carcinoma and its homolog which shared similar biological functions, *TMEM97* has been directly confirmed to be correlated with glioma at transcriptomics level. Considering that methylation at gene body is correlated with gene transcription, it is reasonable for us to regard *TMEM22* associated probes as potential glioma associated probes.

The next identified probe turns out to be **cg11823511**, targeting gene ***BARHL2***. According to two independent studies reported by researchers from University of Birmingham (Dunwell et al., 2010) and Memorial Sloan-Kettering Cancer Center (Shen et al., 2012), respectively, the methylation of *BARHL2* is not only related to hematological and epithelial cancers, but nerve system malignancies including glioma and may play a specific role for the integrative subgrouping of glioma (Shen et al., 2012).

TABLE 3 | Performance of the optimal SVM, RF, and RIPPER classifiers on the independent test dataset.

Classification algorithm	Overall accuracy	MCC
SVM	0.852	0.798
RF	0.877	0.832
RIPPER	0.954	0.937

RASGRF2 targeted by probe **cg06829830** has also been predicted to be contribute to the pathogenesis of glioma at methylation level. According to recent publications, in 2019, a systematic review (Wu et al., 2014) on the cancer methylation biomarkers confirmed that such gene is a specific biomarker for aggressive gliomas at methylation level using liquid biopsy.

Apart from such gene, the next identified biomarker is ***TLX3***, targeted by probe **cg26844246**. The methylation alteration of such gene has been identified in multiple tumor subtypes, like thyroid cancer (Kikuchi et al., 2013), bladder cancer and lung adenocarcinoma (Pradhan et al., 2013). In a systematic study on the whole-genome wide glioma methylation status, *TLX3* has been shown with specific methylation status in level II and III gliomas (Suzuki et al., 2015).

As for gene ***ANKRD34A*** (correlated with probes **cg10178263, cg18280463**, and **cg13947666**), according to related methylation studies (Giri and Aittokallio, 2018; Ding et al., 2020), such gene has shown to have methylation changes during the initiation and progression of multiple tumor subtypes, including lung, colon, bladder, lymphoma, breast and ovarian cancer. Therefore, it is reasonable for us to connect the methylation status of *ANKRD34A* with glioma. Apart from that, a recent publication (Ding et al., 2020) in 2020 also indicated that the transcript of such gene, which is regulated by methylation status, may participate in the RNA regulatory network in low grade glioma. Therefore, the methylation of such gene may be correlated with glioma and performed differentially in different subgroups.

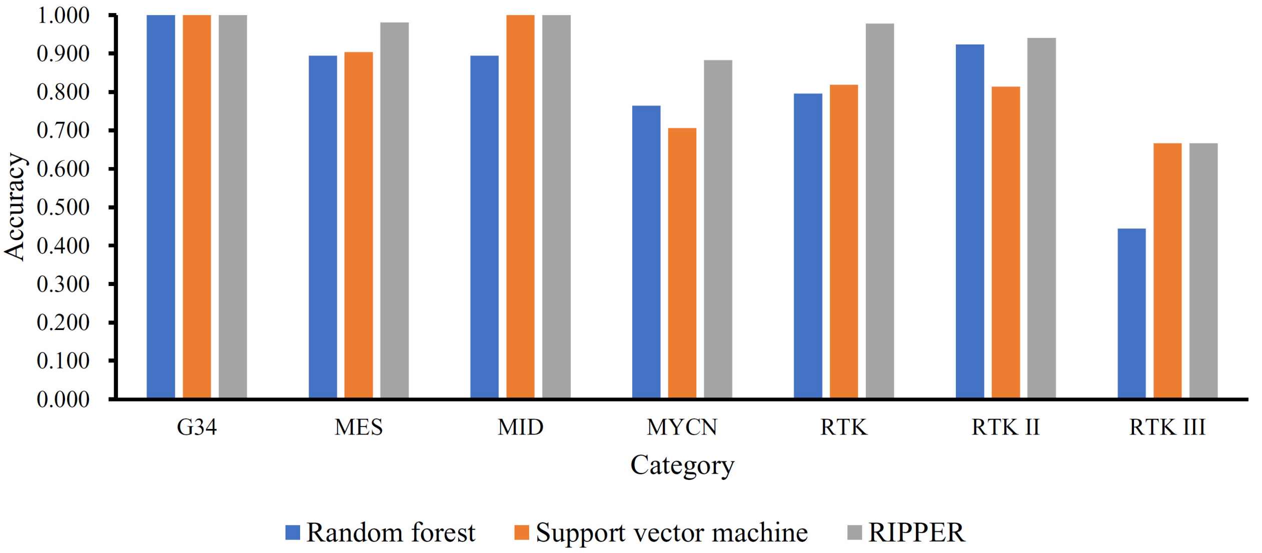

FIGURE 4 | Performance of the optimal SVM, RF, and RIPPER classifiers on different categories in the independent test dataset. The optimal RIPPER classifier gives the best generalizability on the independent test dataset, followed by the optimal RF and SVM classifier.

The last target of the optimal probes is ***MARCH11*** (targeted by probe **cg09017434**), regulating the intracellular transport of lysines. As for its correlations with GBM, according to recent publications, such gene has shown to be correlated with the carcinogenic transformation of cells with different expression levels (Yang et al., 2020). Considering the correlations between gene region methylation and gene expression, it is reasonable for us to speculate that the methylation status of such gene may be correlated with potential malignant alterations, supporting its correlations with GBM.

Biological Functions Relevant to GBM Based on Optimal Genes

Here, to summarize the specific biological functions that may contribute to revealing the differences between different GBM subgroups at methylation level, we performed GO enrichment analyses and pathway analyses on the optimal genes associated with GBM related probes (see **Supplementary Table 4**).

For the GO enrichment analyses results, firstly nervous system development has been screened out. Nervous system development is a biological process related to GBM. The malignant transformation and invasive migration of glioma cells rely on basic cellular components and physical anatomical structure. Therefore, the nervous system may contain proteins that are crucial for GBM. A recent publication confirmed that MT1-MMP, a major component of nervous system development, plays an important role during the pathogenesis of GBM (Beliën et al., 1999). Nervous system development is also associated with DNA methylation. Specific patterns have been seen at the DNA methylation level in the nervous system during the development and pathogenesis of GBM. Some patterns are even shared by two groups (Numata et al., 2012). Therefore, nervous system development, as an effective biological process, can be predicted to contribute to the description of GBM, validating the efficacy and accuracy of our prediction.

Apart from that, the next enriched term calcium ion binding has also been shown to be related to GBM. Various important cells in the central nervous system and the pathogenesis of GBM-like astrocytes participate in complicated metabolite transportation from the blood to the brain. Under pathogenic conditions, glioma cells seize control of the regulation of vascular tone through the Ca^{+}-dependent release of K^{+}, suggesting that calcium ion binding and blood stream in the brain in pathogenic status have important clinical implications (Watkins et al., 2014). Calcium ion binding is also related to methylation. An increase in the ionic strength and a decrease in the methylation reduce the amount of calcium required for the gelation of pectin–calcium systems (Garnier et al., 1993).

Biological Pathways Relevant to GBM Based on Optimal Genes

Apart from GO enrichment analyses, we also performed KEGG pathway analyses on such optimal genes (see **Supplementary Table 4**). The results of this study indicated that alcoholism is related to glioblastoma. Repurposing disulfiram (DSF) is a drug that has been widely used over the past several years to control alcoholism. DSF can inhibit the growth of GBM cells with TMZ resistance without affecting normal cells in the human central nervous system. DSF can suppress the growth and self-renewal of primary cells from GBM tumors, suggesting that an association exists between alcoholism and GBM (Triscott et al., 2012). Alcoholism is also related to the methylation alteration of transporter genes. Methylation status is further affected by alcoholism. The methylation of DAT in peripheral blood has also been validated to be a biomarker for alcohol-dependent patients (Wiers et al., 2015).

CONCLUSION

We found several methylation features (sites) associated with the classification of GBM using our newly presented computational method for classifying glioblastoma subtypes on the basis of gene methylation level. Through the further functional enrichment analysis of dysmethylated genes, such as *CXCR4*, *TBX18*, *SP5*, and *TMEM22*, several potential pathogenic functions are found to participate in the initiation and progression of glioblastoma. These functions include nervous system development, intrinsic plasma membrane component, systemic lupus erythematosus, and alcoholism.

DATA AVAILABILITY STATEMENT

Two sets of methylation profiles of patients with GBM were downloaded from Gene Expression Omnibus (GEO) with the accession numbers GSE90496 and GSE109379.

AUTHOR CONTRIBUTIONS

TH and Y-DC designed the study. Y-HZ, ZL, TZ, and XP performed the experiments. Y-HZ, ZL, LC, DL, and HL analyzed the results. Y-HZ and ZL wrote the manuscript. All authors contributed to the research and reviewed the manuscript.

FUNDING

This work was supported by the Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), National Key R&D Program of China (2018YFC0910403), National Natural Science Foundation of China (31701151), Shanghai Sailing Program (16YF1413800), the Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) (2016245), and the Fund of the Key Laboratory of Tissue Microenvironment and Tumor of Chinese Academy of Sciences (202002).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.604336/full#supplementary-material

Supplementary Table 1 | Features ranked by MCFS.

Supplementary Table 2 | 10-fold cross-validation performance of IFS with SVM, RF, and RIPPER when using different number of features ranked by MCFS.

Supplementary Table 3 | Genes corresponding to the top-ranked features by MCFS.

Supplementary Table 4 | Enrichment analysis results on 1813 coding genes.

REFERENCES

Beliën, A. T. J., Paganetti, P. A., and Schwab, M. E. (1999). Membrane-type 1 Matrix Metalloprotease (MT1-MMP) Enables Invasive Migration of Glioma Cells in Central Nervous System White Matter. *J. Cell Biol.* 144, 373–384. doi: 10.1083/jcb.144.2.373

Breiman, L. (2001). Random forests. *Machine Learning* 45, 5–32.

Bylund, J., Zhang, C., and Harder, D. R. (2002). Identification of a novel cytochrome P450. CYP4X1, with unique localization specific to the brain. *Biochem. Biophys. Res. Commun.* 296, 677–684. doi: 10.1016/s0006-291x(02)00918-x

Capper, D., Jones, D. T. W., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., et al. (2018). DNA methylation-based classification of central nervous system tumours. *Nature* 555, 469–474.

Chen, L., Pan, X. Y., Guo, W., Gan, Z. J., Zhang, Y. H., Niu, Z. B., et al. (2020). Investigating the gene expression profiles of cells in seven embryonic stages with machine learning algorithms. *Genomics* 112, 2524–2534. doi: 10.1016/j.ygeno.2020.02.004

Chen, L., Wang, S., Zhang, Y.-H., Li, J., Xing, Z.-H., Yang, J., et al. (2017). Identify key sequence features to improve CRISPR sgRNA efficacy. *IEEE Access* 5, 26582–26590. doi: 10.1109/access.2017.2775703

Cohen, W. W. (1995). "Fast effective rule induction," in *The Twelfth International Conference on Machine Learning* (San Francisco, SF: Morgan Kaufmann Publishers Inc), 115–123. doi: 10.1016/b978-1-55860-377-6.50023-2

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Machine Learning* 20, 273–297.

Ding, Y., Liu, H., Zhang, C., Bao, Z., and Yu, S. (2020). Comprehensive Analysis of Prognostic lncRNAs, miRNAs, and mRNAs Forming a Competing Endogenous RNA Network in LGG. Preprint.

Dobashi, S., Katagiri, T., Hirota, E., Ashida, S., Daigo, Y., Shuin, T., et al. (2009). Involvement of TMEM22 overexpression in the growth of renal cell carcinoma cells. *Oncol. Rep.* 21, 305–312.

Dramiński, M., Rada-Iglesias, A., Enroth, S., Wadelius, C., Koronacki, J., and Komorowski, J. (2007). Monte Carlo feature selection for supervised classification. *Bioinformatics* 24, 110–117. doi: 10.1093/bioinformatics/btm486

Dunwell, T., Hesson, L., Rauch, T. A., Wang, L., Clark, R. E., Dallol, A., et al. (2010). A genome-wide screen identifies frequently methylated genes in haematological and epithelial cancers. *Mol. Cancer* 9:44. doi: 10.1186/1476-4598-9-44

Frank, E., Hall, M., Trigg, L., Holmes, G., and Witten, I. H. (2004). Data mining in bioinformatics using Weka. *Bioinformatics* 20, 2479–2481. doi: 10.1093/bioinformatics/bth261

Garnier, C., Axelos, M. A. V., and Thibault, J.-F. (1993). Phase diagrams of pectin-calcium systems: Influence of pH, ionic strength, and temperature on the gelation of pectins with different degrees of methylation. *Carbohydrate Res.* 240, 219–232. doi: 10.1016/0008-6215(93)84185-9

Giri, A. K., and Aittokallio, T. (2018). DNMT inhibitors increase methylation at subset of CpGs in colon, bladder, lymphoma, breast, and ovarian, cancer genome. *bioRxiv*:395467. Preprint.

Gorodkin, J. (2004). Comparing two K-category assignments by a K-category correlation coefficient. *Comput. Biol. Chem.* 28, 367–374. doi: 10.1016/j.compbiolchem.2004.09.006

Guadagno, E., Vitiello, M., Francesca, P., Calì, G., Caponnetto, F., Cesselli, D., et al. (2017). PATZ1 is a new prognostic marker of glioblastoma associated with the stem-like phenotype and enriched in the proneural subtype. *Oncotarget* 8, 59282–59300. doi: 10.18632/oncotarget.19546

Guardiola Bagán, S., Seco, J., Varese, M., Díaz Lobo, M., García Arroyo, J., Teixidó Turà, M., et al. (2017). Toward a novel drug to target the EGF-EGFR interaction: design of metabolically stable bicyclic peptides. *ChemBioChem* 19, 76–84. doi: 10.1002/cbic.201700519

Hardee, M. E., and Zagzag, D. (2012). Mechanisms of glioma-associated neovascularization. *Am. J. Pathol.* 181, 1126–1141. doi: 10.1016/j.ajpath.2012.06.030

Jia, Y., Zhao, R., and Chen, L. (2020). Similarity-Based Machine Learning Model for Predicting the Metabolic Pathways of Compounds. *IEEE Access* 8, 130687–130696. doi: 10.1109/access.2020.3009439

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., and Murthy, K. R. K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Comput.* 13, 637–649. doi: 10.1162/089976601300014493

Kikuchi, Y., Tsuji, E., Yagi, K., Matsusaka, K., Tsuji, S., Kurebayashi, J., et al. (2013). Aberrantly methylated genes in human papillary thyroid cancer and their association with BRAF/RAS mutation. *Front. Genet.* 4:271. doi: 10.3389/fgene.2013.00271

Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *International Joint Conference on Artificial Intelligence* (Lawrence: Erlbaum Associates Ltd), 1137–1145.

Lai, R. K., Chen, Y., Guan, X., Nousome, D., Sharma, C., Canoll, P., et al. (2014). Genome-Wide Methylation Analyses in Glioblastoma Multiforme. *PLoS One* 9:e89376. doi: 10.1371/journal.pone.0089376

Liang, H., Chen, L., Zhao, X., and Zhang, X. (2020). Prediction of drug side effects with a refined negative sample selection strategy. *Comput. Math. Methods Med.* 2020:1573543.

Liu, H. A., and Setiono, R. (1998). Incremental feature selection. *Appl. Intell.* 9, 217–230.

Liu, H., Hu, B., Chen, L., and Lu, L. (2020). *Identifying protein subcellular location with embedding features learned from networks. Current Proteomics*. Sharjah: Bentham Science Publishers.

Macyszyn, L., Akbari, H., Pisapia, J. M., Da, X., Attiah, M., Pigrish, V., et al. (2016). Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques. *Neuro Oncol.* 18, 417–425. doi: 10.1093/neuonc/nov127

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica Biophysica Acta BBA Protein Struct.* 405, 442–451. doi: 10.1016/0005-2795(75)90109-9

Muthukrishnan, S., Puri, M., and Lefevre, C. (2014). Support vector machine (SVM) based multiclass prediction with basic statistical analysis of plasminogen activators. *BMC Res. Notes* 7:63. doi: 10.1186/1756-0500-7-63

Numata, S., Ye, T., Hyde, T. M., Guitart-Navarro, X., Tao, R., Wininger, M., et al. (2012). DNA Methylation Signatures in Development and Aging of the Human Prefrontal Cortex. *Am. J. Hum. Genet.* 90, 260–272. doi: 10.1016/j.ajhg.2011.12.020

Pan, X. Y., Zhang, Y. N., and Shen, H. B. (2010). Large-scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features. *J. Proteome Res.* 9, 4992–5001. doi: 10.1021/pr100618t

Platt, J. (ed.) (1998). *Fast training of support vector machines using sequential minimal optimization*. Cambridge, MA: MIT Press.

Pradhan, M. P., Desai, A., and Palakal, M. J. (2013). Systems biology approach to stage-wise characterization of epigenetic genes in lung adenocarcinoma. *BMC Syst. Biol.* 7:141. doi: 10.1186/1752-0509-7-141

Safe, S., and Abdelrahim, M. (2005). Sp transcription factor family and its role in cancer. *Eur. J. Cancer* 41, 2438–2448. doi: 10.1016/j.ejca.2005.08.006

Safe, S., Imanirad, P., Sreevalsan, S., Nair, V., and Jutooru, I. (2014). Transcription factor Sp1, also known as specificity protein 1 as a therapeutic target. *Expert Opin. Therapeut. Targets* 18, 759–769. doi: 10.1517/14728222.2014.914173

Sang, X., Xiao, W., Zheng, H., Yang, Y., and Liu, T. (2020). HMMPred: Accurate Prediction of DNA-Binding Proteins Based on HMM Profiles and XGBoost Feature Selection. *Comput. Math. Methods Med.* 2020:1384749.

Sharma, A., Bendre, A., Mondal, A., Muzumdar, D., Goel, N., and Shiras, A. (2017). Angiogenic Gene Signature Derived from Subtype Specific Cell Models Segregate Proneural and Mesenchymal Glioblastoma. *Front. Oncol.* 7:146. doi: 10.3389/fonc.2017.00146

Shen, R., Mo, Q., Schultz, N., Seshan, V. E., Olshen, A. B., Huse, J., et al. (2012). Integrative subtype discovery in glioblastoma using iCluster. *PLoS One* 7:e35236. doi: 10.1371/journal.pone.0035236

Skiriute, D., Vaitkiene, P., Saferis, V., Asmoniene, V., Skauminas, K., Deltuva, V. P., et al. (2012). MGMT, GATA6, CD81, DR4, and CASP8 gene promoter methylation in glioblastoma. *BMC Cancer* 12, 218–218. doi: 10.1186/1471-2407-12-218

Suzuki, H., Aoki, K., Chiba, K., Sato, Y., Shiozawa, Y., Shiraishi, Y., et al. (2015). Mutational landscape and clonal architecture in grade II and III gliomas. *Nat. Genet.* 47, 458–468. doi: 10.1038/ng.3273

Triscott, J., Lee, C., Hu, K., Fotovati, A., Berns, R., Pambid, M., et al. (2012). Disulfiram, a drug widely used to control alcoholism, suppresses self-renewal of glioblastoma and overrides resistance to temozolomide. *Oncotarget* 3, 1112–1123. doi: 10.18632/oncotarget.604

Wang, C., Chen, Y., Wang, Y., Liu, X., Liu, Y., Li, Y., et al. (2019). Inhibition of COX-2, mPGES-1 and CYP4A by isoliquiritigenin blocks the angiogenic Akt signaling in glioma through ceRNA effect of miR-194-5p and lncRNA NEAT1. *J. Experimen. Clin. Cancer Res.* 38, 1–14.

Wang, C., Li, Y., Chen, H., Huang, K., Liu, X., Qiu, M., et al. (2018). CYP4X1 inhibition by flavonoid CH625 normalizes glioma vasculature through reprogramming TAMs via CB2 and EGFR-STAT3 Axis. *J. Pharmacol. Experimen. Therapeut.* 365, 72–83. doi: 10.1124/jpet.117.247130

Watkins, S., Robel, S., Kimbrough, I. F., Robert, S. M., Ellis-Davies, G., and Sontheimer, H. (2014). Disruption of astrocyte-vascular coupling and the blood-brain barrier by invading glioma cells. *Nat. Commun.* 5:4196.

Wiers, C. E., Shumay, E., Volkow, N. D., Frieling, H., Kotsiari, A., Lindenmeyer, J., et al. (2015). Effects of depressive symptoms and peripheral DAT methylation on neural reactivity to alcohol cues in alcoholism. *Transl. Psychiatry* 5:e648. doi: 10.1038/tp.2015.141

Witten, I. H., and Frank, E. (eds) (2005). Data Mining:Practical Machine Learning Tools and Techniques. San Francisco, SF: Morgan Kaufmann Publishers.

Wu, G., Diaz, A. K., Paugh, B. S., Rankin, S. L., Ju, B., Li, Y., et al. (2014). The genomic landscape of diffuse intrinsic pontine glioma and pediatric non-brainstem high-grade glioma. *Nat. Genet.* 46:444. doi: 10.1038/ng.2938

Würth, R., Bajetto, A., Harrison, J. K., Barbieri, F., and Florio, T. (2014). CXCL12 modulation of CXCR4 and CXCR7 activity in human glioblastoma stem-like cells and regulation of the tumor microenvironment. *Front. Cell. Neurosci.* 8:144. doi: 10.3389/fncel.2014.00144

Yadav, V. N., Zamler, D., Baker, G. J., Kadiyala, P., Erdreich-Epstein, A., Decarvalho, A. C., et al. (2016). CXCR4 increases in-vivo glioma perivascular invasion, and reduces radiation induced apoptosis: A genetic knockdown study. *Oncotarget* 7, 83701–83719. doi: 10.18632/oncotarget.13295

Yang, Q., Li, K., Li, X., and Liu, J. (2020). Identification of Key Genes and Pathways in Myeloma side population cells by Bioinformatics Analysis. *Int. J. Med. Sci.* 17:2063. doi: 10.7150/ijms.48244

Yuan, F., Pan, X. Y., Zeng, T., Zhang, Y. H., Chen, L., Gan, Z. J., et al. (2020). Identifying Cell-Type Specific Genes and Expression Rules Based on Single-Cell Transcriptomic Atlas Data. *Front. Bioengine. Biotechnol.* 8:350. doi: 10.3389/fbioe.2020.00350

Zhao, X., Chen, L., and Lu, J. (2018). A similarity-based method for prediction of drug side effects with heterogeneous information. *Math. Biosci.* 306, 136–144. doi: 10.1016/j.mbs.2018.09.010

Zheng, G., Jia, X., Peng, C., and Zhimin, H. (2015). miR-205 inhibits invasion of glioma cells via targeting TBX18. *Chin. J. Pathophysiol.* 31, 1219–1224.

Zhou, J.-P., Chen, L., and Guo, Z.-H. (2020a). iATC-NRAKEL: An efficient multi-label classifier for recognizing anatomical therapeutic chemical classes of drugs. *Bioinformatics* 36, 1391–1396.

Zhou, J.-P., Chen, L., Wang, T., and Liu, M. (2020b). iATC-FRAKEL: A simple multi-label web-server for recognizing anatomical therapeutic chemical classes of drugs with their fingerprints only. *Bioinformatics* 36, 3568–3569. doi: 10.1093/bioinformatics/btaa166

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.