Integration of Multi-Omics Data for the Classification of Glioma Types and Identification of Novel Biomarkers

Bioinformatics and Biology Insights
Volume 18: 1–14

DOI: 10.1177/11779322241249563

Francisca G Vieira[1], Regina Bispo[1,2] and Marta B Lopes[1,2,3]

[1]Center for Mathematics and Applications (NOVA Math), NOVA School of Science and Technology, Caparica, Portugal. [2]Department of Mathematics, NOVA School of Science and Technology, Caparica, Portugal. [3]UNIDEMI, Department of Mechanical and Industrial Engineering, NOVA School of Science and Technology, Caparica, Portugal.

ABSTRACT: Glioma is currently one of the most prevalent types of primary brain cancer. Given its high level of heterogeneity along with the complex biological molecular markers, many efforts have been made to accurately classify the type of glioma in each patient, which, in turn, is critical to improve early diagnosis and increase survival. Nonetheless, as a result of the fast-growing technological advances in high-throughput sequencing and evolving molecular understanding of glioma biology, its classification has been recently subject to significant alterations. In this study, we integrate multiple glioma omics modalities (including mRNA, DNA methylation, and miRNA) from The Cancer Genome Atlas (TCGA), while using the revised glioma reclassified labels, with a supervised method based on sparse canonical correlation analysis (DIABLO) to discriminate between glioma types. We were able to find a set of highly correlated features distinguishing glioblastoma from lower-grade gliomas (LGGs) that were mainly associated with the disruption of receptor tyrosine kinases signaling pathways and extracellular matrix organization and remodeling. Concurrently, the discrimination of the LGG types was characterized primarily by features involved in ubiquitination and DNA transcription processes. Furthermore, we could identify several novel glioma biomarkers likely helpful in both diagnosis and prognosis of the patients, including the genes *PPP1R8, GPBP1L1, KIAA1614, C14orf23, CCDC77, BVES, EXD3, CD300A*, and *HEPN1*. Collectively, this comprehensive approach not only allowed a highly accurate discrimination of the different TCGA glioma patients but also presented a step forward in advancing our comprehension of the underlying molecular mechanisms driving glioma heterogeneity. Ultimately, our study also revealed novel candidate biomarkers that might constitute potential therapeutic targets, marking a significant stride toward personalized and more effective treatment strategies for patients with glioma.

KEYWORDS: Canonical correlation analysis, classification, glioma, multi-omics, survival analysis

RECEIVED: July 29, 2023. **ACCEPTED:** April 9, 2024.

TYPE: Original Research Article

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded by national funds through the FCT—Fundação para a Ciência e a Tecnologia, I.P., with references CEECINST/00042/2021, UIDB/00297/2020, and UIDP/00297/2020 (NOVA Math), UIDB/00667/2020 and UIDP/00667/2020 (UNIDEMI), and under the scope of the research project "MONET—Multi-omic networks in gliomas" (PTDC/CCI-BIO/4180/2020).

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Marta B Lopes, Center for Mathematics and Applications (NOVA Math), NOVA School of Science and Technology, Caparica 2829-516, Portugal.
Email: marta.lopes@fct.unl.pt

Introduction

Glioma is one of the most common malignant primary brain tumors in adults, comprising multiple types with distinct biomolecular characteristics. Given the low overall patient survival, there is an urgent growing need for new effective treatment approaches. However, the marked differences in prognosis and therapeutic outcomes still represent major clinical challenges.[1] To understand the heterogeneity among gliomas, tremendous efforts have been made to identify the different adult-type diffuse types,[2-4] which can be broadly classified as lower-grade gliomas (LGGs) according to a lower proliferative activity or glioblastoma (GBM) which exhibits high malignancy traits. Deriving from extensive research into the molecular profiling of glioma types, its classification has recently undergone significant changes. The World Health Organization (WHO) Classification of Tumors of the Central Nervous System (CNS) has now started to incorporate additional driver molecular markers in their revised guidelines, obtained via information extracted from next-generation sequencing technologies, such as gene expression and DNA methylome profiling, to further refine their characterization.[5]

Recent advances in such high-throughput technologies have been allowing the extraction of an increasing quantity of biological data, which is not limited to one but instead includes multiple omics modalities collected from the same individuals. The synergy between these technological breakthroughs and publicly accessible multi-omics databases, such as TCGA, prove particularly valuable for acquiring new knowledge about glioma characterization. However, the diagnostic categories reported in the glioma datasets still do not align with the newly published WHO-2021 guidelines. Therefore, using the revised glioma reclassification along with a joint integrative study of the different but linked layers of genetic regulation offers an opportunity to build a more comprehensive landscape of the biological mechanisms underlying heterogeneity and further enhance the molecular understanding of disease, where research has relied mostly on single-omics data.[6] However, given the high-dimensional nature of this type of data, its multivariate analysis is particularly challenging.

Several methods have been proposed to address multi-omics data integration and analysis. Joint dimensionality reduction (JDR) techniques transform the data into a lower-dimensional

space based on matrix decomposition, learning latent factors that capture biological sources of common variability across various omics datasets. These techniques include, for instance, sparse variants of both canonical correlation analysis (CCA)[7,8] and multiblock partial least squares (PLS),[9] joint nonnegative matrix factorization (NMF),[10] joint and individual variation explained (JIVE)[11,12] and multi-omics factor analysis (MOFA).[13,14] Furthermore, supervised extensions of CCA have been recently suggested in a classification framework to identify relevant multiview features that are not only highly correlated but also optimally separate subjects into distinct groups.[15-17] Given that biological processes are inherently interconnected between different layers of genetic regulation, the utilization of correlation-based methods becomes particularly pertinent when working with omics datasets, providing a robust framework to uncover intricate interactions, study dependencies between features and identify co-regulated molecular patterns. In particular, data integration analysis for biomarker discovery using latent components (DIABLO)[15] has been used in multiple omics applications, including the identification of disease mechanisms and biomarkers.[18,19] Moreover, a recent benchmark analysis in cancer type classification[20] revealed that DIABLO performed significantly better than several other methods, including MOFA/MOFA+,[14] principal component analysis (PCA), and iClusterBayes/iClusterPlus/iCluster.[21]

In this work, an integrative multilayer omics study using mRNA, DNA methylation, and miRNA data from The Cancer Genome Atlas (TCGA) was performed to classify and characterize the different glioma types, considering its recently updated reclassification.[22,23] First, in section "Materials and Methods," we describe the data collected and detail all the methods used in this work. We used DIABLO to select a subset of highly correlated features from the different modalities relevant to distinguish either the 3 glioma types (as present in section "Integration of two omics layers for the classification of glioma types") or specifically to discriminate the 2 LGG types (section "Discrimination of LGG types using three omics layers"). This is particularly relevant, considering the ongoing evolution in glioma classification and the need of identifying novel genetic markers that aid in its characterization. Aiming to delineate unknown mechanisms of glioma pathobiology and unveil novel therapeutic targets, we further investigated the correlation between the selected features, the pathways involved, and the prognostic value based on their effect in the survival probability of the patients (section "Investigation on the selected features"). Finally, in sections "Discussion" and "Conclusions," we discuss the obtained results and highlight the main conclusions of the work developed, respectively.

Materials and Methods

Data description

In this study, we used the GBM and LGG datasets from TCGA, available in the Genomic Data Commons Data Portal with project names TCGA-GBM and TCGA-LGG. The TCGA level-3 data on gene expression (Illumina mRNAseq), DNA methylation (Illumina HumanMethylation450), miRNA expression (Illumina miRNAseq, BCGSC miRNA profiling workflow) and clinical information were retrieved. RNA-seq data normalized for gene length and for sequencing depth (Transcript per million, TPM) with upper quantile normalization was extracted for supervised and sparse CCA using DIABLO and survival analysis. RNA-seq RSEM expected counts, without normalization, were extracted for differential gene expression analysis. The R packages RTCGAToolbox v2.28.1[24] and TCGAbiolinks 2.25.3[25] were used to collect the data.

DNA methylation data were first preprocessed by removing nonvalid entries (start = end = −1), probes mapping multiple places, non-CpG sites, sexual chromosome and missing (NA) entries. For each data set, features with zero variance or having very few unique values relative to the number of samples (cutoff of 10%) and a large ratio of the frequency of the most common value to the frequency of the second most common value (cutoff ratio of 95/5) were filtered. The intersection of the data sets was done to keep only the samples that were present in all the data sets. From the 278 GBM individuals, 108 had both mRNA and DNA methylation data, and 5 had miRNA data. From the 262 astrocytoma individuals, 248 had both mRNA and DNA methylation data, and 243 had miRNA data, while from the 169 oligodendroglioma individuals, 166 had mRNA, DNA methylation and miRNA data. Provided that there were only 5 glioblastoma individuals with miRNA data available, we first used mRNA and DNA methylation to separate the 3 glioma classes (Case study 1) and afterwards the 3 data types to separate the 2 LGG classes (Case study 2). The number of patients included in each of these steps of the data analysis, as well as the number of features, are summarized in Table 1.

In the classification task, the data were split into train and test subsets, in a predefined ratio (70% for training and 30% for testing), while preserving relative ratios of the different classes (the different glioma types are in the same proportion as the original dataset).

Data integration analysis for biomarker discovery using latent components

In this work, we applied DIABLO,[15] using the R package mixOmics v6.22.0,[26] to select co-expressed variables from the different datasets that discriminate between the glioma types. This method is based on CCA, a well-established multivariate method, that aims to subtract linear combinations of variables (canonical variates), from 2 data sources, in a way that the canonical variates maximally correlate with each other.[27] After transformation, the complex high-dimensional variable sets are projected into the common latent subspace with the desired lower dimension. On top of that, generalized versions of CCA can perform the integration of multiple (more than 2) datasets, by finding a linear subspace which maximizes the correlations between all the views and the linear combinations among several blocks of variables. In particular, DIABLO discriminant analysis

Table 1. Data description: data types included in each data analysis step (case study), number of total and filtered features, and number of individuals per glioma type.

		FEATURES		INDIVIDUALS		
		TOTAL	FILTERED	ASTRO	OLIGO	GBM
Case study 1	mRNA	20501	19068	248	166	108
	DNA methylation	297602	297602			
Case study 2	mRNA	20501	19049	243	166	-
	DNA methylation	297602	297602			
	miRNA	1881	1118			

Abbreviations: Astro: astrocytoma; Oligo: oligodendroglioma; GBM: glioblastoma.

extends the multivariate methodology generalized canonical correlation analysis (GCCA)[8] to a supervised framework by replacing one data matrix $X^{(q)}$ with the outcome dummy matrix Y. For Q omics datasets measuring the expression levels of the P_q omics variables on the same biological samples, GCCA solves for each component (canonical variate) h:

$$\max_{\mathbf{a}_h^{(1)},\ldots,\mathbf{a}_h^{(Q)}} \sum_{q,j=1,q\neq j}^{Q} c_{q,j}\ \mathrm{cov}\left(X_h^{(q)}\mathbf{a}_h^{(q)}, X_h^{(j)}\mathbf{a}_h^{(j)}\right),$$
$$\text{s.t. } \left\|\mathbf{a}_h^{(q)}\right\|_2 = 1 \text{ and } \left\|\mathbf{a}_h^{(q)}\right\|_1 \leq \lambda^{(q)} \quad (h = 1,\ldots, H)$$

where $X_h^{(q)}$ and $X_h^{(j)}$ represent the datasets, $\mathbf{a}_h^{(q)}$ and $\mathbf{a}_h^{(j)}$ are the loading vectors (component coefficients) on component h with $q, j = 1,\ldots,Q$ $(q \neq j)$. For all $\mathbf{a} \in \mathbb{R}^n$, ℓ_1 and ℓ_2 norms are, respectively, defined by $\|\mathbf{a}\|_1 = \sum_{j=1}^{p} |\mathbf{a}_j|$ and $\|\mathbf{a}\|_2 = \left(\sum_{j=1}^{p} \mathbf{a}_j^2\right)^{1/2}$. $\lambda^{(q)}$ is a ℓ_1 penalisation parameter to induce sparsity on the loading vector $\mathbf{a}_h$ by shrinking some coefficients to zero. One data matrix $X^{(q)}$ is given by the outcome dummy matrix, Y, indicating the class membership of each individual. $C = \{c_{q,j}\}_{q,j}$ is the design matrix, a $Q \times Q$ matrix that specifies whether data sets should be correlated and includes values between zero (data sets are not connected) and one (data sets are fully connected), enabling to model a particular relation between pairs of omics data and to constraint the model to only take into account those specific pairwise covariances, as expected from prior biological knowledge or experimental design. The design matrix was specified based on a data-driven approach. Using PLS as a preliminary analysis, integrating 2 datasets at a time to assess the common information between them, the correlation value was then used in $c_{q,j}$ for $q, j < Q$, while $c_{q,j} = 1$ for $q, j = Q$.

In this work, 2 components were used, since in general, $K - 1$ components are sufficient to achieve the best classification performance, where K is the number of classes.[26] It was also verified, by evaluating the difference in overall misclassification error rate, that there was no gain in performance when adding more components to the sparse model.

When using sparse GCCA, the component scores $t_h = X_h\mathbf{a}_h$ are defined on a small subset of variables with non-zero coefficients, leading to variable selection that aims to optimally maximize the discrimination between the K outcome classes in Y. In DIABLO, a soft-thresholding is used, replacing the non-negative parameter $\lambda^{(q)}$ that controls the amount of shrinkage in $\mathbf{a}_h$ by the number of features to select on each dimension.[26] To select the number of features, a step-by-step approach was used, by assessing the performance of the model (measured via overall misclassification error rate) for each value of features provided as a grid, one component at a time, using 5×5 cross-validation (CV). First, a grid with a higher amplitude of values but low resolution was evaluated, and subsequently finer grids were analyzed around the previously selected values with progressively smaller steps, until the values converged.

Considering an independent test set, the predicted coordinates (scores) are computed for each new observation on the set of H latent components and then used to predict each of the dummy variables. The final predicted class is then obtained, minimizing the distance to the centroid. The predictions are combined by weighted vote, where each omics dataset weight is defined as the correlation between the latent components associated to that particular data set and the outcome, from the training set. The final prediction is the class that obtains the highest weight across all datasets.

In this work, 30 different partitions of the data were considered, retrieving from each trained model the variables with nonzero coefficients and their absolute values, such as the performance measures (accuracy, precision, recall, F1) after predicting the labels in each test set. The receiver operating characteristic (ROC) curves were also used to evaluate the performance of each model and each component in predicting the glioma type, and the respective area under the curve (AUC) was determined.

Selection of features and database search

For further analyzes, we have selected the variables that, within the 30 different models, occurred more frequently in the

components with nonzero coefficient and with higher median absolute value: frequency >15 and median absolute value >0.05 or frequency >10, and median absolute value >0.2. The outcome class of each feature (the glioma type for which the expression of that feature is more distinguished from the other types) was determined as follows: For the RNA features, the respective outcome type was determined by differential gene expression analysis, where, for each gene, the outcome is the class in which it is more differentially expressed (with lower false discovery rate (FDR) and higher log(*FC*)). For the methylation features, the outcome was determined for each methylation site as the class with highest averaged methylation level difference to the other 2 classes. For the miRNA features, the outcome was determined as the class with higher averaged miRNA expression (in this case, there were only the 2 LGG types).

Each selected variable was searched for associations with glioma and cancer over different databases. In the case of DNA methylation features, they were first mapped back to gene symbols, using methylGSA v1.16.0 package[28] and the multisymbol checker tool from HUGO Gene Nomenclature Committee. The search was performed using all the human collections from the Molecular Signatures Database (MolSigDB)[29] and OMIM[30] with the keywords "glioma," "glioblastoma," "oligodendroglioma," "astrocytoma," and in PubMed, using as keywords the symbol of each gene and "AND glioma" or "AND cancer." The gene targets of each miRNA were determined using the MiRTarBase database.[31]

Differential expression analysis

Differential expression analysis was performed using the edgeR package version 3.40.2.[32] Genes with very low counts were first filtered, and trimmed mean of M values (TMM) normalization was applied to account for compositional biases.[33] Differential expression between the experimental groups was tested using quasi-likelihood F-test (considering a negative binomial distribution). Genes with an FDR-adjusted p (based on Benjamini and Hochberg (BH) adjustment)[34] of less than 0.05 were deemed significantly differentially expressed genes (DEGs). Furthermore, to define a compromise between statistical significance and biological variation, the genes for which:

$$-\log_{10}(FDR) > \max(-\log_{10}(FDR)) + \log(FC) \times[-\log_{10}(0.05) - \max(-\log_{10}(FDR))]$$

were considered positively differentially expressed while the genes for which:

$$-\log_{10}(FDR) > \max(-\log_{10}(FDR)) + \log(FC) \times[\log_{10}(0.05) + \max(-\log_{10}(FDR))]$$

were considered negatively differentially expressed. All the remaining DEGs, although significant were considered not relevant given the low log(*FC*).

Overrepresentation and gene set enrichment analysis

Differentially expressed genes that demonstrated significant fold changes between conditions were screened for KEGG pathways and gene ontology (GO) terms with over-representation enrichment analysis (ORA) in the software WebGestalT.[35] Using the same software, the full list of expressed genes ranked by a combination of statistical significance and fold-change:

$$-\log_{10} p \times sign(\log(FC))$$

was used in gene set enrichment analysis (GSEA). Terms and pathways were considered significant for an FDR < 0.05 (BH adjustment).

Correlation and survival analysis

The Pearson correlation (r) between each pair of selected features was computed using the respective values of expression in all the patients, with the R package Hmisc v5.0-1[36] and plotted using the R package pheatmap v1.0.12.[37]

A Kaplan-Meier survival curve was estimated using the clinical data from the patients of the TCGA-LGG and TCGA-GBM projects, with the R packages survival v3.5-3 for the analysis and survminer v0.4.9 for plotting.[38,39] For each evaluated gene (RNA and DNA methylation selected features), the tumor samples were split into 2 groups: low and high expression levels, using either the median value as cut-off or when existing 2 distinct density peaks, the threshold is set as the expression for which the density has the local minimum between the peaks. The statistical significance of survival differences in the analysis was assessed using the log-rank test, using the 0.05 significance level.

Results

Integration of 2 omics layers for the classification of glioma types

To discriminate between the 3 glioma types, namely GBM, astrocytoma, and oligodendroglioma, 2 components of mRNA and DNA methylation features were first obtained using DIABLO.

Regression analysis with PLS showed that the data sets were highly correlated (r = 0.8796), and therefore, the design matrix was set with a weight of 0.8. Based on the grid search, the number of mRNA variables with nonzero coefficient that was chosen were 41 (out of 19 068) for the first component and 7 for the second component, while the number of DNA

Table 2. Performance metrics for the classification of the 3 glioma types using 2 components based on RNA and DNA methylation features.

	PERFORMANCE METRICS			
GLIOMA TYPE	PRECISION	RECALL	F1	ACCURACY
Astrocytoma	0.996	0.964	0.980	
Glioblastoma	0.995	0.992	0.993	0.981
Oligodendroglioma	0.952	1.000	0.976	

methylation variables chosen was 11 (out of 297 602) for the first component and 41 for the second component.

The performance metrics evaluating the classification of the glioma types, based on the obtained components, are summarized in Table 2. Keeping only 2 components was sufficient to achieve a good overall accuracy of 98%, with the first component primarily differentiating GBM from the LGG and the second component mostly distinguishing between astrocytoma and oligodendroglioma, as depicted by Figure 1A. It is clear that GBM is considerably easier to identify because it has the highest F1 value and its ROC curve separating from the LGG types has an AUC close to 1 in all the components (Figure 1B).

Figure 1C shows the averaged expression levels of the selected features generally distinguish one of the classes from the other 2 (denoted as the outcome type). Moreover, features are highly correlated between blocks, where the methylation of some sites is mainly negatively correlated with the expression of other genes. The highly correlated blocks are displayed in Figure 3C and will be further described in section "Selection of features and database search."

We have then selected the variables that more frequently occurred in the components with nonzero coefficients and with higher loading. The selected variables are shown in Figure 1D (mRNA features in Figure 1D1 and DNA methylation features in Figure 1D2), ordered by the respective importance in the components (median absolute value) and grouped by outcome. The barplots also show whether each feature was reported in the literature or in biological databases with connections with glioma or cancer. Indeed, nearly all have been reported in cancer studies, and the vast majority have also been investigated in glioma. Moreover, most of the gene features selected were differentially expressed in GBM, while the selected methylation sites were more relevant for the discrimination of oligodendroglioma. For instance, some of the features that revealed to have the greatest influence in the distinction of GBM were the genes *KIAA0495*, *FBXO17*, *C9orf64*, *MSN* and *ARSD* and the methylated sites cg18222083 (*TMEM106A*), cg05211768 (*FES*), cg17105609 and cg05866411 (*FGFRL1*), and cg15603424 (*ARNTL*), all reported in glioma studies with the exception of the genes *ARSD* and *TMEM106A*, which were only reported in other types of cancer. The major implicated pathways and biological mechanisms encompass crucial processes such as apoptosis regulation, receptor tyrosine kinase (RTK) signaling, extracellular matrix (ECM) organization, and integrin pathway involvement. On the contrary, to discriminate astrocytoma, much fewer features were found to be relevant, including only 2 genes *DRG2* and *THRAP3*, both already reported in glioma studies, and the methylation sites cg24899806 (*KCND2*), cg19093820 (*GPR156*), cg26077062 (*SYBU*), cg04951819 (*CCDC77*) and cg03780927 (*BCL9L*), where *GPR156* and *CCDC77* were never reported in glioma, and the latter also never reported in any other cancer studies. Interestingly, 2 of the most important features to distinguish oligodendroglioma, the gene *FBX042* and the methylated gene *CD300LB* were also never reported in glioma studies. While exploring the molecular underpinnings of astrocytoma discrimination revealed associations with pathways regulating cell growth, differentiation, and neurotransmitter release, the characterization of oligodendroglioma conversely involved features associated with ubiquitination processes, mRNA processing, splicing, and various signaling pathways. A comprehensive exploration of the biological implications of these features, is provided in section "Discussion", shedding light on their roles and delving into the molecular mechanisms linked with glioma development and progression.

Discrimination of LGG types using 3 omics layers

Once shown in the previous analysis that the LGG were more challenging to distinguish, a further analysis keeping only these 2 classes and including an additional data view—miRNA—was performed with DIABLO, again using only 2 components. Regression analysis with PLS showed smaller correlation ($r = 0.6027$) between DNA methylation and miRNA than the one obtained for the other pairs of data, and therefore, the design matrix was set with a weight of 0.6 for this dataset pair and 0.8 for the others. Based on the grid search, the number of mRNA variables with nonzero coefficient that was chosen was 7 in both components, 29 DNA methylation variables for the first component and 42 for the second component, and 11 miRNA variables in both components.

The performance metrics evaluating the classification of the glioma types, based on the obtained components, are present in Table 3, and the AUC obtained are shown in Table 4. An accuracy of 97% was achieved, and the first component was able to distinguish by itself the classes (Figure 2A). Indeed, when observing the circos plot in Figure 2B, we can also see that the averaged expression levels of the selected features do not always

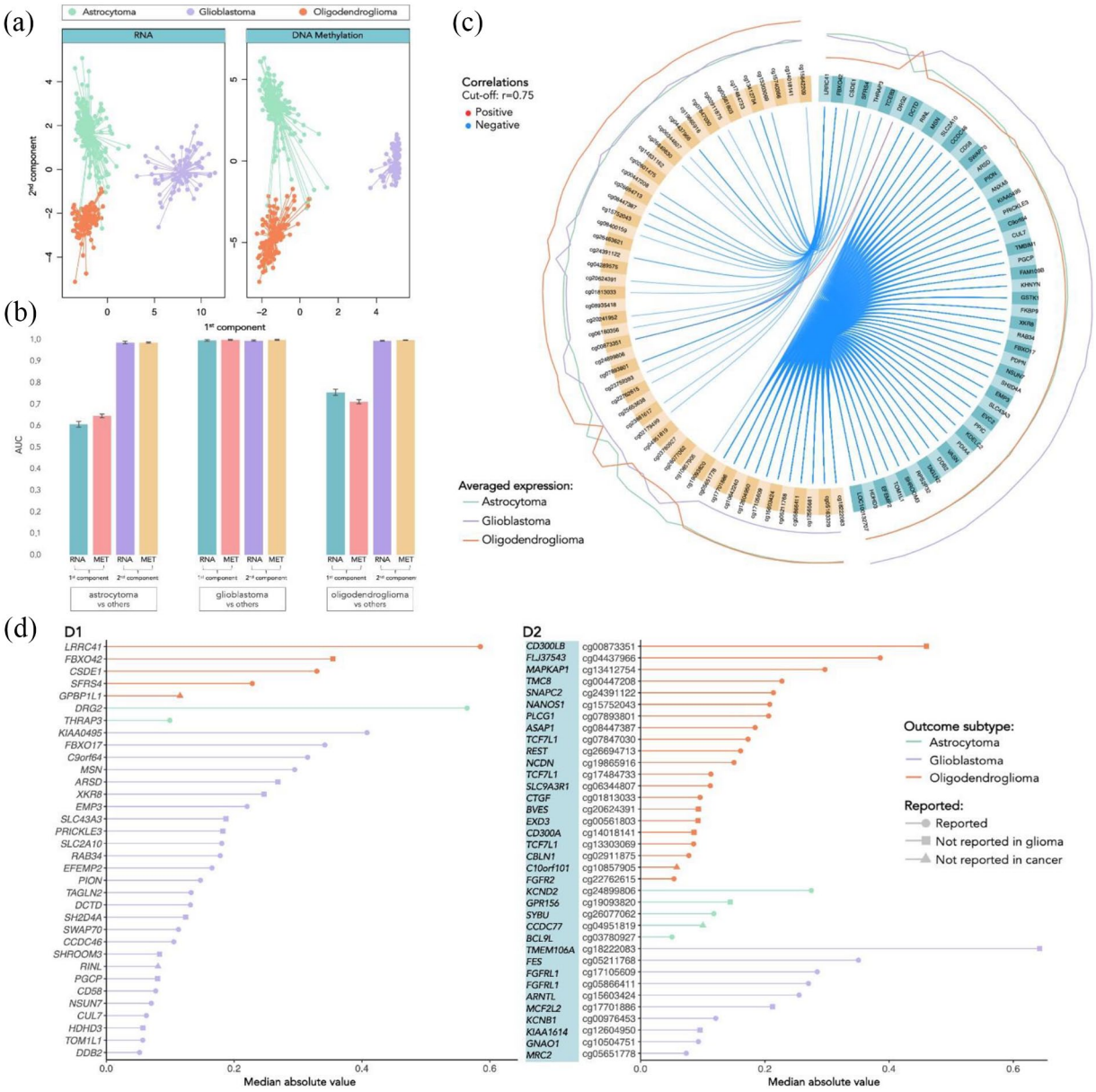


Figure 1. Separation and classification of 3 glioma types based on components with selected mRNA and DNA methylation features and respective performance. A: Separation of the individuals based on the first 2 components (orange dots represent the oligodendroglioma, green dots represent the astrocytoma and lilac dots represent the GBM individuals). B: AUC of the ROC curves to discriminate each class versus the 2 others. C: Circos plot representing the averaged expression of the genes in each glioma type and correlations between the features selected in each of the data blocks (mRNA features colored in turquoise, and DNA methylation features in yellow). D: Features (D1: mRNA; D2: DNA methylation) selected by frequency and absolute median value, categorized by the outcome glioma type (oligodendroglioma colored in orange, astrocytoma in green and GBM in lilac). The symbol at the end of each line indicates: if already reported in glioma studies (circle); if not reported in glioma but in other types of cancer (square); if never reported in cancer at all (triangle), based on search on Human MSigDB Collections, OMIM, and PubMed.

separate the 2 classes (the lines of expression of the second component features tend always to overlap). The features are highly correlated between blocks, where the DNA methylation features are mainly negatively correlated with the expression of the genes, while there are many positive correlations between miRNA and both genes and methylation features.

We have then selected only the variables that occurred more frequently in the first component with nonzero coefficient, given that the ones selected by the second component were not useful to distinguish the classes. The selected variables are shown in Figure 2C (mRNA features in Figure 2C1, DNA methylation features in Figure 2C2, and miRNA features in Figure 2C3), ordered by the respective importance in the components (median absolute value) and grouped by outcome. Note that the selected features are not exactly the ones already selected before when only using mRNA and DNA methylation data (shown in the previous section "Data description"). Indeed, the selection still includes some genes that were already chosen before: *LRRC41*, *SFRS4*, *GPBP1L1*, and *DRG2*, highlighting their relevance for the separation of the classes, while also adding the genes *PSMB2*, *NADK*, and *PPP1R8*, but now discarding *FBX042* and *CSDE1* as relevant genes. Also in the

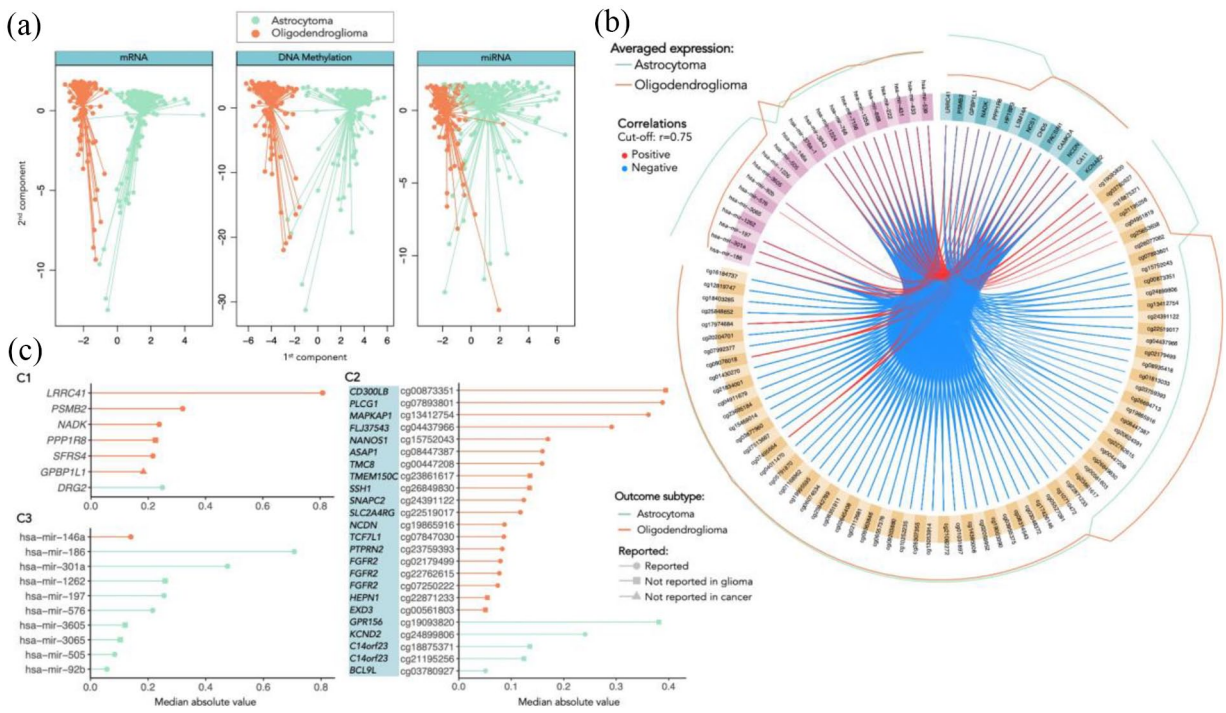


Figure 2. Separation and classification of LGG glioma types based on components with selected mRNA, DNA methylation, and miRNA features. (A): Separation of the individuals based on the first 2 components (orange dots represent the oligodendroglioma individuals, and green dots represent the astrocytoma individuals). (B): Circos plot representing the averaged expression of the genes in each glioma type (orange line for oligodendroglioma and green line for astrocytoma) and correlations between the features selected in each of the data blocks (mRNA features colored in turquoise, DNA methylation features in yellow and miRNA features in pink). (C): Features (C1: mRNA; C2: DNA methylation; C3: miRNA) selected by frequency and absolute median value, categorized by the outcome glioma type (oligodendroglioma colored in orange and astrocytoma in green). The symbol at the end of each line indicates: if already reported in glioma studies (circle); if not reported in glioma but in other types of cancer (square); if never reported in cancer at all (triangle), based on search on Human MSigDB Collections, OMIM, and PubMed.

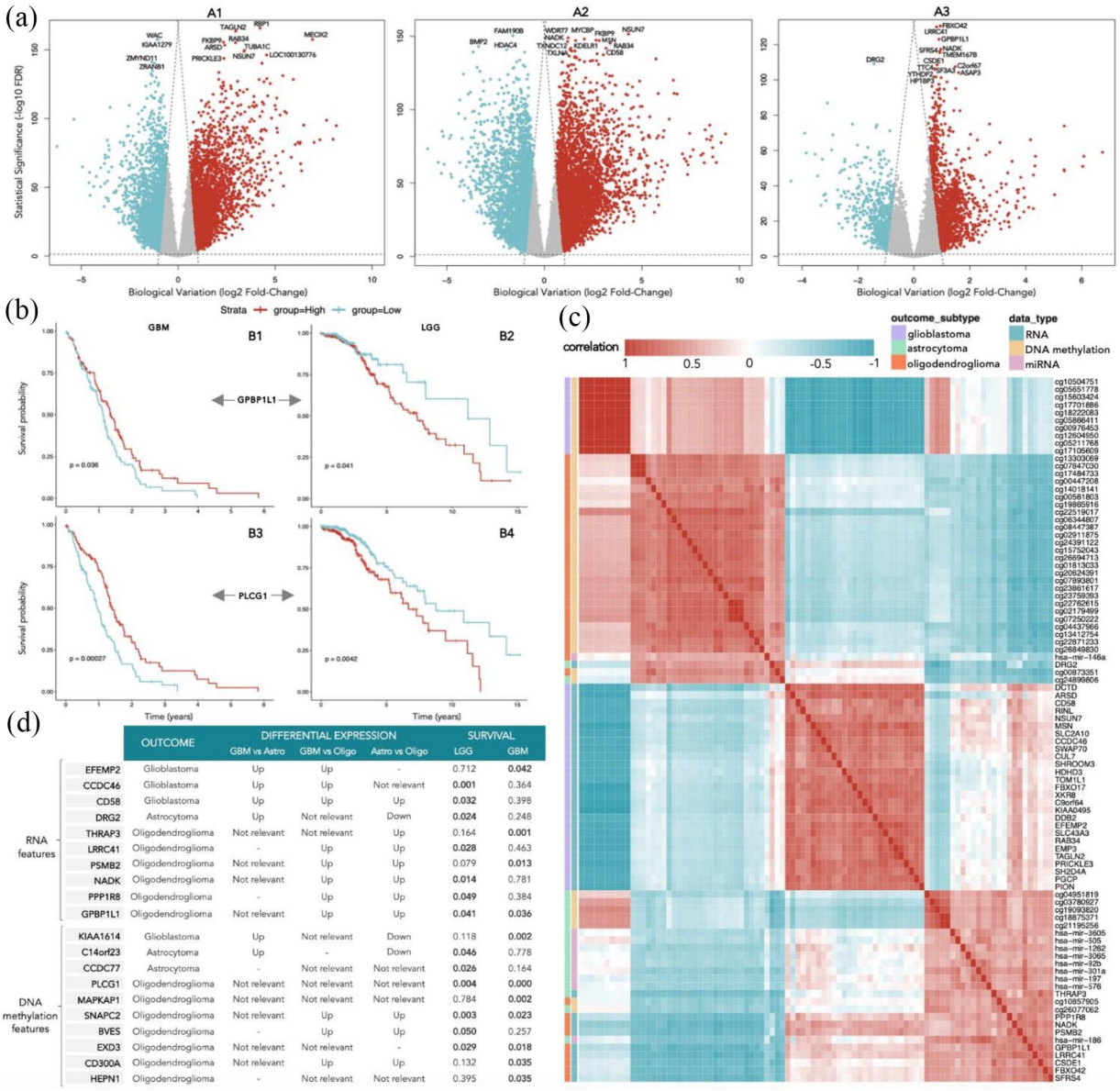
Table 3. Performance metrics for the classification of the LGG types using 2 components based on mRNA, DNA methylation, and miRNA features.

	PERFORMANCE METRICS			
GLIOMA TYPE	PRECISION	RECALL	F1	ACCURACY
Astrocytoma	0.996	0.958	0.977	0.973
Oligodendroglioma	0.943	0.995	0.968	

methylation sites, the more relevant methylated genes, such as *CD300LB*, *PLCG1*, *MAPKAP1*, *FLJ37543*, *NANOS1*, *ASAP1*, *TMC8*, and *SNAPC2* are maintained in the selection, while more differences are detected within the comparison of features with lower absolute value. Furthermore, most of the genes and methylation sites are associated with the distinction of the oligodendroglioma class, while most of the selected miRNA are more expressed in astrocytoma. It is also interesting to note that the gene *GPBP1L1* that was selected in both analyses (2 and 3 views) was never reported in any cancer study.

Investigation on the selected features

Differential gene expression, overrepresentation, and gene set enrichment analyses. Differential gene expression analysis was performed between each pair of glioma types to get a more deep insight about the differences in gene expression between the classes. It was also performed to compare the results with the genes selected by DIABLO and understand whether this method gives additional and more useful information than the one that could be directly obtained by just a single-omics analysis. Figure 3A shows the volcano plots of the DEGs between each pair of glioma types, where the top 15 genes (with the lowest FDR) are labeled. The number of statistical significant DEGs identified between GBM versus astrocytoma, GBM versus oligodendroglioma, and astrocytoma versus oligodendroglioma were 14 084, 14 470 and 11 125, respectively. Given the similarity between the LGG types, the DEGs obtained for astrocytoma versus oligodendroglioma are not only in a lower number but also have a considerably lower log(*FC*) than the



		OUTCOME	DIFFERENTIAL EXPRESSION			SURVIVAL	
			GBM vs Astro	GBM vs Oligo	Astro vs Oligo	LGG	GBM
RNA features	EFEMP2	Glioblastoma	Up	Up	-	0.712	**0.042**
	CCDC46	Glioblastoma	Up	Up	Not relevant	**0.001**	0.364
	CD58	Glioblastoma	Up	Up	Up	**0.032**	0.398
	DRG2	Astrocytoma	Up	Not relevant	Down	**0.024**	0.248
	THRAP3	Oligodendroglioma	Not relevant	Not relevant	Up	0.164	**0.001**
	LRRC41	Oligodendroglioma	-	Up	Up	**0.028**	0.463
	PSMB2	Oligodendroglioma	Not relevant	Up	Up	0.079	**0.013**
	NADK	Oligodendroglioma	Not relevant	Up	Up	**0.014**	0.781
	PPP1R8	Oligodendroglioma	-	Up	Up	**0.049**	0.384
	GPBP1L1	Oligodendroglioma	Not relevant	Up	Up	**0.041**	**0.036**
DNA methylation features	KIAA1614	Glioblastoma	Up	Not relevant	Down	0.118	**0.002**
	C14orf23	Astrocytoma	Up	-	Down	**0.046**	0.778
	CCDC77	Astrocytoma	-	Not relevant	Not relevant	**0.026**	0.164
	PLCG1	Oligodendroglioma	Not relevant	Not relevant	Not relevant	**0.004**	**0.000**
	MAPKAP1	Oligodendroglioma	Not relevant	Not relevant	Not relevant	0.784	**0.002**
	SNAPC2	Oligodendroglioma	Not relevant	Up	Up	**0.003**	**0.023**
	BVES	Oligodendroglioma	-	Up	Up	**0.050**	0.257
	EXD3	Oligodendroglioma	Not relevant	Not relevant	-	**0.029**	**0.018**
	CD300A	Oligodendroglioma	Not relevant	Up	Up	0.132	**0.035**
	HEPN1	Oligodendroglioma	-	Not relevant	Not relevant	0.395	**0.035**

Figure 3. Differential gene expression, correlation, and survival analysis with the selected features. (A): Volcano plots showing the differential expressed genes, where the top 15 genes (with lowest FDR) are labeled, between: A1: GBM and astrocytoma; A2: GBM and oligodendroglioma; A3: Astrocytoma and oligodendroglioma. The red dots represent the positively expressed genes, while the blue dots represent the negatively expressed genes, in the type that is mentioned first relatively to the second. (B): Survival analysis of the gene *GPBP1L1* in GBM (B1) and LGG patients (B2) and of the methylated gene *PLCG1* in GBM (B3) and LGG patients (B4). (C): Correlation plot with all the selected features annotated with the respective outcome and belonging block of data. (D): Selected features with effect on survival probability (*p* of log-rank test) in either LGG or GBM patients (statistical significant results for a .05 significance level are highlighted in bold).

Table 4. AUC obtained for each component and each data layer for the classification of LGG types.

	AUC		
	mRNA	DNA METHYLATION	miRNA
Component 1	0.994	0.987	0.995
Component 2	0.994	0.931	0.887

ones comparing the other classes. As expected, all the mRNA features selected by DIABLO were included in at least one of the sets of DEGs. Nevertheless, it is interesting to verify that not all the top DEGs were selected by DIABLO as the most relevant features to discriminate between the classes.

Furthermore, overrepresentation and GSEA were performed to have a more comprehensive view of the pathways and biological processes that are distinguished between classes. Using GSEA, it was possible to identify 20 GO terms of biological processes significantly enriched (FDR < .05) for GBM in comparison to both astrocytoma and oligodendroglioma, while 10 between the latter ones. On the contrary, when performing ORA, around 1000 biological processes were found, in all the 3 cases, to be significantly enriched for the upregulated genes and less than 500 for the downregulated genes (for more

details, see Supplementary Material). The genes that are overexpressed in GBM when comparing with both LGG types are mainly associated with biological processes such as defense response, cell adhesion, ECM structure organization, and ECM-receptor interaction, while the genes that are underexpressed in GBM enrich for biological processes related with cell-cell signaling, synapse (including its organization, modulation of its transmission, and vesicle cycle), regulation of transport, especially monoatomic ion transmembrane transport, and regulation of membrane potential. The biological processes that are differentially upregulated in GBM when comparing only to astrocytoma are mainly cytokine-mediated signaling and leukocyte migration, while with oligodendroglioma include the regulation of immune system process, cell population and leukocyte proliferation, cell migration, motility and secretion and response to cytokine. When comparing the LGG types, the upregulated genes in astrocytoma enrich for biological processes such as cytokine production, hemopoiesis, regulation of immune and defense response, cell and leukocyte proliferation and migration, while the downregulated genes are associated with cell-cell signaling, homeostasis, membrane potential and transport, synapse, nervous system, and head development. In summary, differential gene expression analysis, ORA, and GSEA revealed multiple statistical differences between each pair of glioma types, in particular genes and biological processes involved in RTK signaling, ECM organization, ubiquitination, and DNA transcription, which will be further discussed in the section "Discussion."

Correlation analysis. In the previous subsections, the presence of several high correlations between features of different modalities was detected, when observing the 2 circos plots in Figures 1 and 2. Therefore, a correlation plot was further obtained with all the selected features, to understand not only the interblock but also the intrablock correlations, as shown in Figure 3. There is a block that clearly stands out, with 10 methylation sites (between cg10504751 and cg17105609 (Figure 3C)) that are located, respectively, in the genes *GNAO1*, *MRC2*, *ARNTL*, *MCF2L2*, *TMEM106A*, *FGFRL1*, *KCNB1*, *KIAA1614*, *FES*, and *FGFRL1*, positively correlated with each other and negatively correlated with the group of 27 genes (between *DCTD* and *PION* (Figure 3C)). All that features are associated with the distinction of GBM, and suggest involvement in common biological processes or cellular components. Indeed, some of these features are known to be involved in RTK signaling. Other smaller blocks with high positive correlations can be detected as well, such as the one composed by the methylation sites cg13303069, cg07847030, and cg17484733, all located in the gene *TCF7L1* and the one composed by the methylation sites cg22762615, cg02179499, and cg07250222, all located in the gene *FGFR2*. Both these blocks are negatively correlated with the group of genes *GPBP1L1*, *LRRC41*, *CSDE1*, *FBXO42*, and *SFRS4*, which are associated with the oligodendroglioma type. Overall, various relevant interblock and intrablock correlations were found within the selected features that are also associated with the pathways identified by the previous analysis (subsection "Differential gene expression, overrepresentation, and gene set enrichment analyses").

Survival analysis. In addition, to understand the possible prognosis value of each of the selected features, the Kaplan-Meier survival curve was then plotted using their gene expression levels. It was performed for GBM and LGG patients separately, given the remarkably different survival times from the TCGA cohorts: generally lower than 5 years for GBM but reaching 15 years in LGG patients, and to compare the different expression profiles in the 2 groups. The p values of the log-rank test used to assess the statistical significance of the survival differences are summarized in Figure 3D, listing only the features which have a significant effect on survival in at least one of the groups. Notably, there are 4 different genes whose expression has an impact on survival probability in both groups of patients: *GPBP1L1* (with $p_{LGG} = .041$ and $p_{GBM} = .036$), *PLCG1* (with $p_{LGG} = .004$ and $p_{GBM} < .001$), *SNAPC2* (with $p_{LGG} = .003$ and $p_{GBM} = .023$), and *EXD3* (with $p_{LGG} = .029$ and $p_{GBM} = .018$). Figure 3B exemplifies the Kaplan-Meier curves of *GPBP1L1*, selected within the mRNA features, and *PLCG1*, selected within the DNA methylation features (with p lower than .01 in both groups of patients). On one hand, the higher expression of these genes reflects a higher survival probability in GBM patients (left side), while on the other hand, it has the opposite effect in LGG patients, by negatively impacting their survival probability. It is also interesting to note that some genes have a very significant effect ($p < .01$) in the survival of one group of patients while its expression is helpful to distinguish another type. For instance, the expression of *CCDC46* gene impacts the survival of LGG patients ($p_{LGG} = .001$) but is useful to discriminate the GBM type, while in the case of *THRAP3* gene, we have the reverse situation ($p_{GBM} = .001$), distinguishing oligodendroglioma. To sum up, 20 features were found to have a statistically significant impact on patient survival probability and might be candidate biomarkers for prognosis outcome.

Discussion

Over the last years, the classification of glioma tumors has been subject of significant alterations as a result of the evolving understanding of its biology derived from extensive research into the molecular profiling of glioma types. The significant role of genetic markers is being increasingly highlighted in the determination of the final diagnosis, and the publicly accessible multiomics databases, such as TCGA, are particularly helpful in gaining new knowledge regarding glioma characterization. However, the diagnostic categories reported for the TCGA-LGG and -GBM projects data sets are still not entirely consistent with the recently published WHO-2021 guidelines.[22,23] In this study, we used the recently revised glioma reclassification and integrate multiple data modalities (including RNA-seq, DNA methylation, and miRNA) to gain more

understanding of the biological mechanisms underlying glioma heterogeneity and discover novel biomarkers that can help further characterize each type and therefore be helpful for diagnosis and prognosis of the patients.

Using DIABLO, we were able to find predictive models that classify the 3 glioma types with high accuracy and identify relevant features in their characterization. This method showed to be more valuable than just applying a single-omics analysis like differential gene expression, where the correlation between features of different modalities is not considered. This network of connections between data layers has thus allowed to extract not only the most DEGs between classes, but the ones that are important for the separation of the 3 types, considering also the importance of DNA methylation and miRNA features and the relation established between each other.

It is known that GBM tumors are characterized by a highly malignant profile with increased cell proliferation and aggressiveness relatively to the LGG types, presenting significantly lower survival times even after chemotherapy and radiation treatments.[40] In this study, 2 highly correlated groups of molecular features were found to distinguish GBM from LGG, with the expression of the methylated sites positively correlated with each other and, in turn, negatively correlated with the selected genes. Interestingly, all these genes were revealed to be upregulated in GBM when compared with both astrocytoma and oligodendroglioma, while the methylation sites showed a considerably lower average level of methylation. It is worth noting that *KIAA0495*, which was selected with the highest loading in the distinction of GBM, codes a long noncoding RNA located at the arm of chromosome 1p, the absence of which is characteristic in oligodendroglioma. Nevertheless, the same gene is also downregulated in astrocytoma, where 1p deletion events are not so prevalent. In fact, *KIAA0495* has been suggested to modulate cellular apoptosis by regulation of p53-dependent antiapoptotic genes and to promote brain glioma proliferation and invasion, being associated with worse patients' prognosis.[41] Other selected features including *XKR8, EMP3*, and *GNAO1* also showed to be involved in the apoptosis process, but only the role of *EMP3* was already elucidated in glioma. The overexpression of *EMP3* in GBM is likely associated with the lack of *EMP3* hypermethylation that is present in LGG types,[42] and its most thoroughly researched function is its regulation of RTK signaling. The phosphorylation of tyrosine residues in signaling proteins is catalyzed by activated RTKs, starting multistep signaling cascades which promote differentiation and proliferation. In particular, *EMP3* has been shown to foster the phosphorylation of the RTKs *EGFR* and ErbB2/HER2, as well as their downstream effectors (ERK, PI3K, and Akt). Moreover, as a result of the *EGFR* gene amplification typically present in IDH-wt GBM, *EGFR* is frequently overactivated in this type, which thus correlates with increased proliferation, apoptosis resistance, and migration of tumor cells.[43] A set of other selected features, including *MSN, EFEMP2 GNAO1, TOM1L1, FES, FGFRL1*, and *TMEM106A*, was also reported to participate in RTK signaling pathways. It is important to highlight the association of *CD58, CUL7, MRC2, SHROOM3* and the aforementioned *FES, MSN, EFEMP2*, and *GNAO1* in the ECM organization, and, in special, the role of the last 3 in the integrin pathway. The ECM act as a biomechanical scaffold and a biochemical regulator of tumor cell homeostasis and is a crucial part of the GBM tumor microenvironment (TME). The ability of GBM tumor cells to migrate by inducing changes in actin cytoskeleton dynamics and ECM remodeling is well documented.[44] Integrins, as cell adhesion transmembrane receptors, act as ECM-cytoskeletal linkers that transmit signals between cells and the environment. Indeed, tumors can take advantage of the integrin-facilitated biological communication to take part in every stage of cancer progression, including tumor initiation, proliferation, and invasiveness.[45] In particular, while the *MSN* gene encodes an Ezrin-radixin-moesin (ERM) family protein that connects the actin cytoskeleton to the plasma membrane, *EFEMP2* was reported to be associated with the expression of matrix metalloproteinases (MMPs),[46] another group of proteins crucial in tissue remodeling. Although never described in glioma studies, *EFEMP2*, and *GNAO1* are reported in pathway databases to be also involved with phospholipase-C (PLC) activity. Phospholipase-C is stimulated by the P2Y2 receptor when coupled to specific G proteins, to hydrolyze PIP2, which is involved in calcium response, modulates a variety of actin binding proteins and activates Rac1 and RhoA.[44] These are small GTP-binding proteins of the Rho family known to act as molecular switches to control actin cytoskeleton dynamics. Interestingly, *GNAO1* is also linked to G-protein signaling and calcium regulation, and other selected features *SWAP70* and *MCF2L2* are players in Rho GTPase cycle. Furthermore, the role of *GNAO1* in purinergic signaling should be further investigated because this pathway has been emerging as an important factor giving glioma cells invasive potential and resistance to adenosine triphosphate (ATP)-induced cell death.[47] In this context, it has been proposed that besides alterations in the extracellular nucleotide/nucleoside metabolism, the disruption of purinergic signaling creates an inflammatory microenvironment with increase in cytokine production (specifically in interleukin [IL]-1β, IL-6, and tumor necrosis factor [TNF]) and regulated platelet function.[47] Notably, several of the selected features are involved in either nucleotide metabolism or transport (*DCTD* and *SLC43A3*), in platelet function (*TAGLN2*) or in cytokine production or regulation (*CD58, TMEM106A, FES, MSN*, and *GNAO1*). It is interesting to highlight that *TMEM106A*, a transmembrane protein found to be expressed on the surface of macrophages, specifically induces the release of IL-1β, IL-6, and TNF upon activation of the MAPK and nuclear factor (NF)-kappaB signaling pathways. This gene was never reported in glioma studies but showed to be a tumor suppressor in gastric, renal, and lung

cancer,[48] and therefore might also be also considered a novel marker candidate in glioma. Most of the remaining selected features are associated with metabolism, hemostasis, and transport. For instance, *ARSD* that encodes the protein arylsulfatase D, essentially involved in sphingolipid, estrogen, and protein metabolism, only gained attention recently for its role in amyloidosis.[49] Given that amyloid build-up is a part of the glioma tumor environment, and it was inclusively indicated as a potential target for developing a novel class of antitumor drugs,[50] more investigation should be dedicated to *ARSD* gene, whose function was never described in glioma.

Although all these genes were revealed to be relevant in biological processes underlying GBM and especially in its highly proliferative and invasive profile, the survival analysis showed that only the expression of the genes *CCDC46* (with updated symbol *CEP112*), *CD58*, *EFEMP2*, and *KIAA1614* were significant in patient survival probability, the first 2 in LGG and the last 2 in GBM types. In fact, even though these genes are upregulated in GBM patients, they might not directly influence their survivability and still have an impact on LGG. Indeed, *CD58* was already reported before to be a prognostic biomarker in LGG.[51] It is worth noting that besides the involvement in the centrosome cycle, the function of *CCDC46* and *KIAA1614* in glioma is still unknown and deserves future research.

In contrast to GBM, LGG tumors typically show a more indolent course. However, many might eventually transform into a more aggressive type.[52] Further characterization of the different omics profiles can possibly offer new insights on TME and development from LGGs to higher-grade gliomas. The current classification of LGG types is based mainly on 2 genetic markers, where the difference between astrocytoma and oligodendroglioma still relies only on the absence of 1p/19q codeletion.[22] In this study, multiple genetic features, including mRNA, DNA methylation, and miRNA, were found to discriminate between astrocytoma and oligodendroglioma. Most of the selected genes were differentially expressed between GBM and oligodendroglioma but not between GBM and astrocytoma, with the exception of the gene *DRG2*, where the opposite was verified. In general, the expression of those genes was downregulated in oligodendroglioma when comparing with either astrocytoma or GBM, while the expression of *DRG2* is upregulated in both oligodendroglioma and GBM when comparing with astrocytoma. Also, there are just a few selected sites where the level of methylation presents a considerable difference between astrocytoma and both the other 2 types. Therefore, astrocytoma seems to be the more difficult type to define and separate, with large similarities with GBM and oligodendroglioma, while these 2 are more easily to distinguish. Nonetheless, it is still possible to highlight some few exceptional features that were selected *DRG2, KCND2*, and *C14orf23*, which allow for the separation of astrocytoma. Indeed, *DRG2* was already reported before to be typically underexpressed in IDH-mutant samples, characteristic of LGG types, explaining its difference to GBM. This gene encodes a GTP-binding protein which catalyzes the conversion of GTP to GDP and is known to be involved in the regulation of cell growth and differentiation. Although its function in glioma is still not completely elucidated, it was recently demonstrated to be positively correlated with several steps in anti-tumor immune response,[53] and its depletion was shown to promote survival.[54] In the case of *KCND2*, it showed a lower level of methylation and a higher expression level in astrocytoma. It encodes a voltage-gated potassium channel that mediates transmembrane potassium transport in excitable membranes, primarily in the brain, regulating neurotransmitter release. Moreover, it also participates in signaling pathways such as ERK,[55] important for cellular proliferation, and GDNF[56] (glial cell line-derived neurotrophic factor), which guides glioma-associated microglia/macrophages (GAMs) recruiting in tumor immune resistance.[57]

Furthermore, *C14orf23*, also known as *LINC01551*, is an lncRNA that influences cell cycle and transcription. It was never described in glioma, but it was suggested to promote metastatic ability by posttranscriptional regulation of miRNA (by "sponge adsorption").[58] It would be interesting to further investigate the relationship between this gene and the miRNA selected as key players in glioma to understand its role in metastic progression.

From the selected features that distinguish oligodendroglioma, with downregulated gene expression and higher level of methylation compared to the other types, one should note that most are related either with ubiquitination processes (*LRRC41, FBXO42*, and *PSMB2*), RNA-binding activity, especially mRNA processing and splicing (*CSDE1, SFRS4, PPP1R8, GPBP1L1, THRAP3, SNAPC2*, and *NANOS1*) or signaling (*LRRC41, PSMB2, PPP1R8, DRG2, GPR156, KCND2, BCL9L, PLCG1, MAPKAP1*, and *FGFR2*). The signaling pathways influenced by these genes are mainly associated with RTK cascades, including the involvement of ERK, PI3K, Akt, EGFR, and Rho GTPases, that were already discussed previously to be associated with increased proliferation and migration of tumor cells, and thus with a more invasive glioma profile. In addition, other signaling pathways such as Notch and Wnt were also found to be essential in type distinction. For instance, *PSMB2* and *BCL9L* are known to be involved in the Wnt pathway, which activation was verified to increase the stemness of glioma cells.[59] Conversely, *FBXO42* not only participates in ubiquitination and degradation of p53/TP53, but it was also shown as a critical regulator of the Notch pathway via modulation of RBPJ-dependent global chromatin landscape changes in leukemia.[60] It would be of great importance to further research the function of *FBXO42* in glioma because it was only reported in other types of cancer. Interestingly, this gene is targeted by (hsa-)miR-186-5p, such as *LRRC41* and *CSDE1*. MicroRNAs (miRNAs) have

recently attracted interest and their potential impact on oncogenic processes has been thoroughly studied. Based on altered miRNA expression profiles, it has been possible to identify and diagnose various tumors, as well as to forecast their development, prognosis, and response to treatment. For example, the selected miR-186 was already described in multiple cancers, including glioma, with involvement in the regulation of inflammatory response and apoptosis.[61] Among the other selected miRNA, it is also highlighted miR-92b, which also targets the selected gene *CSDE1* and *GPBP1L1*. This miRNA was indicated to restrain the proliferation, invasion, and stimulate apoptosis of glioma cells by targeting PTEN/Akt signaling pathway, suggesting a possible antitumor effect in glioma treatment.[62]

Multiple genes revealed to have an important prognostic value, with significant effects on survival probability of LGG (including the genes *DRG2, LRRC41, NADK, PPP1R8*, and *C14orf23*) and GBM patients (including the genes *THRAP3, PSMB2*, and *MAPKAP1*). Notably, 4 genes showed to impact both LGG and GBM survival: *GPBP1L1, PLCG1, SNAPC2*, and *EXD3*. *GPBP1L1, SNAPC2*, and *EXD3* are likely key players in transcription, where the former is predicted to enable DNA and RNA binding and regulate transcription, the second encodes a subunit of the snRNA-activating protein complex, associated with the TATA box-binding protein, which is necessary for RNA polymerase II and III dependent small-nuclear RNA gene transcription, and the latter is involved in genetic stability and correction of DNA polymerase errors. While *SNAPC2* was already identified in glioma studies,[63] *EXD3* was only reported in gastric cancer[64] and *GPBP1L1* was never reported in any cancer study. *PLCG1*, on the contrary, plays an important role in the intracellular transduction of receptor-mediated tyrosine kinase activators, already discussed to be fundamental in actin reorganization and cell migration. It is however very intriguing the relation of the expression profiles of *GPBP1L1* and *PLCG1* with the survival probability observed between GBM and LGG patients, in Figure 3B, with low expression affecting survival in GBM and high expression affecting survival in LGG. This suggests that the role of these genes and/or their interactions might be different in the 2 types, yet impacting patient survival in both cases. In fact, the expression of *PLCG1* was demonstrated to be significantly correlated with IDH status,[65] which is one of the key characteristics that distinguish LGG and GBM.

Overall, the identified candidate biomarkers might have promising implications for patient care, not only by potentially providing a more accurate and tailored diagnosis, but also by optimizing prognostic assessments and guiding treatment strategies. In particular, the molecular features that revealed distinctive expression patterns across types (Figures 1D and 2C) can inform clinicians about the specific characteristics of each patient's glioma, allowing for a more personalized approach to treatment, for instance by influencing the selection of drugs that are known to be more effective according to the tumor profile. On top of that, the genes that exhibited significant prognostic value (Figure 3D), can offer insights into the course of the disease and enable clinicians to more accurately predict patient outcomes. Nevertheless, further experimental research and clinical validation are essential to translate these potential impacts into tangible benefits for patients in practical health care settings.

Conclusions

In this study, we used the recently updated glioma reclassification to further characterize the different glioma types: GBM, astrocytoma, and oligodendroglioma. Based on the integration of multiple omics' layers (mRNA, DNA methylation, and miRNA data from TCGA) using a supervised and sparse variant of CCA (DIABLO), we were able to discriminate the 3 glioma types with very high performance. Indeed, the correlation between blocks showed to be of extreme importance in the selection of the features. The group of correlated features that revealed to be relevant in the distinction of GBM from LGG types were mainly associated with RTK signaling and extracellular matrix organization, which is indeed a crucial part of the GBM TME. Concurrently, the discrimination of the LGG types was characterized mainly by features involved in ubiquitination and DNA transcription processes.

Furthermore, we were able to identify several novel features that deserve future attention. For instance, the methylation of the gene *KIAA1614* is a potential biomarker with both diagnosis and prognosis value in GBM, yet its function is still not elucidated in glioma. Moreover, the gene *GPBP1L1* has a great impact in both LGG and GBM patient survival and showed distinct downregulation in oligodendroglioma when compared to both astrocytoma and GBM, but its role in cancer is still not described. Nonetheless, this gene is known to be targeted by miR-92b, a glioma-reported miRNA which was also selected by our method as a relevant feature to distinguish LGG types. Therefore, the interaction between the selected genes and miRNA features might also be essential in further investigation of these novel biomarkers.

Acknowledgements

The results presented here are based on data generated by The Cancer Genome Atlas (TCGA) Research Network.

Author Contributions

Francisca G. Vieira: Conceptualization, Methodology, Data analysis, Writing—original draft. Regina Bispo: Conceptualization, Methodology, Writing—review and editing, Supervision. Marta B. Lopes: Conceptualization, Methodology, Writing—review and editing, Supervision, Funding acquisition.

SUPPLEMENTAL MATERIAL

Supplemental material for this article is available online.

REFERENCES

1. Nicholson JG, Fine HA. Diffuse glioma heterogeneity and its therapeutic implications. *Cancer Discov*. 2021;11: 575-590.
2. Wiestler B, Capper D, Holland-Letz T, et al. ATRX loss refines the classification of anaplastic gliomas and identifies a subgroup of IDH mutant astrocytic tumors with better prognosis. *Acta Neuropathol*. 2013;126:443-451.
3. Li L, Wei Y, Shi G, et al. Multi-omics data integration for subtype identification of Chinese lower-grade gliomas: a joint similarity network fusion approach. *Comput Struct Biotechnol J*. 2022;20:3482-3492.
4. Sienkiewicz K, Chen J, Chatrath A, et al. Detecting molecular subtypes from multi-omics datasets using SUMO. *Cell Reports Methods*. 2022;2:100152.
5. International Agency for Research on Cancer. *WHO Classification of Tumours of the Central Nervous System*. IARC5 ed. World Health Organization; 2022.
6. Qi L, Wang W, Wu T, Zhu L, He L, Wang X. Multi-omics data fusion for cancer molecular subtyping using sparse canonical correlation analysis. *Front Genet*. 2021;12:607817.
7. González I, Déjean S, Martin PG, Baccini A. CCA: an R package to extend canonical correlation analysis. *J Stat Softw*. 2008;23:1-14.
8. Tenenhaus A, Philippe C, Guillemot V, Le Cao KA, Grill J, Frouin V. Variable selection for generalized canonical correlation analysis. *Biostatistics*. 2014;15: 569-583.
9. Li W, Zhang S, Liu CC, Zhou XJ. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*. 2012;28:2458-2466.
10. Zeng Z, Vo AH, Mao C, Clare SE, Khan SA, Luo Y. Cancer classification and pathway discovery using non-negative matrix factorization. *J Biomed Inform*. 2019;96:103247.
11. Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann Appl Stat*. 2013;7:523-542.
12. Palzer EF, Wendt C, Bowler R, Hersh CP, Safo SE, Lock EF. SJIVE: supervised joint and individual variation explained. *arXiv e-prints*. 2021. https://arxiv.org/abs/2102.13278
13. Argelaguet R, Velten B, Arnol D, et al. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol*. 2018;14:e8124.
14. Argelaguet R, Arnol D, Bredikhin D, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol*. 2020;21:111.
15. Singh A, Shannon CP, Gautier B, et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*. 2019;35:3055-3062.
16. Zhang Y, Gaynanova I. Joint association and classification analysis of multi-view data. *Biometrics*. 2022;78:1614-1625.
17. Safo SE, Min EJ, Haine L. Sparse linear discriminant analysis for multiview structured data. *Biometrics*. 2022;78:612-623.
18. Jin X, Liu L, Wu J, et al. A multi-omics study delineates new molecular features and therapeutic targets for esophageal squamous cell carcinoma. *Clin Transl Med*. 2021;11:e538.
19. Zheng P, Sun S, Wang J, et al. Integrative omics analysis identifies biomarkers of idiopathic pulmonary fibrosis. *Cell Mol Life Sci*. 2022;79:66.
20. Cai Z, Poulos RC, Liu J, Zhong Q. Machine learning for multi-omics data integration in cancer. *iScience*. 2022;25:103798.
21. Mo Q, Shen R, Guo C, Vannucci M, Chan KS, Hilsenbeck SG. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics*. 2018;19:71-86.
22. Louis DN, Perry A, Wesseling P, et al. The 2021 WHO classification of tumors of the central nervous system: a summary. *Neuro-oncology*. 2021;23:1231-1251.
23. Mendonça ML, Coletti R, Gonçalves CS, et al. Updating TCGA glioma classification through integration of molecular profiling data following the 2016 and 2021 WHO guidelines. *bioRxiv*. 2023.
24. Samur MK. RTCGAToolbox: a new tool for exporting TCGA firehose data. *PLoS ONE*. 2014;9:e106397.
25. Colaprico A, Silva TC, Olsen C, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*. 2016;44:e71.
26. Rohart F, Gautier B, Singh A, Lê Cao KA. MixOmics: an R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol*. 2017;13:e1005752.
27. Hotelling H. The most predictable criterion. *J Educ Psychol*. 1935;26:139-142.
28. Ren X, Kuan PF. methylGSA: a Bioconductor package and Shiny app for DNA methylation data length bias adjustment in gene set testing. *Bioinformatics*. 2019;35:1958-1959.
29. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database hallmark gene set collection. *Cell Syst*. 2015;1:417-425.
30. Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res*. 2019;47:D1038-D1043.
31. Huang HY, Lin YCD, Cui S, et al. MiRTarBase update 2022: an informative resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res*. 2022;50:D222-D230.
32. Chen Y, Lun AT, Smyth GK. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Res*. 2016;5:1438.
33. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. 2010;11:R25.
34. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc Series B (Method)*. 1995;57:289-300.
35. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res*. 2019;47:W199-W205.
36. Harrell FE Jr. *Hmisc: Harrell Miscellaneous*. R package version 5.0-1; 2023.
37. Kolde R. *pheatmap: Pretty Heatmaps*. R package version 1.0.12; 2019.
38. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. New York: Springer; 2000.
39. Kassambara A, Kosinski M, Biecek P. *survminer: Drawing Survival Curves using 'ggplot2.'* R package version 0.4.9; 2021.
40. Smith HL, Wadhwani N, Horbinski C. Major features of the 2021 WHO classification of CNS tumors. *Neurotherapeutics*. 2022;19:1691-1704.
41. Xiao S, Wang R, Wu X, Liu W, Ma S. The long noncoding RNA TP73-AS1 interacted with miR-124 to modulate glioma growth by targeting inhibitor of apoptosis-stimulating protein of p53. *DNA Cell Biol*. 2018;37:117-125.
42. Kunitz A, Wolter M, van den Boom J, et al. DNA hypermethylation and aberrant expression of the EMP3 gene at 19q13.3 in human gliomas. *Brain Pathol*. 2007;17:363-370.
43. Martija AA, Pusch S. The multifunctional role of EMP3 in the regulation of membrane receptors associated with IDH-wild- type glioblastoma. *Int J Mol Sci*. 2021;22:5261.
44. Kłopocka W, Korczyński J, Pomorski P. Cytoskeleton and nucleotide signaling in glioma C6 cells. *Adv Exp Med Biol* 2020;1202:109-128.
45. Pang X, He X, Qiu Z, et al. Targeting integrin pathways: mechanisms and advances in therapy. *Signal Transduct Target Ther*. 2023;8:1.
46. Wang L, Chen Q, Chen Z, et al. EFEMP2 is upregulated in gliomas and promotes glioma cell proliferation and invasion. *Int J Clin Exp Pathol*. 2015;8:10385-10393.
47. Braganhol E, Wink MR, Lenz G, Battastini AMO. Purinergic signaling in glioma progression. *Adv Exp Med Biol*. 2020;1202:87-108.
48. Liu J, Zhu H. TMEM106A inhibits cell proliferation, migration, and induces apoptosis of lung cancer cells. *J Cell Biochem*. 2019;120:7825-7833.
49. Lin Y, Fan L, Zhang R, Pan H, Li Y. ARSD is responsible for carcinoma and amyloidosis of breast epithelial cells. *Eur J Cell Biol*. 2022;101:151199.
50. Zayas-Santiago A, Díaz-García A, Nuñez-Rodríguez R, Inyushin M. Accumulation of amyloid beta in human glioblastomas. *Clin Exp Immunol*. 2020;202:325-334.
51. Sato K, Tahata K, Akimoto K. Five genes associated with survival in patients with lower-grade gliomas were identified by information-theoretical analysis. *Anticancer Res*. 2020;40:2777-2785.
52. Binder H, Willscher E, Loeffler-Wirth H, et al. DNA methylation, transcriptome and genetic copy number signatures of diffuse cerebral WHO grade II/III gliomas resolve cancer heterogeneity and development. *Acta Neuropathologica Communications*. 2019;7:59.
53. Chen R, Wu W, Liu T, et al. Large-scale bulk RNA-seq analysis defines immune evasion mechanism related to mast cell in gliomas. *Front Immunol*. 2022;13:914001.
54. Pappula AL, Rasheed S, Mirzaei G, Petreaca RC, Bouley RA. A genome-wide profiling of glioma patients with an idh1 mutation using the catalogue of somatic mutations in cancer database. *Cancers*. 2021;13:4299.
55. Adams JP, Anderson AE, Varga AW, et al. The A-type potassium channel Kv4.2 is a substrate for the mitogen-activated protein kinase ERK. *J Neurochem*. 2000;75:2277-2287.
56. Shinoda M, Fukuoka T, Takeda M, Iwata K, Noguchi K. Spinal glial cell line-derived neurotrophic factor infusion reverses reduction of Kv4.1-mediated A-type potassium currents of injured myelinated primary afferent neurons in a neuropathic pain model. *Mol Pain*. 2019;15:1744806919841196.
57. Catalano M, D'Alessandro G, Trettel F, Limatola C. Role of infiltrating microglia/macrophages in glioma. *Adv Exp Med Biol*. 2020;1202:281-298.
58. Xue MY, Cao HX. LINC01551 promotes metastasis of nasopharyngeal carcinoma through targeting microRNA-132-5p. *Eur Rev Med Pharmacol Sci*. 2020;24:3724-3733.

59. Tao B, Song Y, Wu Y, et al. Matrix stiffness promotes glioma cell stemness by activating BCL9L/Wnt/β-catenin signaling. *Aging*. 2021;13:5284-5296.
60. Jiang H, Bian W, Sui Y, et al. FBXO42 facilitates Notch signaling activation and global chromatin relaxation by promoting K63-linked polyubiquitination of RBPJ. *Sci Adv*. 2022;8:eabq4831.
61. Li Q, Wu M, Fang G, et al. MicroRNA-186-5p downregulation inhibits osteoarthritis development by targeting MAPK1. *Mol Med Rep*. 2021;23:253.
62. Song H, Zhang Y, Liu N, et al. MiR-92b regulates glioma cells proliferation, migration, invasion, and apoptosis via PTEN/Akt signaling pathway. *J Physiol Biochem*. 2016;72:201-211.
63. Ma J, Hou X, Li M, et al. Genome-wide methylation profiling reveals new biomarkers for prognosis prediction of glioblas- toma. *J Cancer Res Ther*. 2015;11:C212-C215.
64. Sun D, Liu M, Huang F. Bioinformatics analysis of expression and function of EXD3 gene in gastric cancer. *Nan Fang Yi Ke Da Xue Xue Bao = J Southern Med Univ*. 2019;39:215-221.
65. Li T, Yang Z, Li H, et al. Phospholipase Cç1 (PLCG1) overexpression is associated with tumor growth and poor survival in IDH wild-type lower-grade gliomas in adult patients. *Lab Invest*. 2022;102:143-153.