BioData Mining

RESEARCH

Open Access

Disclosing transcriptomics network-based signatures of glioma heterogeneity using sparse methods

Sofia Martins[1†], Roberta Coletti[2*†] and Marta B. Lopes[1,2,3,4*]

†Sofia Martins and Roberta Coletti contributed equally to this work.

*Correspondence: roberta.coletti@fct.unl.pt; marta.lopes@fct.unl.pt

[1] NOVA School of Science and Technology, NOVA University of Lisbon, Caparica 2829-516, Portugal
[2] Center for Mathematics and Applications (NOVA Math), NOVA School of Science and Technology, Caparica 2829-516, Portugal
[3] NOVA Laboratory for Computer Science and Informatics (NOVA LINCS), NOVA School of Science and Technology, Caparica 2829-516, Portugal
[4] UNIDEMI, Department of Mechanical and Industrial Engineering, NOVA School of Science and Technology, Caparica 2829-516, Portugal

Abstract

Gliomas are primary malignant brain tumors with poor survival and high resistance to available treatments. Improving the molecular understanding of glioma and disclosing novel biomarkers of tumor development and progression could help to find novel targeted therapies for this type of cancer. Public databases such as The Cancer Genome Atlas (TCGA) provide an invaluable source of molecular information on cancer tissues. Machine learning tools show promise in dealing with the high dimension of omics data and extracting relevant information from it. In this work, network inference and clustering methods, namely Joint Graphical lasso and Robust Sparse K-means Clustering, were applied to RNA-sequencing data from TCGA glioma patients to identify shared and distinct gene networks among different types of glioma (glioblastoma, astrocytoma, and oligodendroglioma) and disclose new patient groups and the relevant genes behind groups' separation. The results obtained suggest that astrocytoma and oligodendroglioma have more similarities compared with glioblastoma, highlighting the molecular differences between glioblastoma and the others glioma subtypes. After a comprehensive literature search on the relevant genes pointed our from our analysis, we identified potential candidates for biomarkers of glioma. Further molecular validation of these genes is encouraged to understand their potential role in diagnosis and in the design of novel therapies.

Keywords: Glioma, Transcriptomics, Biomarkers, Sparse networks, Joint graphical lasso, Robust sparse K-means clustering

Introduction

Gliomas are primary malignant brain tumors, accounting for 28% of all brain tumors and 80% of malignant ones [1]. The large heterogeneity characterizing glioma, at cellular and molecular levels, leads to distinct cancer types with different prognosis, among which glioblastoma (GBM) is the most aggressive one, with a median survival time of about 15 months [2]. Following the advances in molecular and cell technologies, relevant molecular information has been generated through transcriptomics and other 'omics profiling, enabling the definition of novel tumor classification and treatments [1]. Moreover, clinical-specific molecular biomarkers, namely, age and sex, have been pointed out by

bioinformatic analysis disclosing several differences among clinical groups, e.g. at DNA methylation and gene expression levels [3–7]. However, a deeper molecular characterization of this cancer is necessary for an increased understanding of this type of tumors and the development of an effective personalized medicine.

The World Health Organization (WHO) classification of the Central Nervous System (CNS) tumors has been changing throughout the years regarding the classification of glioma. The classification of gliomas has evolved from histological-based [8] to increasingly based on molecular alterations combined to histological features [9, 10]. Glioma subtypes were divided into four main types: glioblastoma (GBM), astrocytoma, oligodendroglioma, and oligoastrocytoma, the latter presenting mixed histological nature, between astrocytoma and oligodendroglioma [8]. In 2016, molecular features were introduced to better define the glioma subtypes [9]. As a consequence, oligoastrocytoma diagnosis became strongly discouraged, and most of those cases could be reclassified as astrocytoma or oligodendroglioma. An updated version of the WHO CNS classification was introduced in 2021 [10], where oligoastrocytomas are no longer considered, and the glioma subtypes are classified mainly based on the sample's molecular profiles.

The technological advances are responsible for the generation of large amounts of genetic information. With an increasing larger number of molecular features than samples, machine learning stands as a powerful tool to retrieve relevant information from these complex datasets. A relevant machine learning task to cope with 'omics data involves dimensionality reduction using, e.g., feature selection methods. Model regularization is a promising way to select features and improve model interpretability by adding constraints to the solutions. Methods like the least absolute shrinkage and selection operator (lasso) [11] and further versions (e.g., fused LASSO [12], adaptive LASSO [13], and group LASSO [14]) and the elastic net [15] are examples of regularization methods. The lasso method was proposed for estimation in linear models by imposing a L_1 -penalty on the cost function. Instead of focusing on subsets, this method defines a continuous shrinking operation capable of producing coefficients exactly equal to zero [11], making the solution sparse, and consequently, reducing the number of variables. The variables selected this way can be regarded as biomarker candidates for both diagnostics and therapeutic purposes [16].

Biological networks are one of the most studied types of networks used to decipher the molecular structures involved in disease development and progression [17]. Biological networks can be represented by graphs, where nodes are the biological entities and the edges the connections between them. Several studies acknowledge the role of networks in the study of cancer, beyond the selection of individual, potentially unrelated, molecular features. Several approaches have been proposed accounting for existing knowledge of gene interactions into classification models to identify network biomarkers and this way better understand the molecular mechanism behind cancer outcomes (e.g., [2, 17–20]).

Graphical lasso (glasso) [21] is a method for estimating undirected sparse graphs, enabling the identification of relevant subnetworks while discarding irrelevant links between the nodes, therefore inducing data dimension reduction. This method has shown successful disclosing gene interactions in 15 types of cancer [22], based on RNA-sequencing (RNA-seq) data. An extension of glasso, the joint graphical lasso

(JGL), jointly estimates graphical models with observations from different classes [23], inducing sparsity and forcing similarity between these classes. JGL relies on the fact that it is expected that different but related classes (such as two different tumor subtypes) share some similarities. JGL determines the joint estimation of separated models, allowing for the exploration of similarities between multiple classes, maintaining the distinctive traits of each. This method stands promising for tackling cancer tumor heterogeneity in glioma, allowing for understanding how genes interact among different cancer glioma types.

Besides the relevance of identifying key network features in known cancer subtypes, the discovery of new patient groups from the increasingly available omics data is also encouraged. This approach aims to evaluate whether the information extracted from available cancer-type-specific molecular data correlates with the established clinical/molecular groups, or supports further cancer reclassification efforts. Finding groups of patients based on molecular information can be tackled by clustering methods. In the context of high-dimensional data, a shape-based method for sparse clustering, SPARCL, was proposed to allow clustering-based feature selection. The motivation behind this technique is to divide observations into a pre-specified number of groups, using only a subset of representative features. Robust K-means clustering (RSKC) is an extension of SPARCL that can handle outliers in the data. A pre-specified proportion of outliers is admitted by the method, which considers that the observations that are not so close to each cluster centers are not representative of any condition [24]. The ability to identify outliers is of high relevance in the context of gliomas, for which the classification has been evolving with successive alterations as far as more molecular information is considered in the definition of glioma subtypes.

The goal of this work is to identify, through sparse methods, transcriptomic network biomarkers of heterogeneity in gliomas. Although differences in glioma types have been investigated in the past [25–28], these studies are based on datasets collected or revised up to 2007. The successive release of updated glioma classification guidelines determined the change in some patients' diagnoses, which could impact the glioma-type characterization. In this light, we are providing the first study exploring the consequence of the 2016 WHO glioma classification in biomarker discovery, aimed at either supporting this glioma classification or disclosing new patient groups.

In a first stage of the network-based methodology proposed, RNA-seq data from glioma patients were obtained from The Cancer Genome Atlas (TCGA), the largest repository of multiple omics data concerning cancer in humans [29]. The TCGA datasets were updated according to more recent glioma classifications. Through the application of JGL, we investigated network differences and similarities between glioma types. In the second stage of the analysis pipeline, the genes involved in the glioma-type specific networks inferred by JGL were used as input in the RSKC algorithm, to assess their ability to group patients into the known glioma subtypes. The potential to disclose new patient groups was also investigated. The gene networks inferred from each glioma subtype, and their most representative features, will potentially lead to a more comprehensive understanding of the molecular landscape of glioma, providing valuable insights to the definition of improved diagnosis, novel therapeutic targets, and ultimately contributing to patient life quality.

Materials and methods

Data

The RNA-seq data used were collected by the TCGA Research Network. We downloaded TCGA-GBM [30, 31] and TCGA-LGG [32] projects, which group the glioma patients according to 2007 WHO classification [33] into the two classes of GBM and Lower Grade Glioma (LGG), the latter comprising all astrocytoma, oligodendroglioma and oligoastrocytoma samples. To avoid considering obsolete glioma types, such as oligoastrocytoma, we updated the dataset to the 2016 WHO classification, by following the procedure explained in Mendonça et al. [34]. In practice, oligoastrocytoma samples were mainly reassigned to astrocytoma or oligodendroglioma, depending on the status of the IDH gene family and 1p/19q codeletion.

The final dataset comprises of 622 patients, divided as 264 astrocytoma, 220 oligodendroglioma and 138 GBM. As required from JGL, only normally distributed variables were considered in our dataset. To ensure this assumption, we selected the features having normal distribution in accordance with the Jarque-Bera test [35], leading to a total of 16338 genes. The LGG and GBM datasets were extracted using the `GDCquery` R function from **TCGAbiolinks** package [36, 37]. The TCGA datasets were already normalized with Transcripts Per Million (TPM) and upper quantile normalization. For the JGL method, nonparanormal normalization with `huge.npn` function from **huge** R package [38] was applied to lead the variables normally distributed [39]. For the RSKC method, z-score [40] was used.

Joint graphical lasso

Let X be the data matrix, $n \times p$, where n and p denote the number of observations and the number of features, respectively. If all the features are independent and identically distributed, glasso algorithm estimates the precision matrix Θ, which is the inverse covariance matrix. 0s in Θ correspond to pairs of features conditionally independent from each other, considering all other variables. These conditional relationships between variables correspond to an undirected graph, where nodes denote features and edges the relationships between pairs of features [21]. Let us consider D distinct but related datasets $Y^{(1)}, ..., Y^{(D)}$, $D \geq 2$. $Y^{(d)}$ is an $n_d \times p$ matrix, where all p features are common to all the D classes. Features are independent and identically distributed within each class. The JGL [23] algorithm estimates the vector $\widehat{\Theta} = \left(\widehat{\Theta}^{(1)}, ..., \widehat{\Theta}^{(D)}\right)$, containing D precision matrices, each one defining the undirected network existing in the corresponding dataset. These networks are estimated jointly, by inducing sparsity and similarity across the different dataset through a penalty function. The authors of JGL method proposed two different penalty functions, the Group Graphical Lasso (GGL), leading to similar pattern of sparsity, and Fused Graphical Lasso (FGL), encouraging similarity between the edges. In this work, we selected FGL, which, beside providing generally better performances in various applications [23], was more suitable for the purposes of our work. FGL regulates the sparsity through the parameter λ_1 (which induces the same degree of sparsity in all the datasets), and encourages similarity among the D datasets by the parameter λ_2.

Sparse clustering

K-means clustering divides the n observations into K clusters $C_1, ..., C_K$, by minimizing the average distance between the observation constituting each cluster. Further extensions of this method defined outlier-robust clusters, which is able to identify outliers by considering an additional parameter α. This pre-defined value represents the percentage of observations that are excluded from a given cluster in each step, as the one having the larger distance from the cluster center (outliers) [41]. This algorithm has been recently modified in order to introduce sparsity in the robust K-means clustering process, by means of a lasso-type penalty, regulated by the parameter $L1$ (the lower the parameter value, the more sparsity will be induced) [42].

Clustering validation

A mathematical validation with random baseline of the RSKC results has been performed. For each case of study, we generated 1000 random datasets with the same dimension, and we applied RSKC, by fixing $L1 = 2$ and testing $K \in \{2,\ 3\}$. For each random dataset, the corresponding unsupervised clustering was evaluated by computing *silhouette* and *Calinski-Harabasz* scores.

Analysis workflow

The analysis started by data preprocessing (dataset update and normalization), and visualization. The latter has been performed by using the Uniform Manifold Approximation and Projection (UMAP)[43], a nonlinear dimensionality reduction technique for data representation, widely used in multi-omics studies for sample visualization [44–46].

Three case studies were created to allow the comparison of different glioma subtypes, namely, 'LGG *vs.* GBM' (case A), 'Astrocytoma *vs.* Oligodendroglioma' (case B), and 'Astrocytoma *vs.* Oligodendroglioma *vs.* GBM' (case C).

The following step consisted of applying JGL to all three cases. JGL was applied using the R package `JGL` [23]. For cases A and B, we are assuming the existence of $D = 2$ distinct datasets, while in case C the number of datasets is $D = 3$. To detect the optimal choice for the JGL parameters, we tested several combinations, based on biological and practical considerations, as suggested by the JGL authors [23]. Given the high dimentionality of the considered starting datasets (16338 variables), a great level of sparsity was desired (high values of λ_1). Conversely, since the aim of this study was to highlight differences between glioma subtypes, we decided to not force similarity (low values of λ_2). Based on these assumptions, we tested $\lambda_1 \in \{0.90, 0.95, 0.97\}$ and $\lambda_2 \in \{0.001, 0.005, 0.01\}$. By comparing the results obtained for all the combinations of the tuning parameters, we detected $\lambda_1 = 0.95$ and $\lambda_2 = 0.01$ as the most suitable combination to discuss the related estimated network, due to the reasonable number of selected variables (easy to discuss but large enough to be biologically meaningful). In the Results and discussion section we focus on this outcome, by performing clustering based on the corresponding variable selection. Indeed, as a way to validate the biological meaning of the inferred networks, RSKC was chosen to evaluate whether distinct patient groups would be obtained based on the features selected by JGL, in an unsupervised way. The corresponding outcome should disclose if the selected features were able

to separate either known glioma subtypes or new patient groups. RSKC was performed using the R package **RSKC** [42], by testing different combinations of parameters. Specifically, the percentage of outliers has been defined as $\alpha = 0.1$, as suggested by the RSKC developers, while the regularization parameter $L1$ has been considered in $\{2, 24\}$, in order to compare clustering results in a setting of strong or weak variable selection (conversely from JGL approach, low values of $L1$ lead to stronger regularization). To set the number of expected clusters K, we observed that the defined cases of study contain two glioma types in case A and B, and 3 glioma types in case C. However, in case C, astrocytoma and oligodendroglioma cases could be also considered as unique class (LGG), so we decided to test $K \in \{2, 3\}$. For each parameter combination, the quality of the clusters obtained was evaluated using the simplified *silhouette* score [47] and the *Calinski-Harabasz* index [48], widely used clustering validity indices in omics studies and top performing indices across several real datasets [45, 49–52]. In the Results and discussion section, we will discuss only the outcomes obtained with the combination leading to the best scores, but the complete analysis is reported in Supplementary Table S4. Clustering validation with a random baseline has been also performed to assess the reliability of our results. For more details we refer to Supplementary Material.

The datasets and R code used for this study will be made available upon request.

Results and discussion

A first visual inspection of glioma transcriptomics data was performed in a 2-dimensional space obtained via UMAP, in order to capture potential preexisting patterns. In this representation (Fig. 1), we can distinguish 2 well-separated groups, with LGG types (astrocytoma and oligodendroglioma) appearing closer compared to GBM cases. This outcome is in agreement with previous literature reports, which highlights some similarities among LGG types which affect tumor evolution and lead to better overall survival compared to GBM patients [53–55]. This preliminary result supports

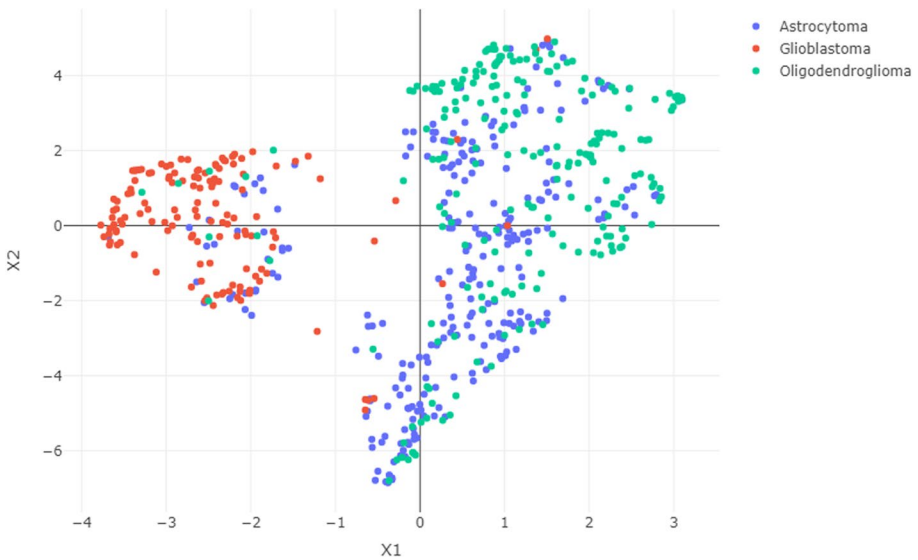

Fig. 1 UMAP representation of the transcriptomics dataset. Labels are assigned based on the 2016 WHO glioma classification guidelines

the need for further disclosing the molecular similarities and uniqueness governing glioma development and progression.

Network inference

Figure 2 illustrates a single network for each case study obtained through JGL with $\lambda_1 = 0.95$ and $\lambda_2 = 0.01$. Each variable is represented by a node (gene), while edges represent relations between nodes. The common edges, as well as the edges exclusive to each class, are highlighted with different colors.

For the 'LGG *vs.* GBM' case (Fig. 2A) the algorithm selects a total of 43 variables. There are some common edges ($n = 14$), but most of the estimated connections are exclusive for LGG ($n = 14$) and GBM ($n = 6$). A large subnetwork can be observed on the right hand side, containing 10 genes and both exclusive (green and blue) and shared connections between LGG and GBM (yellow). The genes included in this subnetwork are *TMEM125*, *ERMN*, *GJB1*, *CARNS1*, *KLK6*, *MAG*, *MOG*, *MBP*, *MOBP* and *CNDP1*. Half of them are involved in relations which are common to the two classes, while *CNDP1* is exclusive to GBM, and *TMEM125*, *ERMN*, *GJB1*, and *CARNS1* are nodes in the LGG network.

The network obtained for the 'Astrocytoma *vs.* Oligodendroglioma' case (Fig. 2B) is composed by 61 variables and shows many common edges between the two datasets (yellow, $n = 65$). There are only a few edges exclusive to one of the two subtypes, namely $n = 9$ for astrocytoma (green) and $n = 1$ for oligodendroglioma (blue). The network highlights the presence of a large subnetwork on the right hand side, represented by genes *BUB1B*, *CENPF*, *TPX2*, *AURKB*, *BIRC5*, *BUB1*, *CKAP2L*, *FAM64A*, *GTSE1*, *HJURP*, *KIFC1*, *MKI67*, *NCAPG*, *NCAPH*, *NUSAP1*, *PBK*, *TOP2A*, *TROAP*, *TTK* and *UBE2C*. All these genes are involved in relations described in both datasets, except for *BUB1B*, *CENPF*, *TPX2*, which are exclusive to astrocytoma.

The 'Astrocytoma *vs.* Oligodendroglioma *vs.* GBM' case (Fig. 2C) results in the most comprehensive network. It comprises 30 edges. Most of them are shared by astrocytoma and oligodendroglioma ($n = 13$, blue), and by all three subtypes ($n = 15$, orange). Only one edge is shared between GBM and astrocytoma (green), as well one

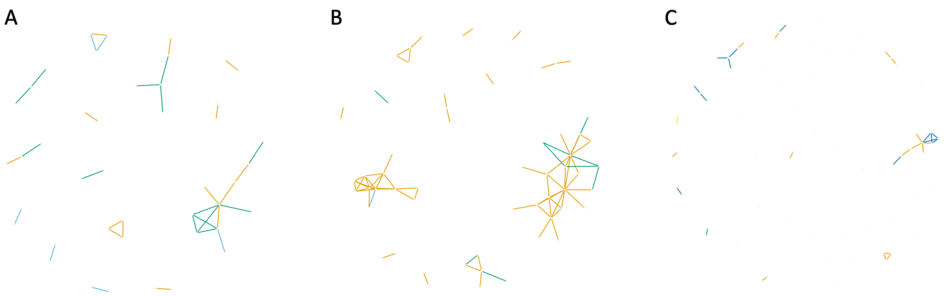

Fig. 2 JGL networks with $\lambda_1 = 0.95$ and $\lambda_2 = 0.01$, for three cases of study: (A) LGG vs GBM, (B) astrocytoma vs oligodendroglioma, and (C) astrocytoma vs oligodendroglioma vs GBM. When considering two classes (A and B), shared edges are colored in orange, while exclusive edges related to the first and second class are green and blue, respectively. When considering three classes (C), shared edges across the three types are highlighted in orange, shared edges between astrocytoma and oligodendroglioma are in blue, shared edges between astrocytoma and GBM are in green, while exclusive GBM edges appear in yellow

edge has been estimated only in GBM dataset (yellow). There are no exclusive edges of astrocytoma or oligodendroglioma, and shared edges between oligodendroglioma and GBM. A subnetwork of shared edges can be observed, composed by the following genes: *TMEM125*, *ERMN*, *GJB1*, *MOG*, *CARNS1*, *KLK6*, *MAG*, *MBP*, and *MOBP*.

For each case, the 5 genes with the highest number of connections were selected. These genes, referred to as *hubs*, are listed in Table 1, where the number of connections is reported (in brackets).

To further explore the estimated joint networks we focused on case C, i.e., the one considering the three glioma types. Figure 3 shows the same joint networks as in Fig. 2C, but emphasizing the nodes constituting the graph, instead of the edges. For consistency, nodes in orange represent genes that have been commonly selected by the three glioma types, while blue and green nodes are shared between LGG (astrocytoma and oligodendroglioma), and astrocytoma-GBM types, respectively. Yellow nodes are exclusive to GBM. In this representation, the relation between *RPSA* and *RPSAP58* appears as potentially relevant for GBM. On the other hand, two

Table 1 Hub genes selected by JGL for the three cases

LGG *vs.* GBM		**Astrocytoma *vs.* Oligodendrogioma**		**Astrocytoma *vs.* Oligodendrogioma *vs.* GBM**		
Hubs$_{LGG}$	**Hubs**$_{GBM}$	**Hubs**$_{Astro}$	**Hubs**$_{Oligo}$	**Hubs**$_{Astro}$	**Hubs**$_{Oligo}$	**Hubs**$_{GBM}$
MAG(6)	*MAG*(3)	*KIFC1*(10)	*KIFC1*(10)	*MAG*(6)	*MAG*(6)	*MAG*(4)
GJB1(3)	*ANXA2P1*(2)	*TOP2A*(10)	*MAG*(7)	*GJB1*(3)	*GJB1*(3)	*C1QA*(2)
KLK6(3)	*ANXA2P2*(2)	*TMEM125*(7)	*TMEM125*(7)	*KLK6*(3)	*KLK6*(3)	*C1QB*(2)
TMEM125(3)	*ANXA2*(2)	*MAG*(6)	*TOP2A*(7)	*TMEM125*(3)	*TMEM125*(3)	*C1QC*(2)
TOP2A(3)	*C1QA*(2)	*UBE2C*(6)	*UBE2C*(6)	*TOP2A*(3)	*TOP2A*(3)	*DOCK2*(2)

LGG lower-grade glioma, *GBM* glioblastoma

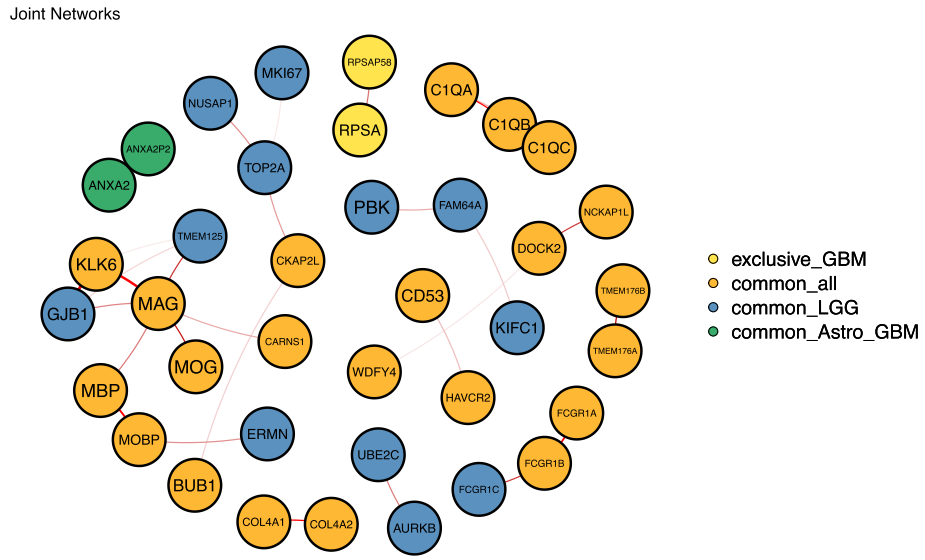

Fig. 3 Case C joint networks representation with gene names. Colors are maintained as the ones provided by the JGL representation. Yellow nodes represent genes that only appear in the GBM network; orange nodes are genes shared across the three types; blue nodes are common between both LGG types (astrocytoma and oligodendroglioma); green nodes are common between astrocytoma and GBM

subnetworks related to LGG are detected, i.e., the one constituted by *PBK*, *FAM64A* and *KIFC1*, and the one involving *UBE2C* and *AURKB*. While the role of *RPSA* in GBM has not yet been investigated, the LGG genes we detected are all known in the context of glioma [56–59]. Interestingly, it has been recently discovered a combined effect of *UBE2C* and *AURKB* genes on glioma histology [60], which is the basis of the 2016 WHO classification. More details about the specific-gene role are provided in Supplementary Materials, Section 0.2.

Clustering

RSKC was applied to the subsets of genes selected through JGL in each case. The rationale of this analysis was to evaluate whether the selected features would identify the known glioma subtypes or disclose new groups according to the RNA-seq data. Table 2 shows the best results obtained by RSKC for each of the combinations of parameters tested (K and $L1$). For all combination of parameters, the best performance was obtained for $L1 = 2$, hence the corresponding parameter value is not reported in Table 2 (complete results in Table S4). The quality of results was evaluated by the *silhouette* and the *Calinski-Harabasz* scores, which have been calculated by considering all cases, including the ones that the method identified as outliers. Larger values of the *silhouette* were obtained when removing the outliers from the dataset, as expected (Supplementary Table S4). For all the cases of study, the *silhouette* scores were higher than 0.5, indicating the good quality of the identified clusters [61]. Overall, the two indexes are in agreement, with the only exception of case A, where the *silhouette* indicates a better cluster with $K = 2$, while *Calinski-Harabasz* is higher for $K = 3$. This result is reasonable, since the starting dataset in case A was based on the variables selected by considering 2 classes, but the LGG class includes two distinct glioma types. In case B, the scores obtained for $K = 2$ and $K = 3$ are comparable, though both indicate a slightly better cluster division with $K = 3$. Interestingly, in case C the better cluster division is the one provided by $K = 2$, even if we are considering the 3 glioma types, separately.

To test the clustering robustness, we performed a validation with a random baseline. For all the cases of study, 1000 datasets composed of randomly selected subsets of variables of the same dimension of the variable sets selected by JGL were created. For each random dataset, RSKC was applied and the clustering performances were evaluated by computing the *silhouette* and *Calinski-Harabasz* indexes. Table 3 shows

Table 2 Summarized results of RSKC applied to the three cases considering the genes selected by JGL (λ_1= 0.095 andλ_2= 0.01), forα= 0.1 and $L1 = 2$. The clusters were evaluated by Silhouette and Calinski-Harabasz scores. ARI was computed to quantify the agreement between clusters and the glioma types, considering the most relevant parameters combination

Case	K	*Silhouette*	*Calinski-Harabasz*	*ARI*
(A) LGG *vs.* GBM	2	0.70	640.95	0.54
	3	0.62	786.37	–
(B) Astrocytoma *vs.* Oligodendroglioma	2	0.58	513.29	0.15
	3	0.59	526.92	–
(C) Astrocytoma *vs.* Oligodendroglioma *vs.* GBM	2	0.73	843.31	0.49
	3	0.61	753.07	0.21

LGG Lower-Grade Glioma, *GBM* glioblastoma

Table 3 Validation of RSKC through random baseline clustering. This table shows the scores corresponding to the best clustering result, the average and the mean value from the 1000 random datasets, comparing it with the reference scores

Case	K	Reference score		Random score (Best – Average – Median)	
		Silhouette	*Calinski-Harabasz*	*Silhouette*	*Calinski-Harabasz*
(A) LGG vs GBM	2	0.70	640.95	0.72 — 0.52 — 0.52	668.21 — 170.93 — 158.06
	3	0.62	746.37	0.59 — 0.33 — 0.34	247.07 — 66.21 — 60.47
(B) Astrocytoma vs Oligodendroglioma	2	0.58	513.29	0.74 — 0.45 — 0.46	264.50 — 57.91 — 46.17
	3	0.59	526.92	0.60 — 0.32 — 0.31	61.16 — 8.24 — 6.92
(C) Astrocytoma vs Oligodendroglioma vs GBM	2	0.73	843.31	0.73 — 0.51 — 0.52	497.87 — 153.68 — 142.14
	3	0.61	753.07	0.64 — 0.33 — 0.33	300.62 — 62.19 — 55.98

Cases A, B and C represent our cases of study, respectively, 'LGG vs GBM', 'Astrocytoma vs Oligodendroglioma', and 'Astrocytoma vs Oligodendroglioma vs GBM'

LGG lower-grade glioma, *GBM* glioblastoma

the results of the best random clusters, as well as the average and the median values compared to the reference score, i.e., the one reported in Table 2. This approach allows the exploration of the complete transcriptomics datasets, providing a graphical representation of the overall score distributions (Figs. S1 and S2). Due to the fact that the *Calinski-Harabasz* index does not have a defined cut-off value, this representation also serves to evaluate the goodness of the corresponding indexes obtained in the different cases of study, by comparing them with the general distribution. In particular, in all cases, the computed *Calinski-Harabasz* scores appear as the highest values compared with all the 1000 random subsets from the complete dataset.

To further investigate the configuration of the identified unsupervised clusters, we compared them with the pre-assigned diagnostic labels to verify if they are in agreement. Table 4 summarizes the result of this comparison. For cases A and B, $K = 2$ is considered, while in case C we decided to show both outcomes obtained with $K = 2$ (the best according to both the considered score) and $K = 3$ (the actual classes we were taking into account).

In case A (LGG vs. GBM), the identified clusters support the defined glioma types, since cluster 1 and 2 are mainly composed by LGG and GBM cases, respectively. In case B, clustering places the majority of oligodendroglioma samples into cluster 2, while cluster 1 is mainly composed by astrocytoma samples. However, 40% of astrocytoma cases were also assigned to cluster 2. In case C, if we consider $K = 3$ there is

Table 4 Cross-comparison between the clusters obtained by RSKC on the datasets composed by the genes selected through JGL (applied to the three cases of study) and the pre-assigned glioma types (according to 2016 WHO CNS classification)

	Case A		Case B		Case C ($K = 3$)			Case C ($K = 2$)	
Cluster	**1**	**2**	**1**	**2**	**1**	**2**	**3**	**1**	**2**
LGG (Astrocytoma)	425	59	156	108	192	27	45	60	424
LGG (Oligodendroglioma)			40	180	179	12	29		
GBM	18	120	-	-	7	92	39	127	11

LGG lower-grade glioma, *GBM* glioblastoma

not a clear distinction of the glioma types in the three clusters. Cluster 1 is mainly composed of LGG samples, and cluster 2 contains mostly GBM, while cluster 3 is a combination of all glioma types. However, by setting $K = 2$ the obtained clustering reproduces the outcome of case A, with a clear distinction of LGG in cluster 1 and GBM in cluster 2.

To quantify how much the identified clusters are in agreement with the diagnostic labels, we computed the Average Rate Index (ARI) in each case of study, by considering the combination of parameters leading to the best results. ARI score provides a measure of how much the clusters agree or disagree with the known labels, by varying in a range of $[-1, 1]$, where the extreme values mean complete disagreement or agreement, respectively, and 0 corresponds to the random assignments. The results (Table 2) are in line with the previous observations. The computed coefficients highlight an overall agreement in cases A and C ($K = 2$), while in cases B and C ($K = 3$) are associated to very low values.

Clustering has been also used to assess the biological information carried by the set of selected variables. To this aim, we compared the clusters obtained by considering the complete dataset with the one related to our cases of study, by fixing $L1 = 2$. We observed that, for $K = 2$, the *silhouette* score computed by considering the complete dataset with was totally comparable with the one related to cases A and C . In all these cases, the values are higher than 0.7 (Table 2), indicating good performances of the clustering method. Table 5 compares the cluster assignments by considering the complete variable set vs case A and C. Most samples are systematically associated to the same group, meaning that no relevant information might be lost despite a considerable dimension reduction (from 16K to around 40 variables). For $K = 3$ (Table S4), the lower values of *silhouette* indicate that the three glioma type are not easily distinguished based on transcriptomics data. With our variable selection (case C, $K = 3$) we are able to slightly improve the quality of clustering (Table S4), but we cannot assess that we have a good clustering performance. We hypothesize this could depend on the labels assigned by following the 2016-WHO classification, which could not be properly explained by the transcriptomics layer. Indeed, while the ARI of case C ($K = 3$) is considerably low (ARI $= 0.21$, S4), the one computed for the clustering taking into account the complete set of variables was very close to the random assignment (ARI $= 0.0055$, Table S4), proving that our variable selection is defining the three glioma types, though these not represent the best clusters based on transcriptomics data.

Table 5 Comparison of the number of samples constituting the clusters obtained by considering the complete set of variables (rows) and the subset of variables in the case studies A and C (columns)

		Case A		**Case C ($K = 2$)**	
		C1	**C2**	**C1**	**C2**
Complete dataset	C1	390	38	387	41
	C2	53	141	59	135

These outcomes refer to the best clustering performances, obtained for $K = 2$ and $L1 = 2$

Potential biomarker discovery

Our analysis highlighted 27 interesting genes, which have been identified either as nodes in subnetworks or as hubs. Literature research revealed that 17 of these genes have been already investigated in the context of glioma, and they are recognized to be involved in many common processes. For instance, 41% of them influence glioma cell proliferation and/or migration [59, 62–65], whereas 47% resulted as differentially regulated in glioma [66–71]. Other genes, namely *C1QA*, *C1QB*, *C1QC*, *ANXA2*, *CENPF*, *NCAPH*, *ERMN*, and *MOBP* have been pointed out as relevant through bioinformatic analyses on glioma datasets [66, 72–75], while *CARNS1* and *DOCK2* have not yet been linked with adult glioma, but they are known to play role in cancer-related processes [76, 77]. More details about the specific processes in which these genes are involved are reported in the Supplementary Material, Section 0.2 [78–92]. These genes represent a possibility for biomarker discovery, but further biological evaluations are needed to assess their potential.

Conclusions

This work aimed at finding potential biomarkers of glioma heterogeneity. The results obtained confirm that astrocytoma and oligodendroglioma are more similar to each other at a transcriptomics level compared to GBM. In particular, our estimated networks show many common relations between the two LGG subtypes, while GBM shares few edges. The K-means clustering also confirms this outcome, since the lowest *silhouette* and *Calinski-Harabasz* scores have been obtained in case B (comparing the two LGG subtypes). Overall, clustering results have been used as a validation of JGL variable selection. Indeed, both the considered scores confirmed good clustering performances, suggesting that a representative subset of genes might have been identified. Clustering outcomes also indicate that the expression of few genes can distinguish different glioma conditions and disclose new groups of patients based on transcriptomics data, since better performances were obtained with lower values of $L1$. Interestingly, in case C, which compares the three glioma types, the best clustering was obtained by considering only $K = 2$ classes. This result highlights the difficulty to distinguish between astrocytoma and oligodendroglioma groups, and it is in agreement with the preliminary UMAP outcome. The investigation of the cluster composition highlights a general agreement between clustering results and pre-assigned diagnosis in distinguishing LGG and GBM. Despite this, the unsupervised clusters do not entirely reflect the patients' glioma types. In particular, the two LGG cases are not coherently distributed into the two clusters in case B, which, compared to the other cases of study, provides worst RSKC performances, suggesting that the used diagnostic labels are not well described by transcriptomics. This assumption is also supported by the results obtained in the comparison between clustering from the complete dataset and the one related to our case C ($K = 3$). Indeed, despite our variable selection provides a slight improvement, this is not enough to obtain a good distinction of the three glioma types. For future studies, it would be interesting to compare the present results with the ones using an updated dataset according to the 2021 classification, to evaluate whether the latest classification yields better separation between known glioma types or if it might reveal new findings. Moreover, multinomial classification models based on transcriptomics data, possibly combined with relevant clinical data (e.g., sex and age) will enable assessing the concordance of the updated

glioma classes with the groups here estimated in an unsupervised way, and further evaluating the features explaining the differences between the groups.

Although our study leads to a list of potentially interesting genes, further analysis is necessary to sustain the already performed literature search. Indeed, on one hand, the existence of previous studies about the role of the identified genes in glioma processes can be seen as a preliminary validation, supporting our findings and the great potentiality of this study. On the other hand, the genes that have not been yet described in the context of glioma might be regarded as candidates for experimental validation and therapy research. Biologically testing the most promising candidates will be the natural next step to validate their role in the genesis, development, and progression of glioma.

Abbreviations

TCGA	The Cancer Genome Atlas
RNA-seq	RNA-sequencing
LASSO	Least absolute shrinkage and selection operator
GBM	Glioblastoma
WHO	World Health Organization
CNS	Central Nervous System
LGG	Lower-grade glioma
glasso	Graphical lasso
JGL	Joint Graphical Lasso
FGL	Fused graphical lasso
SPARCL	Shape-based clustering
RSKC	Robust k-means clustering
TPM	Transcripts per million
UMAP	Uniform manifold approximation and projection
ARI	Average Rate Index

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13040-023-00341-1.

Additional file 1. Supplementary Material.

Acknowledgements
The results shown in this work are based upon data generated by the TCGA Research Network (`https://www.cancer.gov/tcga`).

Authors' contributions
MBL and RC: conceptualization of the study and work supervision. RC: dataset preprocessing, formal analysis, RSKC validation, implementation and visualization. SM: implementation, formal analysis and visualization. MBL: funding acquisition. All authors interpreted and analysed the results, read, edit and approved the final manuscript.

Funding
This work was supported by Fundação para a Ciência e a Tecnologia (FCT) through references CEECINST/00102/2018, CEECINST/00042/2021, UIDB/00297/2020 and UIDP/00297/2020 (NOVA MATH), UIDB/04516/2020 (NOVA LINCS), UIDB/00667/2020 and UIDP/00667/2020 (UNIDEMI), and developed within the project "MONET: Multiomic networks in gliomas (PTDC/CCI-BIO/4180/2020)".

Availability of data and materials
The methodology used to build the updated TCGA glioma dataset used in this study is described in [34].

Declarations

Competing interests
The authors declare no competing interests.

Received: 21 March 2023 Accepted: 13 August 2023
Published online: 26 September 2023

References

1. Weller M, Wick W, Ken Aldape MB, Pfister MBSM, Nishikawa R, Rosenthal M, et al. Glioma. Nat Rev Dis Primers. 2015;1:15017.
2. Lopes MB, Martins EP, Vinga S, Costa BM. The Role of Network Science in Glioblastoma. Cancers. 2021;13(5):1045.
3. Chatsirisupachai K, Lesluyes T, Paraoan L, Van Loo P, de Magalhães JP. An integrative analysis of the age-associated multi-omic landscape across cancers. Nat Commun. 2021;12:2345. https://doi.org/10.1038/s41467-021-22560-y.
4. Bozdag S, Li A, Riddick G, Kotliarov Y, Baysan M, Iwamoto F, et al. Age-specific signatures of glioblastoma at the genomic, genetic, and epigenetic levels. PLoS ONE. 2013;8(4):e62982. https://doi.org/10.1371/journal.pone.0062982.
5. Khan M, Prajapati B, Lakhina S, Sharma M, Prajapati S, Chosdol K, et al. Identification of Gender-Specific Molecular Differences in Glioblastoma (GBM) and Low-Grade Glioma (LGG) by the Analysis of Large Transcriptomic and Epigenomic Datasets. Front Oncol. 2021;11:699594. https://doi.org/10.3389/fonc.2021.699594.
6. Johansen ML, Stetson LC, Vadmal V, Waite K, Berens ME, Connor JR, et al. Gliomas display distinct sex-based differential methylation patterns based on molecular subtype. Neuro-Oncol Adv. 2020;2(1):vdaa002. https://doi.org/10.1093/noajnl/vdaa002.
7. Yang W, Warrington NM, Taylor SJ, Whitmire P, Carrasco E, Singleton KW, et al. Sex differences in GBM revealed by analysis of patient imaging, transcriptome, and survival data. Sci Transl Med. 2019;11(437):eaao5253.
8. Louis DNL, Ohgaki H, Wiestler OD, Cavenee WK, Burger PC, Jouvet A, et al. The 2007 WHO Classification of Tumours of the Central Nervous System. Acta Neuropathol. 2007;114:1432–0533.
9. Louis DN, Perry A, Reifenberger G, von Deimling A, Figarella-Branger D, Cavenee WK, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathol. 2016;131:803–20.
10. Louis DN, Perry A, Wesseling P, Brat DJ, Cree IA, Figarella-Branger D, et al. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. Neuro-Oncol. 2021;23(8):1231–51.
11. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. J R Stat Soc Ser B (Methodol). 1996;58(1):267–88.
12. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso. J R Stat Soc Ser B (Stat Methodol). 2004;67(1):91–108.
13. Zou H. The Adaptive Lasso and Its Oracle Properties. J Am Stat Assoc. 2006;101(476):1418–29.
14. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B (Stat Methodol). 2006;68:49–67.
15. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B (Stat Methodol). 2005;67(2):301–20.
16. Torres R, Judson-Torres RL. Research Techniques Made Simple: Feature Selection Biomarker Discovery. J Investig Dermatol. 2019;139(10):2068-2074.e1. https://doi.org/10.1016/j.jid.2019.07.682.
17. Lopes MB, Vinga S. In: Pham TD, Yan H, Ashraf MW, Sjöberg F, editors. Learning Biomedical Networks: Toward Data-Informed Clinical Decision and Therapy. Cham: Springer International Publishing; 2021. p. 77–92.
18. Lopes MB, Vinga S. Tracking intratumoral heterogeneity in glioblastoma via regularized classification of single-cell RNA-Seq data. BMC Bioinformatics. 2020;21(59).
19. Jubair S, Alkhateeb A, Tabl AA, Rueda L, Ngom A. A novel approach to identify subtype-specific network biomarkers of breast cancer survivability. Netw Model Anal Health Inform Bioinforma. 2020;9:43.
20. Lopes MB, Casimiro S, Vinga S. Twiner: correlation-based regularization for identifying common cancer gene signatures. BMC Bioinformatics. 2019;20(356):1–15.
21. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008;9(3):432–41.
22. Zhao H, Duan ZH. Cancer Genetic Network Inference Using Gaussian Graphical Models. Bioinforma Biol Insights. 2019;13:1177932219839402.
23. Danaher P, Wang P, Witten DM. The joint graphical lasso for inverse covariance estimation across multiple classes. J R Stat Soc. 2014;76(2):373–97.
24. Kondo Y, Salibian M, Zamar RH. RSKC: An R Package for Robust and Sparse K-Means Clustering Algorithm. J Stat Softw. 2016;72:1–26.
25. Ji W, Liu Y, Xu B, Mei J, Cheng C, Xiao Y, et al. Bioinformatics Analysis of Expression Profiles and Prognostic Values of the Signal Transducer and Activator of Transcription Family Genes in Glioma. Front Genet. 2021;12. https://doi.org/10.3389/fgene.2021.625234.
26. Wang R, Wei J, Li Z, Tian Y, Du C. Bioinformatical analysis of gene expression signatures of different glioma subtypes. Oncol Lett. 2018;15(3):2807–14. https://doi.org/10.3892/ol.2017.7660.
27. Wang GM, Cioffi G, Patil N, Waite KA, Lanese R, Ostrom QT, et al. Importance of the intersection of age and sex to understand variation in incidence and survival for primary malignant gliomas. Neuro-Oncol. 2022;24:302–10. https://doi.org/10.1093/neuonc/noab199.
28. Sharma N, Saxena S, Agrawal I, Singh S, Srinivasan V, Arvind S, et al. Differential Expression Profile of NLRs and AIM2 in Glioma and Implications for NLRP12 in Glioblastoma. Sci Rep. 2019;9(1):8480. https://doi.org/10.1038/s41598-019-44854-4.
29. Tomczak K, Czerwińska P. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol Współczesna Onkologia. 2015;19(1A):68–77.
30. Brennan CW, Verhaak RG, McKenna A, Campos B, Noushmehr H, Salama SR, et al. The somatic genomic landscape of glioblastoma. Cell. 2014;155(2):462–77.
31. TCGA. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455(23):1061–8.
32. TCGA. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. New Ingland J Med. 2015;372(26):2481–2498.
33. Louis DN, Ohgaki H, Wiestler OD, Cavenee WK, Burger PC, Jouvet A, et al. The 2007 WHO Classification of Tumours of the Central Nervous System. Acta Neuropathol. 2007;114:97–109.

34. Mendonça ML, Coletti R, Gonçalves CS, Martins EP, Costa BM, Vinga S, et al. Updating TCGA glioma classification through integration of molecular profiling data following the 2016 and 2021 WHO guidelines. 2023. bioRxiv 20230219529134.
35. Jarque CM. In: Lovric M, editor. Jarque-Bera Test. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. p. 701–702.
36. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016;44(8):e71.
37. Mounir M, Lucchetta M, Silva TC, Olsen C, Bontempi G, Chen X, et al. New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. PLoS Comput Biol. 2019;15(3):e1006701.
38. Zhao T, Liu H, Roeder K, Lafferty J and Wasserman L. The huge package for high-dimensional undirected graph estimation in R. J Mach Learn Res. 2012;13:1059–62.
39. Liu H, Lafferty J, Wasserman L. The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs. J Mach Learn Res. 2009;10:12295–328.
40. Cheadle C, Vawter MP, Freed WJ, Becker KG. Analysis of Microarray Data Using Z Score Transformation. J Mol Diagn. 2003;5:73–81.
41. Cuesta-Albertos JA, Gordaliza A, Matrán C. Trimmed *k*-means: an attempt to robustify quantizers. Ann Stat. 1997;25(2):553–76.
42. Kondo Y. Robust Sparse K-Means. CRAN. 2016. Available at: https://cran.r-project.org/web/packages/RSKC/RSKC.pdf. Version 2.4.2.
43. Diaz-Papkovich A, Anderson-Trocmé L, Gravel S. A review of UMAP in population genetics. J Hum Genet. 2021;66:85–91.
44. Gao H, Zhang B, Liu L, Li S, Gao X, Yu B. A universal framework for single-cell multi-omics data integration with graph convolutional networks. Brief Bioinforma. 2023;bbad081:1–11.
45. Khadirnaikar S, Shukla S, Prasanna SRM. Machine learning based combination of multi-omics data for subgroup identification in non-small cell lung cancer. Sci Rep. 2023;13:4636. https://doi.org/10.1038/s41598-023-31426-w.
46. ElKarami B, Alkhateeb A, Qattous H, Alshomali L, Shahrrava B. Multi-omics Data Integration Model Based on UMAP Embedding and Convolutional Neural Network. Cancer Inform. 2022;21:1–7.
47. Kaufman L, Rousseeuw PJ. Finding Groups in Data. An Introduction to Cluster Analysis. New York: Wiley Inter-Science; 1990.
48. Caliński T, Harabasz J. A dendrite method for cluster analysis. Commun Stat. 1974;3(1):1–27. https://doi.org/10.1080/03610927408827101.
49. Dhakan DB, Maji A, Sharma AK, Saxena R, Pulikkan J, Grace T, et al. The unique composition of Indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches. GigaScience. 2019;8(3). https://doi.org/10.1093/gigascience/giz004.
50. Pan X, Burgman B, Wu E, Huang JH, Sahni N, Stephen Yi S. i-Modern: Integrated multi-omics network model identifies potential therapeutic targets in glioma by deep learning with interpretability. Comput Struct Biotechnol J. 2022;20:3511–21. https://doi.org/10.1016/j.csbj.2022.06.058.
51. Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, Perona I. An extensive comparative study of cluster validity indices. Pattern Recog. 2013;46(1):243–56. https://doi.org/10.1016/j.patcog.2012.07.021.
52. Hassani M, Seidl T. Using internal evaluation measures to validate the quality of diverse stream clustering algorithms. Vietnam J Comput Sci. 2017;4(3):171–83. https://doi.org/10.1007/s40595-016-0086-9.
53. Alban TJ, Alvarado AG, Sorensen MD, Bayik D, Volovetz J, Serbinowski E, et al. Global immune fingerprinting in glioblastoma patient peripheral blood reveals immune-suppression signatures associated with prognosis. JCI Insight. 2018;3(21). https://doi.org/10.1172/jci.insight.122264.
54. Khan MT, Prajapati B, Lakhina S, Sharma M, Prajapati S, Chosdol K, et al. Identification of Gender-Specific Molecular Differences in Glioblastoma (GBM) and Low-Grade Glioma (LGG) by the Analysis of Large Transcriptomic and Epigenomic Datasets. Front Oncol. 2021;11. https://doi.org/10.3389/fonc.2021.699594.
55. Bulakbaşı N, Paksoy Y. Advanced imaging in adult diffusely infiltrating low-grade gliomas. Insights Imaging. 2019;10:122. https://doi.org/10.1186/s13244-019-0793-8.
56. Quan C, Xiao J, Duan Q, Yuan P, Xue P, Lu H, et al. T-lymphokine-activated killer cell-originated protein kinase (TOPK) as a prognostic factor and a potential therapeutic target in glioma. Oncotarget. 2018;9(8):7782–95. https://doi.org/10.18632/oncotarget.23674.
57. Fu M, Zhang J, Zhang L, Feng Y, Fang X, Zhang J, et al. Cell Cycle-Related FAM64A Could be Activated by TGF-β Signaling to Promote Glioma Progression. Cell Mol Neurobiol. 2023. https://doi.org/10.1007/s10571-023-01348-2.
58. Xiao YX, Yang WX. KIFC1: a promising chemotherapy target for cancer treatment? Oncotarget. 2016;7(30):48656–70.
59. Ma R, Kang X, Zhang G, Fang F, Du Y, Lv H. High expression of UBE2C is associated with the aggressive progression and poor outcome of malignant glioma. Oncol Lett. 2016;11(3):2300–4.
60. Alafate W, Zuo J, Deng Z, Guo X, Wu W, Zhang W, et al. Combined elevation of AURKB and UBE2C predicts severe outcomes and therapy resistance in glioma. Pathol Res Pract. 2019;215(10):152557. https://doi.org/10.1016/j.prp.2019.152557.
61. Kaufman L, Rousseeuw P. Finding Groups in Data: An Introduction To Cluster Analysis. 1990. https://doi.org/10.2307/2532178.
62. Yang H, Liu X, Zhu X, Zhang M, Wang Y, Ma M, et al. GINS1 promotes the proliferation and migration of glioma cells through USP15-mediated deubiquitination of TOP2A. iScience. 2022;25(9):104952.
63. Zhu L, Zheng Y, Hu R, Hu C. CKAP2L, as an Independent Risk Factor, Closely Related to the Prognosis of Glioma. BioMed Res Int. 2021;2021:5486131.
64. Wu X, Xu B, Yang C, Wang W, Zhong D, Zhao Z, et al. Nucleolar and spindle associated protein 1 promotes the aggressiveness of astrocytoma by activating the Hedgehog signaling pathway. J Exp Clin Cancer Res. 2017;36(1):127.
65. Huang Y, Ouyang F, Yang F, Zhang N, Zhao W, Xu H, et al. The expression of Hexokinase 2 and its hub genes are correlated with the prognosis in glioma. BMC Cancer. 2022;2(900):900.

66. Ma K, Chen X, Liu, Weihai, Yang Y, Chen S, et al. ANXA2 is correlated with the molecular features and clinical prognosis of glioma, and acts as a potential marker of immunosuppression. Sci Rep. 2021;11:20839.
67. Golan N, Adamsky K, Kartvelishvily E, Brockschnieder D, Möbius W, Spiegel I, et al. Identification of Tmem10/Opalin as an oligodendrocyte enriched gene using expression profiling combined with genetic cell ablation. Glia. 2008;56(11):1176–86.
68. Li S, Zou H, Shao YY, Mei Y, Cheng Y, Hu DL, et al. Pseudogenes of annexin A2, novel prognosis biomarkers for diffuse gliomas. Oncotarget. 2017;8(63):106962–75.
69. Wu J, Wang X, Yuan X, Shan Q, Wang Z, Wu Y, et al. Kinesin Family Member C1 Increases Temozolomide Resistance of Glioblastoma Through Promoting DNA Damage Repair. Cell Transplant. 2021;30:1–13.
70. Hao Z, Zhang H, Cowell J. Ubiquitin-conjugating enzyme UBE2C: molecular biology, role in tumorigenesis, and potential as a biomarker. Tumor Biol. 2012;33:723–30.
71. Zheng G, Han T, Hu X, Yang Z, Wang J, Wen Z, et al. NCAPG promotes tumor progression and modulates immune cell infiltration in glioma. Front Oncol. 2022;12:770628.
72. Zhang M, Zhang Q, Bai J, Zhao Z, Zhang J. Transcriptome analysis revealed CENPF associated with glioma prognosis. Math Biosci Eng. 2021;18:2077–96.
73. Chen L, Sun T, Li J, Zhao Y. Identification of hub genes and biological pathways in glioma via integrated bioinformatics analysis. J Int Med Res. 2022;50(6):03000605221103976.
74. Yang Y, Chu L, Zeng Z, Xu S, Yang H, Zhang X, et al. Four specific biomarkers associated with the progression of glioblastoma multiforme in older adults identified using weighted gene co-expression network analysis. Bioengineered. 2021;12(1):6643–54.
75. Mangogna A, Belmonte B, Agostinis C, Zacchi P, Iacopino DG, Martorana A, et al. Prognostic Implications of the Complement Protein C1q in Gliomas. Front Immunol. 2019;10:2366.
76. Zhang L, Zhang Y, Zhang X, Li X, He M, Qiao S. Combining bioinformatics analysis and experiments to explore CARNS1 as a prognostic biomarker for breast cancer. Mol Genet Genomic Med. 2021;9(2):e1586.
77. Huang Y, Luo W, Chen S, Su H, Zhu W, Wei Y, et al. Association of a novel DOCK2 mutation-related gene signature with immune in hepatocellular carcinoma. Front Genet. 2022;13:872224.
78. Ohashi T, Komatsu S, Ichikawa D, Miyamae M, Okajima W, Imamura T, et al. Overexpression of PBK/TOPK relates to tumour malignant potential and poor outcome of gastric carcinoma. Br J Cancer. 2017;116(2):218–26. https://doi.org/10.1038/bjc.2016.394.
79. Qiao L, Ba J, Xie J, Zhu R, Wan Y, Zhang M, et al. Overexpression of PBK/TOPK relates to poor prognosis of patients with breast cancer: a retrospective analysis. World J Surg Oncol. 2022;20(1):316. https://doi.org/10.1186/s12957-022-02769-x.
80. Dong C, Fan W, Fang S. PBK as a Potential Biomarker Associated with Prognosis of Glioblastoma. J Mol Neurosci. 2020;70(1):56–64. https://doi.org/10.1007/s12031-019-01400-1.
81. Yao Z, Zheng X, Lu S, He Z, Miao Y, Huang H, et al. Knockdown of FAM64A suppresses proliferation and migration of breast cancer cells. Breast Cancer. 2019;26(6):835–45. https://doi.org/10.1007/s12282-019-00991-2.
82. Kleylein-Sohn J, Pöllinger B, Ohmer M, Hofmann F, Nigg EA, Hemmings BA, et al. Acentrosomal spindle organization renders cancer cells dependent on the kinesin HSET. J Cell Sci. 2012;125(22):5391–402. https://doi.org/10.1242/jcs.107474.
83. Fu X, Zhu Y, Zheng B, Zou Y, Wang C, Wu P, et al. KIFC1, a novel potential prognostic factor and therapeutic target in hepatocellular carcinoma. Int J Oncol. 2018;52(6):1912–22. https://doi.org/10.3892/ijo.2018.4348.
84. Kostecka LG, Olseen A, Kang K, Torga G, Pienta KJ, Amend SR. High KIFC1 expression is associated with poor prognosis in prostate cancer. Med Oncol. 2021;38(5):47. https://doi.org/10.1007/s12032-021-01494-x.
85. Giet R, Petretti C, Prigent C. Aurora kinases, aneuploidy and cancer, a coincidence or a real link? Trends Cell Biol. 2005;15(5):241–50. https://doi.org/10.1016/j.tcb.2005.03.004.
86. Du R, Huang C, Liu K, Li X, Dong Z. Targeting AURKA in Cancer: molecular mechanisms and opportunities for Cancer therapy. Mol Cancer. 2021;20(1):15. https://doi.org/10.1186/s12943-020-01305-3.
87. Wang T, Wang Z, Niu R, Wang L. Crucial role of Anxa2 in cancer progression: highlights on its novel regulatory mechanism. Cancer Biol Med. 2019;16(4):671–87.
88. Liu X, Ma D, Jing X, Wang B, Yang W, Qiu W. Overexpression of ANXA2 predicts adverse outcomes of patients with malignant tumors: a systematic review and meta-analysis. Med Oncol. 2015;32(1):1–9.
89. Sharma K, Singh J, Pillai PP, Frost EE. Involvement of MeCP2 in Regulation of Myelin-Related Gene Expression in Cultured Rat Oligodendrocytes. J Mol Neurosci. 2015;57:176–84.
90. Brockschnieder D, Sabanay H, Riethmacher D, Peles E. Ermin, A Myelinating Oligodendrocyte-Specific Protein That Regulates Cell Morphology. J Neurosci. 2006;26(3):757–62. https://doi.org/10.1523/JNEUROSCI.4317-05.2006.
91. Chen Y, Meng F, Wang B, He L, Liu Y, Liu Z. Dock2 in the development of inflammation and cancer. Eur J Immunol. 2018;48(6):915–22.
92. Zhao H, Cai W, Su S, Zhi D, Lu J, Liu S. Screening genes crucial for pediatric pilocytic astrocytoma using weighted gene coexpression network analysis combined with methylation data analysis. Cancer Gene Ther. 2014;21(10):448–55.

Publisher's Note