Computational and Structural Biotechnology Journal 20 (2022) 3482–3492

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Multi-omics data integration for subtype identification of Chinese lower-grade gliomas: A joint similarity network fusion approach

Lingmei Li [a], Yifang Wei [a], Guojing Shi [a], Haitao Yang [b], Zhi Li [c], Ruiling Fang [a], Hongyan Cao [a,d,*], Yuehua Cui [e,*]

[a] Division of Health Statistics, School of Public Health, Shanxi Medical University, Taiyuan, Shanxi 030001, PR China
[b] Division of Health Statistics, School of Public Health, Hebei Medical University, Shijiazhuang, Hebei 050017, PR China
[c] Department of Hematology, Taiyuan Central Hospital of Shanxi Medical University, Taiyuan, Shanxi 030001, PR China
[d] Shanxi Medical University-Yidu Cloud Institute of Medical Data Science, Taiyuan, Shanxi 030001, PR China
[e] Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA

ARTICLE INFO

Article history:
Received 19 April 2022
Received in revised form 30 June 2022
Accepted 30 June 2022
Available online 2 July 2022

Keywords:
Joint-SNF
LGG
Multi-omics data integration
Subtypes identification

ABSTRACT

Lower-grade gliomas (LGG), characterized by heterogeneity and invasiveness, originate from the central nervous system. Although studies focusing on molecular subtyping and molecular characteristics have provided novel insights into improving the diagnosis and therapy of LGG, there is an urgent need to identify new molecular subtypes and biomarkers that are promising to improve patient survival outcomes. Here, we proposed a joint similarity network fusion (Joint-SNF) method to integrate different omics data types to construct a fused network using the Joint and Individual Variation Explained (JIVE) technique under the SNF framework. Focusing on the joint network structure, a spectral clustering method was employed to obtain subtypes of patients. Simulation studies show that the proposed Joint-SNF method outperforms the original SNF approach under various simulation scenarios. We further applied the method to a Chinese LGG data set including mRNA expression, DNA methylation and microRNA (miRNA). Three molecular subtypes were identified and showed statistically significant differences in patient survival outcomes. The five-year mortality rates of the three subtypes are 80.8%, 32.1%, and 34.4%, respectively. After adjusting for clinically relevant covariates, the death risk of patients in Cluster 1 was 5.06 times higher than patients in other clusters. The fused network attained by the proposed Joint-SNF method enhances strong similarities, thus greatly improves subtyping performance compared to the original SNF method. The findings in the real application may provide important clues for improving patient survival outcomes and for precision treatment for Chinese LGG patients. An R package to implement the method can be accessed in Github at https://github.com/Sameeerer/Joint-SNF.

1. Introduction

Lower-grade gliomas (LGG), including diffuse low-grade and intermediate-grade gliomas (World Health Organization grades II and III) are the most common infiltrative neoplasms that occur in adult cerebral hemispheres [1]. Most patients exhibit high postoperative recurrence risk [2], and may further deteriorate into glioblastomas (grade IV, GBM). Historically, histologic classifications and tumor grades of LGG have been used to assist therapeutic interventions. However, patients with the same grade often have distinct molecular characteristics and prognosis [3]. With the rapid development of molecular biology research on LGG, identification of molecular subtypes and biomarkers have been explored to guide clinical decision-making [4,5]. The identification of a group of genetic lesions including isocitrate dehydrogenase 1/2 (*IDH*1/2) mutation and codeletion of chromosome 1p and 19q (1p/19q) [6,7] has been a major progress in recent years. Based on these two genetic alterations, cumulative evidence shows that LGG can be classified into three subtypes with different clinical outcomes [8]. Patients with *IDH* mutation (*IDH*-mut) have longer survival than those with *IDH* wild-type (*IDH*-WT) [9]. Nonetheless, the current biomarkers still cannot adequately predict the overall survival for all LGG patients. For example, *IDH*-WT may occur in WHO grades II gliomas or in recurrent gliomas [10]. Moreover, due to

* Corresponding authors at: Division of Health Statistics, School of Public Health, Shanxi Medical University, Taiyuan, Shanxi, PR China.
E-mail addresses: caohy@sxmu.edu.cn (H. Cao), cuiy@msu.edu (Y. Cui).

https://doi.org/10.1016/j.csbj.2022.06.065
2001-0370/© 2022 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.

substantial heterogeneity between LGG, the elaboration of optimal therapeutic strategies at the individual level is still a great challenge [11]. Thus, there is a pressing need to develop reliable approaches to identify patients with high risk of deterioration and find new molecular targets for developing effective treatment strategies.

Recent technological advances allow us to understand the onset and progression of tumors and identify the risk factors, molecular basis and prognostic biomarkers underlying invasive tumors [12,13]. Similar to other subtyping studies, multi-omics data integration remains the preferred approach to obtain the accurate subtype of LGG patients. Multi-omics data integration enables the joint analysis of multiple data types to provide a comprehensive understanding of the biological system and offer insights into the crucial associations between different omics data types [14]. It has been a well-established strategy for identifying molecular subtypes and elucidating pathogenesis in cancer [15]. However, multi-omics integration faces several major challenges, such as the curse of dimensionality and the modeling of interactions between the different types of omics data [16,17]. Methods that can address the potential challenges in multi-omics integration can be largely classified as multivariate, concatenation-based and transformation-based methods [18]. Multivariate methods such as partial least squares or canonical correlation analysis, treat different omics data individually to discover associations between them. Concatenation-based integration combines omics data into a single matrix which is then used as input for low rank-based approximation or latent factor analysis in a low-dimensional space. Focusing on the shared information and integrative dimension reduction of multi-omics data, Lock *et al.* [19] proposed the Joint and Individual Variation Explained (JIVE) method, a typical example of the concatenation-based method. It uses a decomposition method and segregates the combined omics data matrix into three terms, a low-rank joint variation matrix between data sets, a low-rank individual specific matrix and the residual noise. The method can separate synergistic activities common to all data types from individual ones specific to a particular data type. This method was applied to gene expression and miRNA of GBM tumor samples from the TCGA database, showing better characterization of tumor types and better understanding of the biological interactions between different data types.

The transformation-based approaches integrate omics data after transforming each omics data type into an intermediate and common form (e.g. graph or kernel matrix). They have the advantage of capturing individual omics characteristics in the transformation step and are robust to different data measurement scales [20]. One of the popular methods is the Similarity Network Fusion (SNF) algorithm [21] which creates an individual sample similarity matrix for each data type and then fuses these into a single similarity network using a message passing theory, making the combined networks more coherent during each iteration. Like any transformation-based methods, no feature selection step is required [22]. However, due to the limitation of measurement technology and inherent natural variation, unavoidable noise features can dilute clustering signals, leading to potential spurious associations between samples [23].

Collectively, the two methods are largely complementary and individually, and they have their own merits in certain aspects. To fully utilize the strength of the two methods and achieve better subtyping performance, in this work, we proposed a Joint-SNF method which employs SNF to obtain the fused sample similarity matrix by integrating the joint structure extracted by the JIVE method. The fused network matrix enhances strong similarities and weakens spurious associations between samples while reducing the noise. We performed simulation studies to compare the performances of the proposed Joint-SNF method with the original SNF method. We further applied the Joint-SNF method to integrate mRNA, DNA methylation and miRNA expression data obtained from the CGGA database, aiming to discover molecular subtypes of Chinese LGG patients with different prognoses. For the identified subtype with the worst prognosis, in-depth bioinformatics analysis was conducted to uncover key pathways and biomarkers that could explain the underlying molecular mechanism. Our method offers a promising new strategy to the toolbox of cancer subtyping with multi-omics data integration.

2. Materials and methods

2.1. Study cohort

The data in this study included clinical data and three types of omics data, downloaded from the Chinese Glioma Genome Atlas database [24] (CGGA, https://www.cgga.org.cn). CGGA, an open-access database, was launched by a team from Beijing Tiantan Hospital, Capital Medical University in 2012 and opened in 2019. LGG patients (WHO grade II and III) with survival time (from initial diagnoses to death, or to the last follow up) and survival status were included for further analysis. All three types of omics data including gene-level mRNA expression data (mRNA-array_301), gene-level DNA methylation data (methyl_159) and gene-level miRNA expression data (microRNA_198) were available for 86 LGG patients.

2.2. Statistical method

2.2.1. Joint similarity network fusion (Joint-SNF)

Joint-SNF method uses SNF to integrate the joint structure extracted by JIVE method to obtain the fusion matrix. The fused network enhances strong similarities and weakens spurious associations between samples while reducing the noise. The realization of Joint-SNF method relies on the following two important algorithms.

Suppose there are k data types and each is measured on p_i features over n samples and is represented by a data matrix $X_i (i = 1, \cdots, k)$ with dimension $p_i \times n$. The k data matrices are merged to form a single data matrix X, i.e.,

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix}_{p \times n}, p = p_1 + p_2 + \cdots + p_k$$

To eliminate baseline differences caused by different data dimensions and scales, each data type is centered by row-wise subtraction of its means, then scaled by applying Frobenius norm, i.e., $X_i^{scaled} = \frac{X_i}{\|X_i\|}$.

(1) Extraction of joint structures by JIVE.

JIVE is a method of integrating multiple datasets via a general decomposition of variation. The decomposition is composed of three parts: a low-rank approximation capturing the joint variation of different data types, low-rank approximations reflecting individual variation of each data type and the residual noise [19].

Each appropriately scaled matrix X can be decomposed into three terms: joint structure matrix J_i associated with X_i, individual structure matrix A_i of X_i and the error matrix E_i. This gives the factorized model,

$$X_i = J_i + A_i + E_i$$

The model assumes that $J_i A_i^T = 0_{p \times p_i}$ for $i = 1, \cdots, k$, that is, the joint and individual terms are uncorrelated. Low-rank constraints are imposed on both the joint and individual structure matrix (i.e., $rank\ (J_i) = r < rank\ (X_i)$, $rank\ (A_i) = r_i < rank\ (X_i)$). The rank ($r$)

of the joint structure J_i associated with X_i is assumed to be the same for different data types. The rank r and r_i are estimated via a permutation testing approach. Then, the matrices J_i and A_i can be obtained using singular value decomposition (SVD). The specific procedure is summarized in Algorithm 1.

Algorithm 1: JIVE decomposition

```
Require: a scaled matrix $X$
Output:
Lists of the joint and individual structure matrices
Details:
1. A singular value decomposition is performed for each data.
   Using the singular values ($\Sigma$) and the right singular values
   (V), the reduced data set is $\Sigma V'$.
2. Set J = $\{J_1, \ldots, J_k\}$ by a rank $r$ singular value decomposition of
   a scaled matrix $X$. Save the right singular values (V).
3. Set $A_i = A_i \prod_{k \neq i} (I - V_i V'_i)$ by a rank $r_i$ singular value
   decomposition of $X_{Individual}$ = $(X - J)(I - VV')$ if orthogonality
   is enforced between individual structures. This is the first
   iteration.
4. $A_i$ is obtained by a rank $r_i$ singular value decomposition of
   $X_{Individual}$ = $(X - J)(I - VV') \prod_{k \neq i} (I - V_i V'_i)$ if the orthogonality
   constraint is imposed between individual structures. Save
   the right singular values ($V_i$).
5. Repeat steps 2–4 until the Frobenius norm of the difference
   between the current and previous iteration in both $J$ and $A$
   is less than some threshold.
6. Return results (J, A, and the ranks used in the
   decomposition).
```

Since the joint structure matrix J_i associated with X_i captures the common structures shared between different data types, it contains common information that can potentially enhances the subtyping performance. Thus, they are used as the input matrix into SNF method for subsequent subtyping.

(2) Similarity Network Fusion (SNF).

SNF is a similarity-based method to integrate multi-omics data by constructing and fusing sample-sample similarity networks of patients [21]. Suppose we have n samples and r joint structures. A patient similarity network is denoted as a graph $G = (V, E)$. The nodes V are patients and weighted edges E form an $n \times n$ similarity matrix $W(i,j)$ measuring the similarity between patients x_i and x_j. $W(i,j)$ is computed by a scaled exponential similarity kernel as follows,

$$W(i,j) = exp\left(-\frac{\rho^2(x_i, x_j)}{\mu \varepsilon_{i,j}}\right)$$

where $\rho(x_i, x_j)$ denotes the Euclidean distance between patients x_i and x_j, μ is a hyperparameter that can be empirically set, $\varepsilon_{i,j} = \frac{mean(\rho(x_i,N_i)) + mean(\rho(x_j,N_j)) + \rho(x_i,x_j)}{3}$ which is used for removing the scaling problem, and $mean(\rho(x_i, N_i))$ is the average value of the distances between x_i and each of its neighbors.

After constructing sample-sample similarity matrices from multiple data sources, we then fuse these similarity matrices into one similarity network. Procedure of fusion is summarized in Algorithm 2. Given matrix W, a normalized kernel matrix P carrying the full information about the similarity to all others for each patient and a local kernel matrix S encoding the similarity of each patient to its nearest neighbors are obtained. Then, P^v and S^v for the vth joint structure $(v = 1, 2, \cdots, r)$ can be obtained. A message-passing process is then used to iteratively update similarity networks to realize the fusion of networks.

Algorithm 2: Similarity network fusion

```
Input: a similarity matrix W with $W(i,j)$
Output: the final fused network $P_{final}$
1. if $j \neq i$ then
    normalize the weight matrix $P(i,j) = \frac{W(i,j)}{2\sum_{k \neq i} W(i,k)}$
else
    $P(i,j) = 1/2$
end if
if $j \in N_i$ then
    $S(i,j) = \frac{W(i,j)}{\sum_{k \in N_i} W(i,k)}$, where $N_i$ is a set of neighbors for patient
    $x_i$
else
    $S(i,j) = 0$
end if
2.obtain$P^v$ and $S^v$ for the $v$th joint structure $(v = 1, 2, \cdots, r)$.
3. Iteratively update similarity network
    $P^v = S^v \times \left(\frac{\sum_{k \neq v} P^{(k)}}{r-1}\right) \times \left(S^{(v)}\right)^T$
4.$P_{final} = \frac{\sum_{v=1}^{r} P^v}{r}$
return $P_{final}$
```

This fusion process converges to a single similarity network that summarizes the similarity between samples across all omics data types sharing the common structures. P_{final} represents the final fused network. The network obtained by the Joint-SNF method is used for further spectral clustering analysis which can capture the global structure of a graph [25].

3. Simulation study

We carried out simulation study to demonstrate the performance of the Joint-SNF method by comparing it with the original SNF method. The simulation design follows the following principles: (i) Each data type has an independent clustering structure, as well as overlapping parts with other omics data types; (ii) The overall clustering structure can be obtained only by integrating information from all omics data types; and (iii) All data types are contaminated with Gaussian noises.

3.1. Simulation settings

The generation of simulated datasets is similar to those reported elsewhere [26,27]. Here, we considered 200 samples including three types of omics data with 1000 features each. These 200 samples were pre-defined as four subtypes, each with 50 samples. To equip the simulated data matrix with a preset clustering structure, three types of omics data matrices were constructed by setting $X_i^s = mean^s + \varepsilon_i$, where ε_i $N(0, \sigma^2)$ represents random noises; $s \in \{1, 2, 3\}$ is the data type index; $mean^s = \{0,\ 1,\ 2\,\text{or}\,3\}$ is the mean expression level of the features in data type s. The four mean groups represent four subtypes among the samples. Specifically, samples 1–50, 51–150, and 151–200 in X^1 with $mean^1 \in \{0,\ 1,\ 3\}$; samples 1–50/101–150, 51–100, and 151–200 in X^2 with $mean^2 \in \{0,\ 2,\ 3\}$; samples 1–100, 101–150, and 151–200 in X^3 with $mean^3 \in \{2,\ 1,\ 3\}$. We also varied the noise level to make the clustering more challenging by generating three datasets named SimData1 ($\sigma^2 = 8$), SimData2 ($\sigma^2 = 12$) and SimData3 ($\sigma^2 = 16$). It is expected that high variance (hence high noise level) will make it more difficult to separate the four clusters. To evaluate the performance of each method at different proportions of signal features, three different signal levels of low, moderate and high signal (5%, 10% and 15%) were considered for each simulated dataset. Each simulation scenario was repeated 1000 times.

3.2. Simulation results

The standardized mutual information (NMI) was considered as a criterion to evaluate the performance. The larger the NMI value, the closer the relationship between the clustering structure and the real label. Shown in Table 1 are the averaged NMI values out of the 1000 simulation runs together with the standard error. The method of using SNF to integrate both joint and individual structure is referred to as JIVE-SNF. Additionally, we have made comparisons with other popular multi-omics integrative clustering methods such as Cancer Integration via Multikernel Learning (CIMLR) [28] and integrative non-negative matrix factorization (IntNMF) [29]. Overall, the Joint-SNF method shows superior performance over the other methods in different simulation scenarios in terms of NMI measures. In particular, when the noise level and the percentage of signal features were high, the NMI of Joint-SNF, JIVE-SNF, SNF, IntNMF and CIMLR are 0.650, 0.339, 0.325, 0.328 and 0.381, respectively, showing great advantage of Joint-SNF over other methods in recovering the true clustering structures. As expected, the NMI obtained by most methods (except JIVE-SNF) increases with increasing signal features at the same noise level.

4. Real data applications

In this study, we used the data of 86 LGG patients from the CGGA database, aged from 17 to 65 years old, with an average age of 38.5 years. Their baseline characteristics were presented in Table 2. A total of 52 patients (60.5%) of histopathologically confirmed grade II and 34 patients (39.5%) of histopathologically confirmed grade III were included. In addition, the gender composition of the patients was about 46.5% for female and 53.5% for male. The majority of patients were primary and only 5 patients were recurrent. By the last follow-up, 44 patients survived and 42 patients died, the survival time ranged from 90 to 5159 days.

We applied Joint-SNF to a total of 86 Chinese LGG patients using three data types including mRNA expression (19,416 mRNAs), miRNA expression (827 miRNAs) and DNA methylation (14,476 genes). Fig. 1 shows the flowchart of the LGG subtyping analysis using the Joint-SNF method. Specifically, the first step is to extract the joint structures among mRNA, miRNA and DNA methylation data with a low rank approximation, then fuse these structures to construct a network to boost similarities and weaken spurious associations between samples for further spectral clustering. Finally, the molecular subtypes of LGG patients can be obtained based on the fused network using the SNF algorithm.

4.1. Subtyping of LGG using Joint-SNF

Applying Joint-SNF, we obtained three subtypes. We further conducted Kaplan-Meier survival analysis to test whether the survival risks among different subtypes identified by Joint-SNF were clinically significant. The log-rank test was performed. The Kaplan–Meier curves constructed by Joint-SNF and SNF are shown in Fig. 2. Clearly, compared to the result of SNF, survival curves of different types obtained by Joint-SNF do not overlap, revealing significant difference. Combining with the p-value result and the previous report [1], we divided Chinese LGG patients into three subtypes. The survival rate of patients with different subtypes is significantly different (χ^2 = 32.8, P = 7.48E-08).

Table 2
Baseline characteristics of 86 LGG patients.

Item	Classification	n (%)/mean ± SD
Age, years		38.56 ± 11.60
Gender	Female	40(46.5)
	Male	46(53.5)
WHO grade	II	52(60.5)
	III	34(39.5)
Sample type	Primary	81(94.2)
	Recurrent	5(5.8)
Survival outcome	Dead	42(48.8)
	Alive	44(51.2)
IDH_mutation_status	Mutant	59(68.6)
	Wildtype	26(30.2)
	NA	1(1.2)

We further explored prognostic value of the subtypes in LGG patients identified by Joint-SNF. Fig. 2A demonstrates the overall survival of different subtypes. A total of 26 patients (30.2%) in Cluster 1 had a 5-year mortality rate of 80.8%, 28 patients (32.6%) in Cluster 2 with a 5-year mortality rate of 32.1%, and 32 patients (37.2%) in Cluster 3 with a 5-year mortality rate of 34.4%. In addition, clinical characteristics are different among different clusters. The results in Table 3 show that compared with the other two clusters, patients in Cluster 1 with the worst prognosis tend to be more older, and most patients are histopathologically confirmed grade III.

4.2. Comparison of Joint-SNF with SNF in subtyping

We compared the subtyping results of Joint-SNF and SNF on LGG to evaluate the differences among the identified subtypes. The p-value result showed that Joint-SNF performed better in identifying clusters significantly associated with patient survival for a fixed number of clusters (see Table 4).

Considering the small sample size in the LGG dataset, we further conducted stability analysis to check the robustness of the subtyping performance with Joint-SNF and SNF following the work by [30,31]. Specifically, we randomly sampled 75% of the LGG patients and performed subtyping using Joint-SNF and SNF assuming different number of clusters (e.g., 3, 4 and 5) and repeated this process 20 times. For each sample split, we conducted a log-rank test to test the difference of the survival curves under the assumed number of clusters. The distribution of the log-rank test p-values obtained by the two methods is displayed in Fig. 3. Overall, Joint-SNF performs better than SNF, though the difference is subtle when

Table 1
The averaged NMI on simulated dataset with the standard errors given in the parenthesis.

Method	SimData1($\sigma^2 = 8$)			SimData2($\sigma^2 = 12$)			SimData3($\sigma^2 = 16$)		
	Low	Moderate	High	Low	Moderate	High	Low	Moderate	High
Joint-SNF	**0.597** (0.059)	**0.660** (0.056)	**0.670** (0.059)	**0.484** (0.053)	**0.641** (0.058)	**0.656** (0.056)	**0.372** (0.069)	**0.598** (0.057)	**0.650** (0.057)
JIVE-SNF	0.364 (0.094)	0.346 (0.076)	0.360 (0.067)	0.307 (0.087)	0.346 (0.093)	0.339 (0.072)	0.235 (0.080)	0.358 (0.098)	0.339 (0.080)
SNF	0.265 (0.032)	0.356 (0.033)	0.466 (0.054)	0.196 (0.034)	0.308 (0.031)	0.362 (0.034)	0.131 (0.035)	0.277 (0.031)	0.325 (0.031)
IntNMF	0.294 (0.031)	0.351 (0.036)	0.414 (0.049)	0.251 (0.030)	0.310 (0.028)	0.339 (0.023)	0.205 (0.039)	0.285 (0.037)	0.328 (0.031)
CIMLR	0.293 (0.042)	0.449 (0.058)	0.668 (0.052)	0.170 (0.036)	0.353 (0.044)	0.452 (0.062)	0.091 (0.043)	0.303 (0.042)	0.381 (0.038)

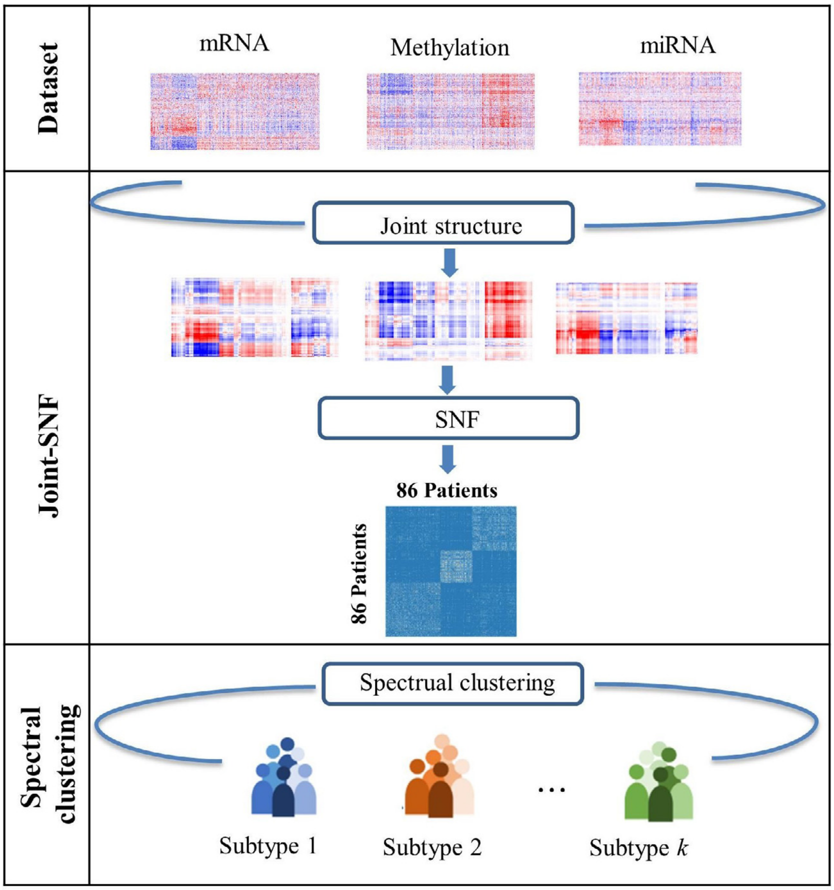

Fig. 1. Schematic representation of the Joint-SNF method used for LGG subtyping.

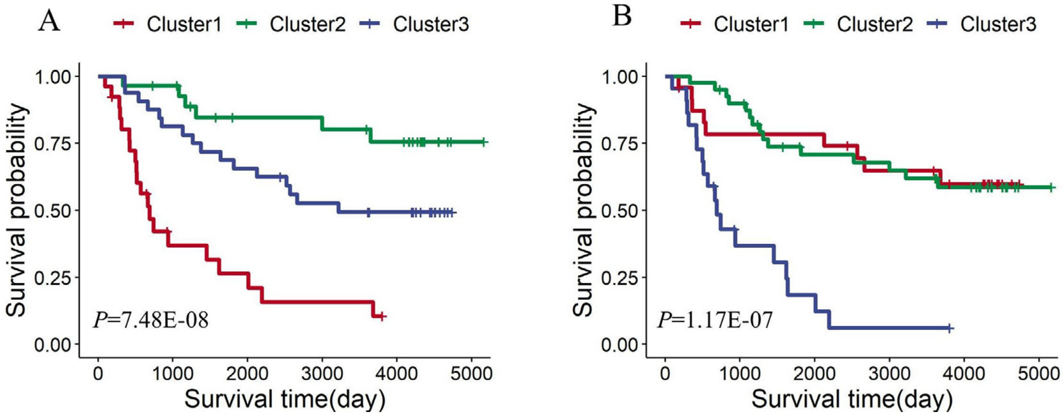

Fig. 2. Kaplan-Meier curves showing overall survival for the three subtypes of LGG obtained by Joint-SNF (**A**) and SNF (**B**).

Table 3
Clinical and pathological characteristics of different subtypes.

Characteristic	Cluster 1(n = 26)	Cluster 2 (n = 28)	Cluster 3(n = 32)
Age, years	42.65 ± 14.61	37.61 ± 8.75	36.06 ± 10.42
Female, n (%)	12(46.1)	14(50.0)	14(43.8)
WHO grade, n (%)			
Grade II	1(3.8)	26(92.9)	25(78.1)
Grade III	25(96.2)	2(7.1)	7(21.9)
Death event, n (%)	20 (76.9)	6 (21.4)	16 (50.0)

Table 4
Comparison of log-rank test p-value of Joint-SNF and SNF across different numbers of subtypes.

Method	p-value under different numbers of clusters		
	3	4	5
Joint-SNF	7.48E-08	6.28E-07	3.10E-07
SNF	1.17E-07	6.37E-07	4.29E-04

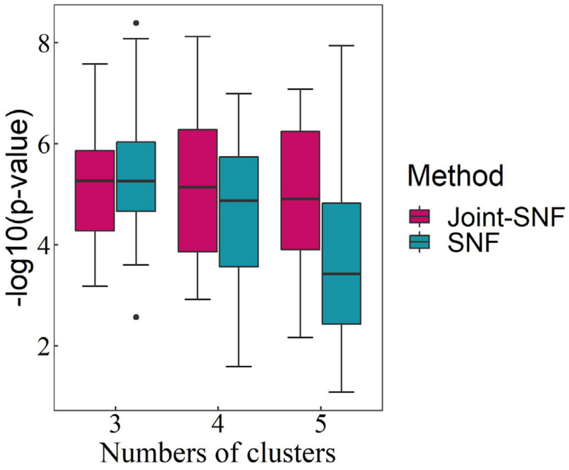

Fig. 3. Boxplots of the -log10(p-value) obtained with the log-rank test for the difference of the survival curves assuming different numbers of clusters using Joint-SNF and SNF over 20 random sample splits.

Table 5
The mean p-value of the log-rank test over 20 random sample splits.

Method	mean p-value of the log-rank test		
	3	4	5
Joint-SNF	8.61E-05	1.83E-04	6.71E-04
SNF	1.68E-04	1.64E-03	1.11E-02

the number of clusters is 3. The mean p-values over the 20 repetitions are summarized in Table 5. The results show that the mean performance of Joint-SNF is better than SNF, in the sense that the survival curves obtained by Joint-SNF can be better differentiated. This stability analysis shows the robustness of Joint-SNF in subtyping the 86 LGG patients.

4.3. Association of prognosis with the identified molecular subtypes

Controlling clinic pathological variables such as age, gender and grade, we performed Cox regression analysis to assess the association between the three subtypes and LGG survival outcomes. As depicted in Table 6, patients in Cluster 1 were 5.06 times higher in risk of death than patients in Cluster 2.

4.4. Biological implications between the identified molecular subtypes

To elucidate the differential manifestations of different molecular subtypes, we performed pathway activity analysis using PROGENy. Kruskal–Wallis test was used to assess biological pathways that show different activities between subtypes. The threshold was set as $P_{adj} < 0.05$.

As shown in Fig. 4, five pathway activities showed significant differences between the three clusters (p < 0.05), with Cluster 1 showing the highest activity in EGFR, VEGF, MAPK, p53 pathways and the lowest Androgen activity. Various signaling pathways are linked to the pathogenesis of different cancers and are considered as potential hallmarks for cancer targeted therapy. Inhibition of certain disease-related signaling pathways may be a promising strategy in cancer prevention or treatment. Thus, the inhibition of EGFR, VEGF, MAPK and p53 pathway activities might lead to improved prognosis of Cluster 1 patients.

4.5. Co-expression network construction and core module identification

We carried out weighted gene co-expression network analysis (WGCNA) to identify gene modules associated with prognosis of LGG patients focusing on the mRNA expression data. A total of top 5000 genes (according to median absolute deviation) were screened to construct the mRNA co-expression network using the R package WGCNA [32]. Briefly, the adjacency matrix was converted into a topological overlap matrix (TOM) when setting power of β to 6 (R^2 = 0.86). Then, we used a dynamic shear tree algorithm to identify gene modules and further merged the relevant modules following a height cutoff of 0.25. Finally, the core modules that may be highly correlated with prognosis in patients were selected for subsequent analyses by associating module eigengenes which summarize the expression of each module with clinical traits.

Fifteen co-expression modules were identified (see Fig. 5A), not including the grey module. A heat map showing the module-trait relationship was used to assess the relationship between each co-expression module with the LGG subtype traits (Cluster 1, Cluster 2, Cluster 3) and other clinical features (WHO grade, Gender, Age, Overall survival). As shown in Fig. 5B, the yellow module was strongly correlated with Cluster 1 (r = 0.74, P = 7E-16) and overall survival (r = -0.52, P = 3E-07). Given that the study goal is to find new therapeutic targets and prolong survival time of patients with extremely poor prognosis, we selected the yellow module for subsequent analysis.

4.6. Functional annotation and enrichment analysis of the core module

To identify the potential biological processes and pathways for 478 genes in the yellow module, Gene Ontology [33] (GO) and Kyoto Encyclopedia of Genes and Genomes [34] (KEGG) analysis were carried out using the R package clusterProfiler [35], to obtain the relevant biological function categories and signaling pathways. The cutoff criterion is set to p-value < 0.05 and q-value < 0.01. As presented in Fig. 6A, these genes were mainly enriched for the following GO terms: nuclear division, organelle fission, chromosome segregation and negative regulation of cell cycle process. In addition, KEGG analysis revealed that these genes were enriched in 11 pathways including cell cycle, p53 signaling pathway, small cell lung cancer and oocyte meiosis (Fig. 6B).

Table 6
Cox regression results of 86 LGG patients.

Item	*Coefficient (SE)*	*Wald Z*	*P*	*HR (95% CI)*
Subtypes				
Cluster1*	1.622(0.626)	2.591	0.009	5.062(1.484,17.260)
Cluster3	0.885(0.488)	1.814	0.070	2.423(0.931,6.306)
Age	0.347(0.325)	1.067	0.286	1.415(0.747,2.679)
Gender	0.020(0.320)	0.064	0.949	1.020(0.545,1.911)
WHO grade	0.603(0.487)	1.240	0.215	1.829(0.704,4.748)

Note: *Showing statistical significance at the 0.05 significance level. Cluster 2 was used as the reference group for subtype comparison. When considering the influence of age, patients were divided into two groups with 36 years old as the cutoff value ($\geq$ 36 vs. $<$ 36).

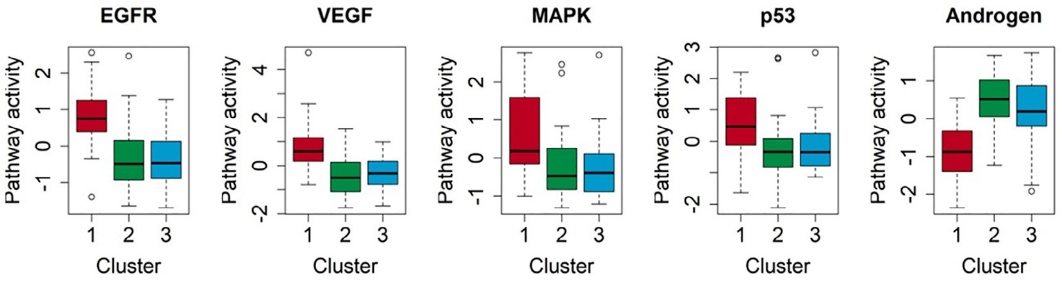

Fig. 4. Boxplots of the pathway activity for 5 pathways in different subtypes.

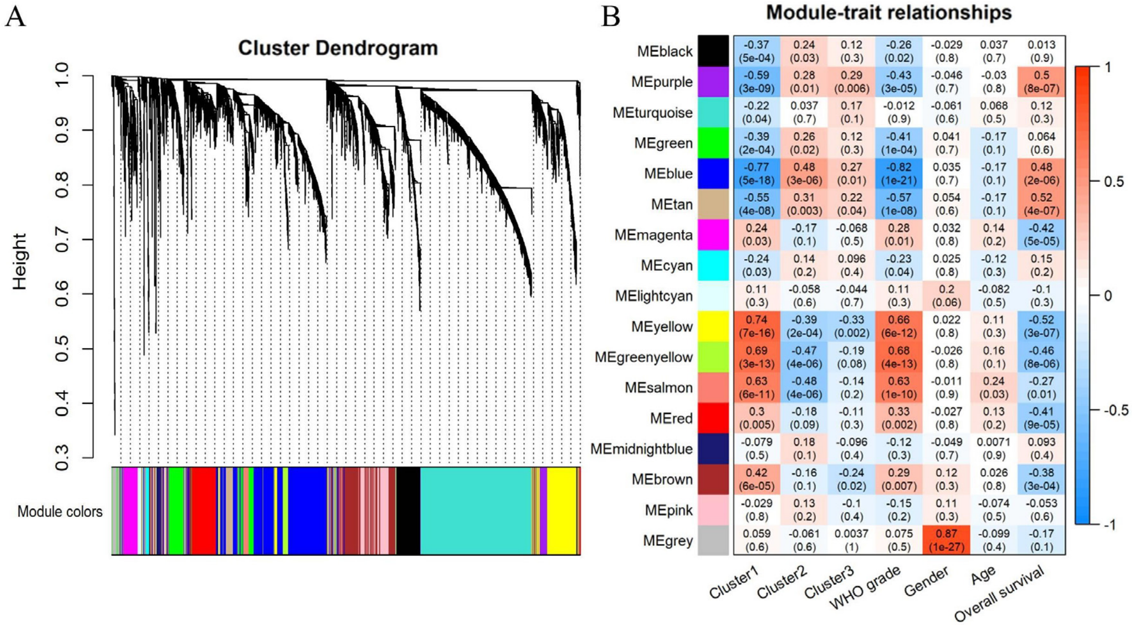

Fig. 5. (**A**) Dendrogram representing hierarchical clustering of identified co-expressed modules. (**B**) Heatmap visualizing the correlation between Eigengene of modules and clinical traits of LGG. Each row represents a color module, and each column represents a clinical feature. Each cell is filled with the correlation and p-value.

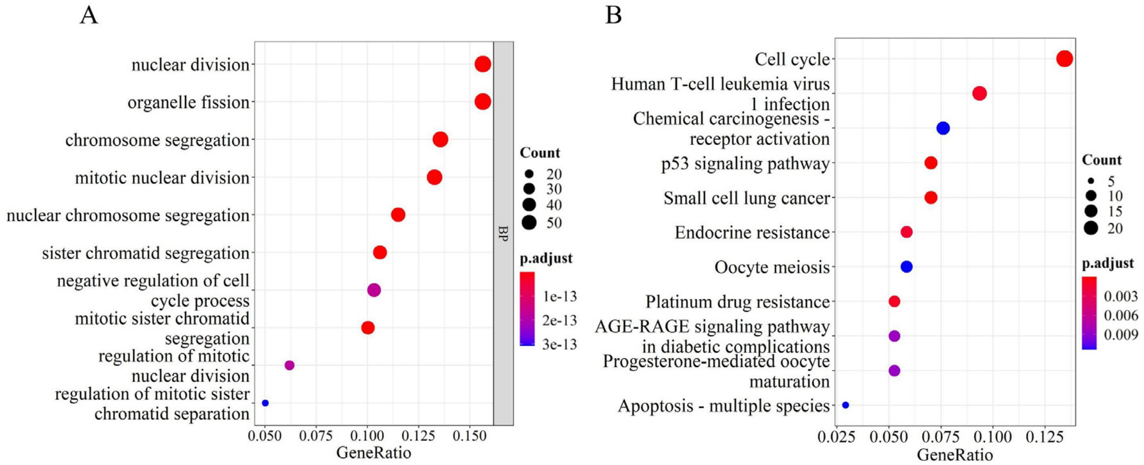

Fig. 6. GO biological process enrichment analysis (**A**) and KEGG enrichment analysis (**B**) for 478 genes in yellow module. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.7. Hub gene identification

Candidate genes were defined as genes correlated with module eigengenes (cor. gene ModuleMembership > 0.9) and clinical traits (cor. gene TraitSignificance > 0.3). As such, we screened 28 candidate genes from the yellow module according to the criteria. The expression heatmap of candidate genes in different subtypes was presented in Fig. 7A. It can be seen that the expression levels of

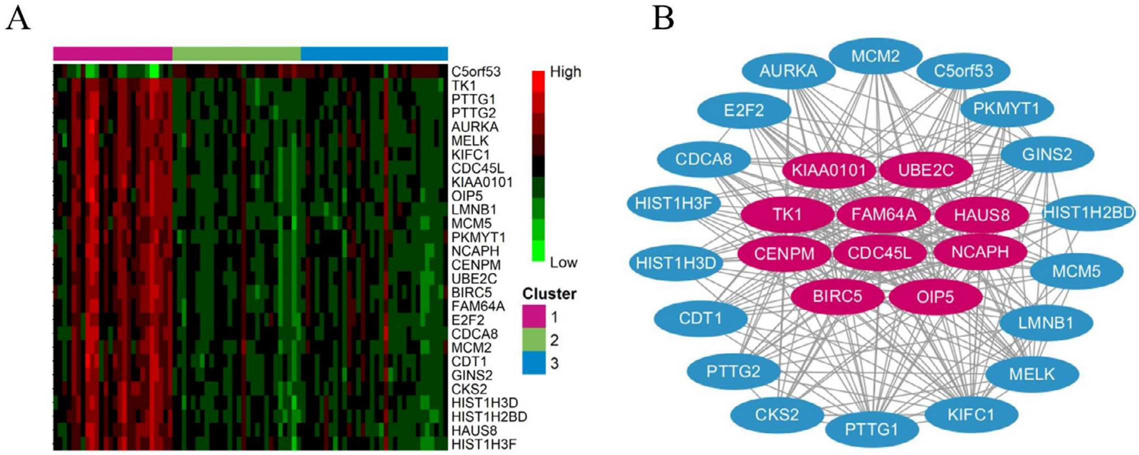

Fig. 7. (**A**) Heatmap reflecting the expression level of candidate genes in the three subtypes. Each row corresponds to a gene feature and each column corresponds to a patient. Red and blue colors indicate relatively high and low gene expressions. The three colored bars on the top indicate subtype cluster 1, 2 and 3 from left to right. (**B**) Network diagram of the interactions between hub genes (red) and candidate genes (blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

candidate genes vary in different subtypes. More specifically, the expression of candidate genes was higher in Cluster 1 which has the worst prognosis.

To identify hub genes, CytoHubba plugin in the Cytoscape software was employed to measure the Maximal Clique Centrality (MCC) score of candidate genes. MCC has been considered as a powerful indicator for identifying central nodes in co-expression networks [36]. The top 10 highly connected genes were used as the hub genes for further analysis, namely *NCAPH*, *CENPM*, *CDC45L*, *TK1*, *FAM64A*, *UBE2C*, *BIRC5*, *KIAA0101*, *OIP5*, *HAUS8*. The interaction between hub genes and candidate genes was visualized by using the Cytoscape software. Shown in Fig. 7B, we can see that candidate genes are less connected to all other candidate genes and more connected to the hub genes, indicating the importance of the hub genes in regulating other genes.

4.8. Evaluation of the prognostic value of hub genes

To verify the prognostic value of the 10 hub genes, all patients were divided into two groups based on the median expression value of hub genes, with patients higher than or equal to the median value assigned to the high-level group and patients lower than the median value assigned to the low-level group. We performed survival analysis to evaluate the statistical significance of survival outcomes between two groups using R package survival and survminer. Kaplan-Meier analysis showed that 9 of the 10 hub genes (*FAM64A*, *OIP5*, *NCAPH*, *KIAA0101*, *UBE2C*, *TK1*, *CDC45L*, *BIRC5*, *CENPM*) were significantly correlated with prognosis ($P < 0.05$). The relationship between all 9 hub genes and the prognosis of LGG patients was such that the higher expression of the gene, the poor the prognosis (see Fig. 8).

5. Discussion

In this study, we proposed a new method called Joint-SNF to integrate multi-omics data to identify molecular subtypes. The Joint-SNF method is a disease subtyping method making use of the correlation and complementary information between different omics data types. The fused network obtained by the Joint-SNF method enhances strong similarities and weakens some spurious associations between samples while reducing the noise. This method separates signals common to all data types from individual ones and avoids the negative impact of irrelevant omics data on cancer subtyping. By extracting the joint structure between omics data types, the original data can be effectively reduced in dimensionality without losing key information. Both simulation studies and LGG subtyping application have demonstrated that Joint-SNF achieves efficient and accurate subtyping compared to the original SNF method based on original features.

Three LGG subtypes (Cluster 1, 2 and 3) identified by Joint-SNF differed significantly in survival outcome. We observed that Cluster 1 with 26 subjects had the worst survival rate with a high 5-year mortality rate of 80.8%, compared to the other two clusters with a 5-year mortality rate of 32.1% and 34.4%. Furthermore, after adjusting for the effects of covariates, patients in Cluster 1 were 5.062 times higher in mortality compared to patients in Cluster 2.

Focusing on subtypes, we further investigated some unique manifestations of different subtypes through bioinformatics analysis and explored their clinical value, especially whether they could help improve the survival time of patients with poor prognosis. We obtained gene modules that affect the prognosis of LGG patients through WGCNA analysis, of which the yellow module had the highest correlation with prognosis. This indicates that the critical genes in the yellow module may serve as potential biomarkers affecting the progression of Chinese LGG patients. We further analyzed a total of 478 genes with co-expression trends identified in the yellow module. GO functional annotation analysis of these genes showed that they were mainly enriched in nuclear division, organelle division, chromosome separation and negative regulation of cell cycle process. The above biological process are involved in regulating the growth and proliferation of cancer cells and are associated with the recurrence of LGG [37,38]. These genes were subjected to the KEGG pathway enrichment analysis which showed that they were associated with various cancer pathways, such as the p53 signaling pathway, the small cell lung cancer and the cell cycle pathway. The cell cycle and p53 signaling pathway have been reported to play a crucial role in the development of LGG [39].

To identify critical genes in the yellow module, we first obtained candidate genes based on the association among genes and the association between the gene set and the clinical subtypes. The results showed that the expression levels of candidate genes were different in different groups, with high expression levels in Cluster 1, showing the importance of these genes with poor prognosis in Cluster 1. We further screened 10 hub genes according to the MCC score to further investigate their prognostic value. We

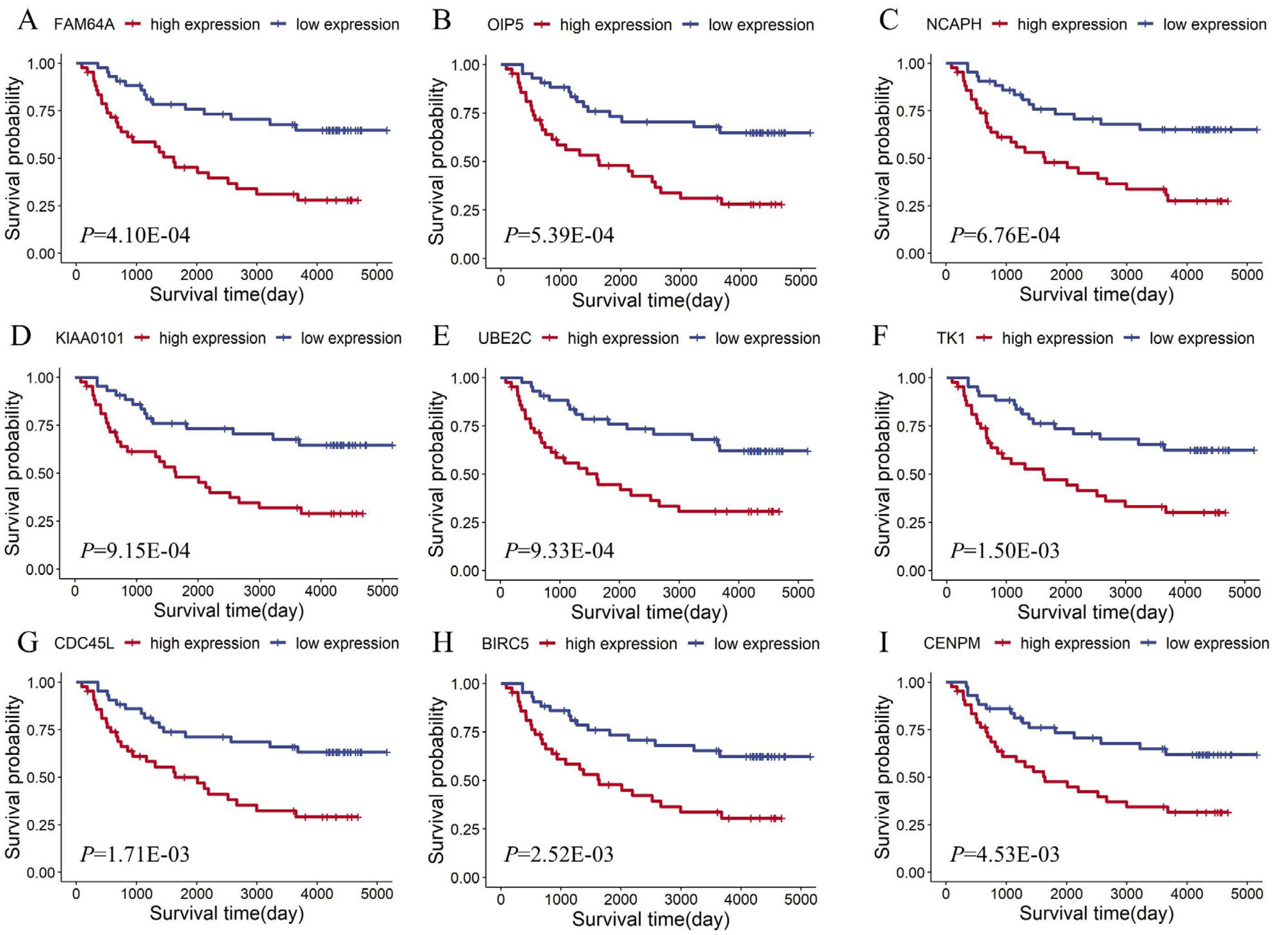

Fig. 8. Plots showing prognostic survival curves of the 9 hub genes sorted in ascending order by p-value.

analyzed the survival of LGG patients with high and low expression of these genes and found that 9 of 10 hub genes were associated with prognosis.

Four of these 9 genes have been reported to be related to gliomas. *UBE2C*, a member of the E2 ubiquitin-conjugating enzyme family, plays a key role in cell cycle control, cell signal transduction and cell differentiation. Additionally, the previous study has shown that *UBE2C* is overexpressed in LGG and its overexpression can lead to poor prognosis [40]. This is consistent with the results of our study, the high expression level of *UBE2C* is associated with poor prognosis. *BIRC5*, also known as survivin, is an immune-related gene belonging to the apoptotic gene family. It has been reported that *BIRC5* may be a potential biomarker and therapeutic target for LGG [41]. Overall, the *UBE2C* and *BIRC5* might be promising candidate biomarkers for improving prognostic outcomes of Chinese LGG patients, although further biological validations are needed. *KIAA0101* encodes a conserved protein which plays an essential role in the regulation of various biological processes [42]. Recently, Liu et al. [43] reported that *KIAA0101* is overexpressed in gliomas, and its expression level was positively correlated with the grade of gliomas. Opa Interacting Protein 5 (*OIP5*) is a cancer-testis specific gene participated in various tumor biological processes [44]. Recent research has shown that it is upregulated in glioblastoma patients and correlated with poor prognosis [45].

Non-SMC condensin I complex subunit H (*NCAPH*) encodes a member of the Barr gene family and a regulatory subunit of the condensin complex. In addition, *NCAPH* was reported to promote tumor formation, proliferation and metastasis [46,47]. Centromere protein M (*CENPM*) is a component of the CENPA-NAC (nucleosome-associated) complex, which plays a central role in the assembly of kinetochore proteins, mitotic progression, and chromosome segregation [48]. It has been reported as a novel biomarker of hepatocellular carcinoma [49], melanoma [50] and bladder cancer [51]. *FAM64A* (also known as *RCS1*, *PIMREG*) plays important biological functions in various cells by accelerating the cell cycle and is abnormally expressed in many tumor tissues [52]. Current studies have found that *FAM64A* was remarkably highly expressed in tumor tissues and cells of patients with Lung Adenocarcinoma [53], and pancreatic cancer [54]. Thymidine kinase l (*TK1*) has been found to be closely related to cancer proliferation [55]. Cell division cycle 45-like (*CDC45L*) has a critical role in the initiation and elongation steps of DNA replication [56], and it is regarded as a promising proliferation marker in tumor cell biology [57]. Although these genes have not been directly reported in gliomas studies, their basic biological functions and carcinogenic properties have been elucidated. This also suggests that they have the potential to affect the occurrence and development of LGG. Moreover, our results demonstrate that the high expression of these genes leads to poor prognosis of patients. Further studies to elucidate their specific mechanisms in LGG are needed.

Our proposed Joint-SNF method provides a new strategy for integrated analysis of multi-omics data and has been successfully applied to LGG patients subtyping. The fused network obtained by Joint-SNF enhances strong similarities and weakens spurious associations between samples while reducing the noise. In addition, this method separates signals common to all data from individual ones and effectively reduces the dimension of original data without losing key information. Overall, our findings may provide novel insights into the subtype of LGG patients and provide important clues for improving patient survival outcomes and for the option of individualized treatment.

CRediT authorship contribution statement

Lingmei Li: Formal analysis, Methodology, Software, Writing – original draft. **Yifang Wei:** Formal analysis. **Guojing Shi:** Formal analysis. **Haitao Yang:** Visualization. **Zhi Li:** Conceptualization. **Ruiling Fang:** Data curation. **Hongyan Cao:** Conceptualization, Methodology, Writing – review & editing. **Yuehua Cui:** Conceptualization, Methodology, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors wish to thank the two anonymous reviewers for their insightful comments that greatly helped improve the manuscript, and the CGGA for maintaining critical public databases and services. The data we used were obtained from the CGGA database (http://www.cgga.org.cn).

Funding

This work was supported in part by grants from NSFC (71403156 to HC, 81872717 to HY), Applied Basic Research Project of Shanxi Province (201901D111204 to HC), Open Fund from Shanxi Medical University-Yidu Cloud Institute of Medical Data Science (IMDS/SXMU-2020 to HC), Natural Science Foundation of Hebei Province (H2019206558 to HY), Startup Foundation for Doctors of Shanxi Medical University (BS201722 to HC), the Education Department of Hebei Province (ZD2018022 to HY), and a fund from Michigan State University (to YC).

References

[1] Brat DJ, Verhaak R, Aldape KD, et al. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. N Engl J Med 2015;372(26):2481–98.
[2] Adenis L, Plaszczynski S, Grammaticos B, et al. The effect of radiotherapy on diffuse low-grade gliomas evolution: confronting theory with clinical data. J Personalized Med 2021;11(8):818.
[3] Kloosterhof NK, de Rooi JJ, Kros M, et al. Molecular subtypes of glioma identified by genome-wide methylation profiling. Genes Chromosom Cancer 2013;52(7):665–74.
[4] Eckel-Passow JE, Lachance DH, Molinaro AM, et al. Glioma groups based on 1p/19q, IDH, and TERT promoter mutations in tumors. New Engl J Med 2015;372(26):2499–508.
[5] Hou Z, Zhang K, Liu X, et al. Molecular subtype impacts surgical resection in low-grade gliomas: a Chinese Glioma Genome Atlas database analysis. Cancer Lett 2021;522:14–21.
[6] Yan H, Parsons DW, Jin G, et al. IDH1 and IDH2 mutations in gliomas. N Engl J Med 2009;360(8):765–73.
[7] Ohgaki H, Kleihues P. Genetic profile of astrocytic and oligodendroglial gliomas. Brain Tumor Pathol 2011;28(3):177–83.
[8] Aoki K, Nakamura H, Suzuki H, et al. Prognostic relevance of genetic alterations in diffuse lower-grade gliomas. Neuro-Oncology 2018;20(1):66–77.
[9] Kloosterhof NK, Bralten LBC, Dubbink HJ, et al. Isocitrate dehydrogenase-1 mutations: a fundamentally new understanding of diffuse glioma? Lancet Oncol 2011;12(1):83–91.
[10] Yuan Y, Qi P, Xiang W, et al. Multi-omics analysis reveals novel subtypes and driver genes in glioblastoma. Front Genet 2020;11:565341.
[11] Duffau H. Paradoxes of evidence-based medicine in lower-grade glioma: To treat the tumor or the patient? Neurology 2018;91(14):657–62.
[12] Maddalena L, Granata I, Manipur I, et al. A Framework Based on Metabolic Networks and Biomedical Images Data to Discriminate Glioma Grades. In International Joint Conference on Biomedical Engineering Systems and Technologies (pp. 165–189). Springer, Cham.
[13] Maddalena L, Granata I, Manipur I, et al. Glioma Grade Classification via Omics Imaging In Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2020) – Volume 2: BIOIMAGING, pages 82–92.
[14] Sathyanarayanan A, Gupta R, Thompson EW, et al. A comparative study of multi-omics integration tools for cancer driver gene identification and tumour subtyping. Briefings Bioinf 2019;21(6):1920–36.
[15] Woo HG, Choi JH, Yoon S, et al. Integrative analysis of genomic and epigenomic regulation of the transcriptome in liver cancer. Nat Commun 2017;8:839.
[16] Palsson B, Zengler K. The challenges of integrating multi-omic data sets. Nat Chem Biol 2010;6(11):787–9.
[17] Wu C, Zhou F, Ren J, et al. A selective review of multi-level omics data integration using variable selection. High-Throughput 2019;8(1):4.
[18] Tini G, Marchetti L, Priami C, et al. Multi-omics integration-a comparison of unsupervised clustering methodologies. Briefings Bioinf 2019;20(4):1269–79.
[19] Lock EF, Hoadley KA, Marron JS, et al. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. Ann Appl Statistics 2013;7(1):523–42.
[20] Ritchie MD, Holzinger ER, Li R, et al. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet 2015;16(2):85–97.
[21] Wang B, Mezlini AM, Demir F, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods 2014;11(3):333–7.
[22] Ruan P, Wang Y, Shen R, et al. Using association signal annotations to boost similarity network fusion. Bioinformatics 2019;35(19):3718–26.
[23] Wu Y, Wang H, Li Z, et al. Subtypes identification on heart failure with preserved ejection fraction via network enhancement fusion using multi-omics data. Comput Struct Biotechnol J 2021;19:1567–78.
[24] Zhao Z, Zhang KN, Wang Q, et al. Chinese glioma genome atlas (CGGA): a comprehensive resource with functional genomic data from Chinese Glioma patients. Genom, Proteom Bioinf 2021;19(1):1–12.
[25] Luxburg UV. A tutorial on spectral clustering. Stat Comput 2007;17:395–416.
[26] Shi Q, Zhang C, Peng M, et al. Pattern fusion analysis by adaptive alignment of multiple heterogeneous omics data. Bioinformatics 2017;33(17):2706–14.
[27] Song W, Wang W, Dai DQ. Subtype-WESLR: identifying cancer subtype with weighted ensemble sparse latent representation of multi-view data. Brief Bioinform 2022;23(1):bbab398.
[28] Ramazzotti D, Lal A, Wang B, et al. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. Nat Commun 2018;9(1):4453.
[29] Chalise P, Fridley BL. Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. PLoS ONE 2017;12(5):e0176278.
[30] Li W, Li Q, Kang S, et al. CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. Nucleic Acids Res 2018;46(15):e89.
[31] Yi T, Ab Ü, Fm A, Pamogk OT. A pathway graph kernel-based multiomics approach for patient clustering. Bioinformatics 2021;36(21):5237–46.
[32] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinf 2008;9:559.
[33] Gaudet P, Dessimoz C. Gene ontology: pitfalls, biases, and remedies. Methods Mol Biol 2017;1446:189–205.
[34] Kanehisa M, Furumichi M, Tanabe M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 2017;45(D1):D353–61.
[35] Yu G, Wang LG, Han Y, et al. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics: J Integrative Biolgy 2012; 16(5): 284–287.
[36] Chin CH, Chen SH, Wu HH, et al. cytoHubba: identifying hub objects and sub-networks from complex interactome. BMC Syst Biol 2014;8(Suppl 4):S11.
[37] Deng T, Gong YZ, Wang XK, et al. Use of genome-scale integrated analysis to identify key genes and potential molecular mechanisms in recurrence of lower-grade brain glioma. Med Sci Monitor: Int Med J Exp Clin Res 2019;25:3716–27.
[38] Qi C, Lei L, Hu J, et al. Serine incorporator 2 (SERINC2) expression predicts an unfavorable prognosis of low-grade glioma (LGG): evidence from bioinformatics analysis. J Mol Neurosci 2020;70(10):1521–32.
[39] Xu J, Hou X, Pang L, et al. Identification of dysregulated competitive endogenous RNA networks driven by copy number variations in malignant gliomas. Front Genet 2019;10:1055.
[40] Dastsooz H, Cereda M, Donna D, et al. A comprehensive bioinformatics analysis of UBE2C in cancers. Int J Mol Sci 2019;20(9):2228.
[41] Wang C, Qiu J, Chen S, et al. Prognostic model and nomogram construction based on autophagy signatures in lower grade glioma. J Cell Physiol 2021;236(1):235–48.
[42] Chang CN, Feng MJ, Chen YL, et al. p15(PAF) is an Rb/E2F-regulated S-phase protein essential for DNA synthesis and cell cycle progression. PLoS ONE 2013;8(4):e61196.
[43] Liu J, Gao L, Liao J, et al. Kiaa0101 serves as a prognostic marker and promotes invasion by regulating p38/snail1 pathway in glioma. Ann Transl Med 2021;9(3):260.
[44] Wang D, Chen Z, Lin F, et al. OIP5 promotes growth, metastasis and chemoresistance to cisplatin in bladder cancer cells. J Cancer 2018;9(24):4684–95.
[45] He J, Zhao Y, Zhao E, et al. Cancer-testis specific gene OIP5: a downstream gene of E2F1 that promotes tumorigenesis and metastasis in glioblastoma by stabilizing E2F1 signaling. Neuro-Oncology 2018;20(9):1173–84.

[46] Ma Q, Xu Y, Liao H, et al. Identification and validation of key genes associated with non-small-cell lung cancer. J Cell Physiol 2019;234(12):22742–52.

[47] Zhan SJ, Liu B, Linghu H. Identifying genes as potential prognostic indicators in patients with serous ovarian cancer resistant to carboplatin using integrated bioinformatics analysis. Oncol Rep 2018;39(6):2653–63.

[48] Izuta H, Ikeno M, Suzuki N, et al. Comprehensive analysis of the ICEN (Interphase Centromere Complex) components enriched in the CENP-A chromatin of human cells. Genes Cells 2006;11(6):673–84.

[49] Xiao Y, Najeeb RM, Ma D, et al. Upregulation of CENPM promotes hepatocarcinogenesis through multiple mechanisms. J Exp Clin Cancer Res: CR, BioMed Central 2019;38(1):458.

[50] Chen J, Wu F, Shi Y, et al. Identification of key candidate genes involved in melanoma metastasis. Mol Med Rep 2019;20(2):903–14.

[51] Kim WT, Seo SP, Byun YJ, et al. The anticancer effects of garlic extracts on bladder cancer compared to cisplatin: a common mechanism of action via centromere protein M. Am J Chin Med 2018;46(3):689–705.

[52] Zhou Y, Ou L, Xu J, et al. FAM64A is an androgen receptor-regulated feedback tumor promoter in prostate cancer. Cell Death Dis 2021;12 (7):668.

[53] Mizuno K, Tanigawa K, Nohata N, et al. FAM64A: a novel oncogenic target of lung adenocarcinoma regulated by both strands of miR-99a (miR-99a-5p and miR-99a-3p). Cells 2020;9(9):2083.

[54] Jiao Y, Fu Z, Li Y, et al. Aberrant FAM64A mRNA expression is an independent predictor of poor survival in pancreatic cancer. PLoS ONE 2019;14(1): e0211291.

[55] Aufderklamm S, Todenhöfer T, Gakis G, et al. Thymidine kinase and cancer monitoring. Cancer Lett 2012;316(1):6–10.

[56] Aparicio T, Guillou E, Coloma J, et al. The human GINS complex associates with Cdc45 and MCM and is essential for DNA replication. Nucleic Acids Res 2009;37(7):2087–95.

[57] Pollok S, Bauerschmidt C, Sänger J, et al. Human Cdc45 is a proliferation-associated antigen. FEBS J 2007;274(14):3669–84.