

Investigating Intra-group Molecular Heterogeneity in IDH-mut Lower-Grade Gliomas Using Integrative Multi-Omics Clustering

Yunting Jiang

Shanghai United International
School, Gubei Campus, Minhang
District, Shanghai 201103, China
E-mail: jyt631218@163.com

Abstract:

This study investigated intra-group molecular heterogeneity in IDH-mutant lower-grade gliomas (LGG) by implementing integrative multi-omics clustering, to provide further evidence of the LGG subtype classification and the molecular heterogeneity within groups. This research collects clinical, mutation, copy number alteration (CNA) and messenger RNA (mRNA) data from the TCGA Lower Grade Glioma dataset via the cBioPortal platform. Hence, two dimensionality reduction methods Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) were used to process the raw data. Then, combining K-means clustering algorithm, samples were grouped. The results showed that, among the overall LGG cohort, 2 molecular types are classified, with significant differences in the proportion of IDH-mutated samples and overall survival. Further clustering of the IDH-mutated group yielded 2 subtypes, with significant differences on CNA burden, mRNA expression and gene markers. Although the overall survival rate did not show significant difference with the 2 subclusters, treatment stratification can be done according to the risk level, involving de-escalation/surveillance-leaning therapy and standard combined therapy respectively. This study demonstrated that, integrative multi-omics clustering methods could effectively reveal the intra-group molecular heterogeneity within the IDH-mutated LGG. Moreover, CNA variation is a more important source of heterogeneity than transcriptional variation. These results provided new guidance in precise classification of gliomas and personalized clinical treatments.

Keywords: lower-grade gliomas; multi-omics clustering; copy number alteration; K-means algorithm

1. Introduction

Diffuse gliomas represent 80% of the primary malignant tumor in the central nervous system [1] exhibiting high invasiveness and recurrence rates, posing great pressure to the global health. Epidemiological data indicate that the five-year survival rate for patients with malignant glioblastoma is about 23.5%, [2]. While even relatively indolent low-grade gliomas (LGGs) often face recurrence risks due to difficulties in complete resection. Traditional histopathological classification primarily relies on histological features, and divided diffuse gliomas into grades II-IV based on histological features, including oligodendroglioma, astrocytoma, and glioblastoma, [3].

However, treatments and prognosis applied relying solely on the histopathologic traits, suffers from the high intra- and inter-observer variability, particularly amongst grade II-III tumors [4]. With the advancement of molecular biology, the World Health Organization (WHO) classification standards for central nervous system tumors have undergone a revolutionary transformation. The fifth edition (WHO CNS5) published in 2021 officially integrated molecular markers into the diagnostic framework, marking the entry of glioma diagnosis and treatment into the "integrated diagnostics" era [5]. For example, isocitrate dehydrogenase (IDH) gene mutations have been found a key molecular stratification marker for differentiating subsets in LGGs: IDH-mutant gliomas exhibit unique metabolic characteristics and relatively favorable prognoses, whereas IDH-wildtype tumors are more aggressive. [6] And the focus of this research will be specifically evaluating the prognosis value of IDH mutation in LGGs (grade II glioma).

Previous studies have several limitations. Most studies focus on the general genomic landscape of the whole glioma, few of them deeply investigate LGGs, which is especially essential at the early stage of prognosis and prevention of tumor progression. Moreover, previous research rarely took IDH as the primary molecular feature for classifying subsets. Furthermore, translating molecular characteristics into personalized treatment strategies remains a significant challenge in precision medicine [7].

This study aims to address the research gap by systematically constructing a precision-stratified treatment framework for LGGs through comprehensive bioinformatics analysis. Objectives include: building a glioma mutation database based on the cBioPortal cancer genomics platform; identifying prognostically valuable gene combination patterns by using K-means clustering algorithm; establishing a multivariate prognostic model based on Cox proportional hazards regression; and proposing personalized treatment strategies based on molecular char-

acteristics. By integrating multi-omics data with machine learning approaches, this study aims to provide reference and tools for precision diagnosis and treatment of gliomas.

2. Literature review

2.1 Genomic landscape of LGG

Numerous studies have consistently demonstrated that glioma is a molecularly heterogeneous disease, which cannot be directly classified or prognosed by traditional histopathological methods. Additionally, several genomic profiling research found recurring mutated genes IDH and tumor protein 53 (TP53), alpha thalassemia (ATRX) to be highly related to the prognosis of LGG. The research of The Cancer Genome Atlas (TCGA) performed an unsupervised clustering and multiplatform integration to classify 3 molecular subtypes, identified by IDH mutation and 1p/19q codeletion status. And it showed a stronger concordance with prognosis than histopathology, Rand index 0.79 vs. 0.19 [8]. And similarly, in pan-glioma research done by Ceccarelli et al., they expanded the framework by clustering the IDH mutant LGG tumors into four subtypes by the epigenomic status (Glioma-CpG Island Methylator Phenotype (G-CIMP) level), and revealed the existence of additional intra-LGG heterogeneity [3]. Compared with the broader genomic framework of LGG given by TCGA, it gives a deeper understanding of the differences in epigenetic characteristics of a specific gene (IDH), which also have a stronger relevance with this research. In a study carried out by Eckel-Passow et al., based on medical records, they found that the simple combination of IDH mutation, telomerase reverse transcriptase (TERT) promoter mutation and 1p/19q codeletion could classify glioma into 5 molecular groups, each associated with distinct age at diagnosis and survival outcomes [9]. In addition to the previous researches, this study explored the clinical applicability of targeted genomic markers, which may contribute to the future treatment regimens. But as a medical record-based study, its focus is not on LGG but the whole glioma cancer, which is not fully responsive to this research topic. And it has a relatively small patient cohort, which becomes an apparent limitation of its statistical power.

2.2 Molecular subtypes and treatment stratification in LGG

Despite the molecule subtypes that had been proved valid for prognosis and classification, the mutation in IDH1 and IDH2 are found an especially essential biomarker for LGG prognosis and had been certified as one of the clin-

ical indicators. In an early 2009 observational study carried out by Hai Yan et. Al, they had reported that IDH1/2 mutations are prevalent in early development of glioma tumors, especially in grade II-III tumors. However, rarely found in grade I tumors. Also, by using Kaplan-Meier (KM) survival curves and log-ranking test, they found that IDH-mutant patients have longer overall survival, 31 month versus 15 month in anaplastic astrocytoma (grade III) and 65 months versus 20 months in glioblastoma (grade IV) [10]. This research provided early evidence, indicating that IDH could act as a biomarker for identifying the stage of tumor. Also, it first established the statistical association between IDH status and clinical outcomes. However, covariates such as age and sex were not fully controlled, limiting the strength of the results of this study. Building upon this discovery, van den Bent et al. found a positive correlation between O⁶-methylguanine-DNA methyltransferase (MGMT) methylation and IDH mutation status, and demonstrated an improvement in overall survival in the group with both MGMT methylation and IDH mutation [11]. Comparing with the previous research, it additionally evaluated the clinical outcomes of implementing different treatments on glioma patients, Procarbazine, CCNU (Lomustine), Vincristine+ Radiation Therapy (PCV+RT) treatment have a higher overall survival comparing with only RT treatment, median overall survival (OS) 42.3 months versus 30.6 months [1]. This advanced the research from molecular level to the therapeutic and prognostic level. In addition, by using Cox proportional hazards regression analysis, adding 1p/19q codeletion status and MGMT methylation status as covariates, it found that in IDH-mut group have hazard rate (HR)=0.53(PCV+RT vs. RT, 95% confidence interval (CI)), indicating that adding PCV could reduce death risk; and in IDH-wild group HR=0.78, indicating that adding PCV yielded no significant difference [11]. This Cox analysis improved the accuracy and reliability of the results compared with survival testing alone.

In the 2021 WHO Classification, the significance of survival difference is again validated. The evidence was synthesized from numerous international cohorts with a strong data coherence which could increase its authority and generalizability. As mentioned in previous studies, the guideline proposed that IDH mutation is not merely a biomarker, but can define distinct diagnostic entities. "Diffuse astrocytic tumors are classified primarily according to IDH mutation status." [5]. Moreover, "astrocytoma", "IDH-mutant" and "Glioblastoma, IDH-wildtype" are officially classified as distinct tumor types, which demonstrated the significance of IDH gene. Since survival analyses had established the association of IDH mutation with prolonged OS, this comprehensive paper argued that the

aim for classification is not academic analysis but clinical decision-making. These diagnoses could provide clinically relevant information for risk stratification and personalized treatments. However, as a consensus-based guideline rather than primary resources, the WHO classification does not generate original survival data or demonstrating deduction method by statistical models, and relies on the quality and limitations of the underlying literature.

2.3 K-means clustering in cancer genomics

K-means clustering algorithm, an unsupervised clustering method, have been utilized in the research of glioma in order to find molecular subtypes, backing up the histological classification. Gene expression-based clustering is applied in the TCGA 2008 GBM study, to identify transcriptional groups classified by different pathway activation patterns. This suggested that instead of solely relying on predefined clinical categories, clustering methods could be an effective tool in capturing intrinsic tumor heterogeneity. With this biological characterization, a study carried out in 2010, refines the GBM classification, and revealed the association between the gene expression-based subtypes like proneural and mesenchymal and different overall survival and treatment response, by also using unsupervised clustering [12]. Comparing those two pieces of literature, the former has a main focus on molecular pathways, while the later targeted on linking the clusters with clinical prognosis, being more relevance with this research and provided a valid reference for the methodology. However, both studies involved a common limitation, which is relying too heavily on transcriptional features and is lacking subjective decisions, for example the number of clusters to improve the stability of subtypes.

K-means clustering algorithm is not only implemented in glioma cancer, but a great range of other cancers to support its generality. The TCGA 2012 breast cancer research uses consensus K-means clustering to classify molecular subtypes. And investigated the relationship between the subtypes and the therapeutic responses. It found that data-driven subtypes discovery was not specific to tumor. This analysis integrated multiple molecular platforms and was more comprehensive, unlike disease-focused studies. In 2003, Monti et al. found that K-means are sensitive to initial conditions and resampling variability by systematically assessing the stability of clustering [13]. This result raised concerns about the reproducibility and also suggested possible limitations of the K-means method. Overall, these studies together proved K-means clustering as a powerful and applicable explanatory tool for assisting meaningful molecular classification, but also stated the necessity of comprehensive validations. This balance

between interpretability and limitation supports the use of K-means clustering in the current study, especially when combined with survival-based validation such as Kaplan Meier and Cox regression analyses.

2.4 Research gap

Firstly, although IDH mutation is widely identified as a key prognostic biomarker in LGG, majority of the classification frameworks still recognize IDH-mut tumors as a homogeneous group after separation from IDH-wildtype gliomas. Secondly, even if the intra-molecular heterogeneity was highlighted in the studies within the IDH-mut populations, such classification was often based on single gene mutation combinations or epigenetic characteristics, rather than combining integrated structural features and transcriptomic dimension.

Most importantly, the potential role of copy number alterations and multi-omics variation in identification of IDH-mut subclusters remains unknown. Thirdly, previous researches tend to emphasize the protective effects of IDH-mut gene, but did not discuss the possibility of discovering meaningful clinical outcomes by implementing an additional molecular stratification within the IDH-mut group.

Lastly, most researches focused on the whole glioma cancer instead of the low-grade glioma with a broad scope. The results may be highly valuable and groundbreaking, however the accuracy of early prognosis, effective prevention of tumor development and the clinical applicability could not see a significant improvement.

3. Methodology

3.1 Data collection

The raw data used in this research is extracted from Public Cancer Genome Database cBioPortal. The dataset selected is the Lower Grade Glioma dataset from the PanCancer Atlas of The Cancer Genome Atlas (TCGA). Four categories of data have been downloaded: clinical data (involves overall survival event, overall survival time, tumor grade, age, Karnofsky Performance Status (KPS) and Eastern Cooperative Oncology Group (ECOG)); mutation data (for identifying IDH mutation status); copy number alteration (CNA) and messenger RNA (mRNA) data. All data are downloaded in txt. format and then processed in the R studio. Different types of data are matched by identifying patient ID, so that clinical data could effectively correspond with the molecular data.

$$x'_{ij} = x_{ij} - \bar{x}_j \quad (\text{Equation 1})$$

is the raw CNA or mRNA molecular feature of the i^{th} LGG

sample; is the mean value of feature j across all patient specimens; is the feature value after mean preprocessing.

$$\Sigma = \frac{1}{n-1} \mathbf{X}^T \mathbf{X}' \quad (\text{Equation 2})$$

is the matrix of centered multi-omics data; is the transpose matrix of ;

is the total number of samples.

$$J = \sum_{i=1}^n \sum_{j=1}^k z_{ij} \|x_i - \mu_j\|^2 \quad (\text{Equation 3})$$

is the total LGG samples; is the predefined molecular cluster count; is the integrated multi-omics vector of sample i ; is the centroid vector of cluster j ; is the binary assignment indicator; is the total within-cluster squared error minimized iteratively.

$$s(i) = \frac{b(i)}{\max(a(i), b(i))} - \frac{a(i)}{\max(a(i), b(i))} \quad (\text{Equation 4})$$

is the average intra-cluster distance of sample i ;

is the minimum average distance to the nearest alternative cluster;

is silhouette coefficient ranging from -1 to 1 for clustering quality evaluation.

3.2 Data analysis

Due to the complexity of the raw data, dimensionality reduction methods appear to be especially essential for initial processing. And in this research, two types of models are used, Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP). PCA is a linear dimensionality reduction method which reduce the multiple dimensions of data by orthogonalizing features, and to maximize the variance retention. First, the integrated CNA and mRNA feature matrix need to be centered (Equation 1). Each raw value is adjusted by subtracting the mean of the corresponding feature and calculating the covariance matrix to find correlated features. (Equation 2) By carrying out the calculations, high-dimensional data could be condensed into individual points, therefore be easily plotted on a 2 dimensional planar based on the 2 most obvious features that the model has identified (PC1 and PC2). Differently, UMAP is a non-linear manifold learning technique. It involves the preservation of local and global data features. The neighbor relationships among each sample and the nearby are constantly obtained to ensure the detection of the optimal non-molecular groupings. The PCA effectively gives a global linear overview of the distribution, while the UMAP remains the partial local relationships and adds detailed features to the distribution. Together, they are used as exploratory tools for assessing potential structure within the integrated CNA

and mRNA feature before the actual clustering step. Unsupervised K-means clustering was implemented to divide samples into homogeneous groups. The objective function is optimized in the algorithm (Equation 3) to minimize the Euclidean distance between samples and cluster centroids which refers to the intra-cluster variance, while maximizing inter-cluster distribution. Multiple random candidate values were tested to determine the optimal number of clusters (k). For each k, clustering was initialized with 25 random center initializations in order to avoid inaccuracy or systematic errors. Silhouette score (s) equation is used to evaluate the cluster quality. (figure-Equation 4) It quantifies the ratio of intra-cluster cohesion to inter-cluster distribution for each sample. The range of Silhouette score is from -1 to 1, and the closer the score is to 1, the more effective the cluster. And, as the silhouette score across all samples appears to be the biggest when $k=2$, so 2 sets of clusters are determined. In this research, two clustering stages were implemented. First, global clustering across all samples using integrated CNA and

mRNA features, followed by an intra-group sub clustering within the IDH-mutant subgroup to investigate molecular heterogeneity.

Kaplan–Meier survival analysis was used to estimate survival probabilities across clusters to assess clinical relationships. Multivariable Cox proportional hazards modelling was applied to adjust for confounding factors. Moreover, associations between the clusters and clinical variables (age, tumor grade, KPS, ECOG) were evaluated using non-parametric tests to account for non-normal variable distributions. Eventually, due to the incompleteness of treatment data, a risk-based framework was developed using survival outcomes and clinical indicators to suggest potential treatment intensity patterns across molecular subtypes. And to transform the results from dimensionality reduction and clustering to treatment stratifications.

4. Analysis and Discussion

4.1 Overview

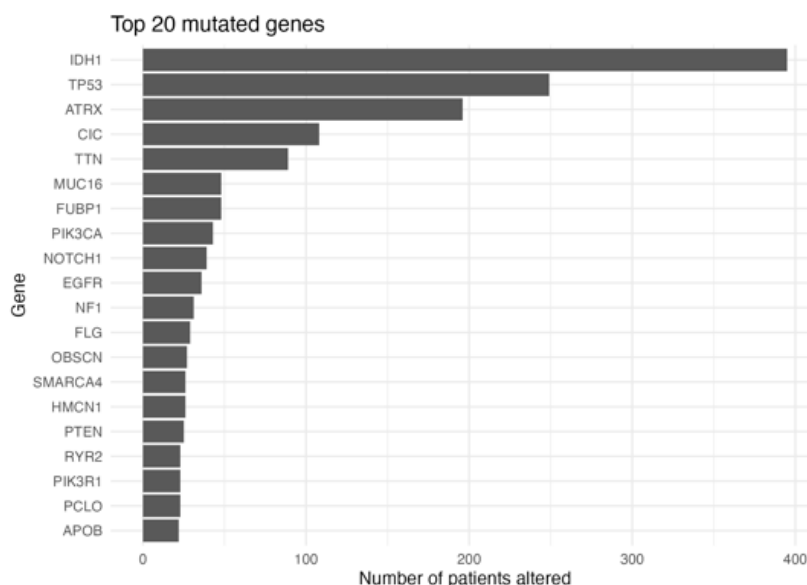


Figure 4.1 The frequency of mutation of the top 20 genes in the lower-grade glioma cohort.

As shown in this figure, IDH1 is the most frequently mutated gene in the cohort, followed by TP53 and ATRX. As these mutations are commonly associated with glioma

development proven in previous research, it is confirmed that the data in this cohort have typical glioma genomic features and provides a foundation for clustering analyses.

4.2 Global clustering

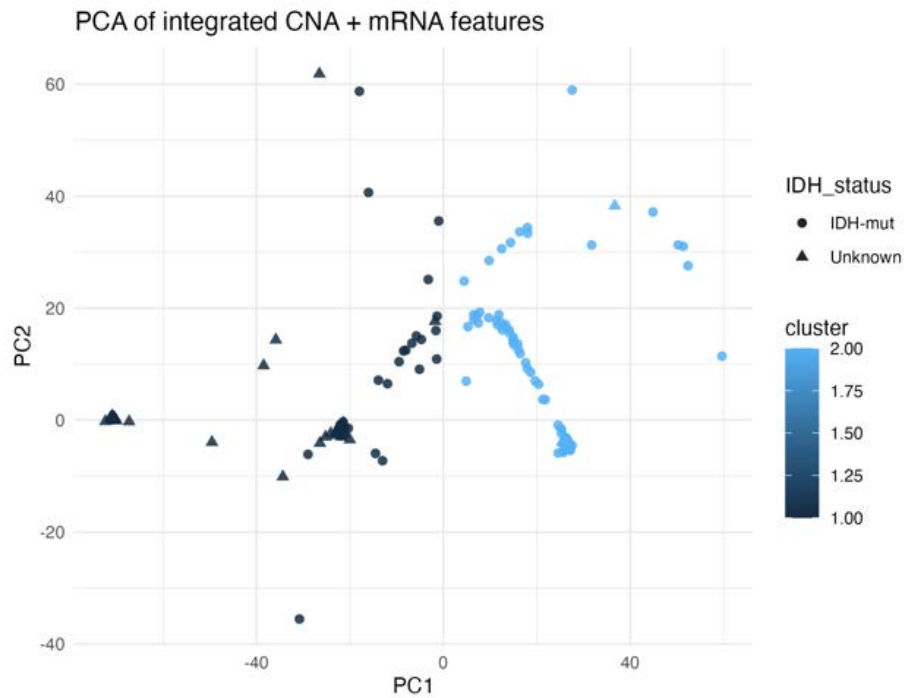


Figure 4.2.1 Principal Component Analysis (PCA) of integrated CNA and mRNA features.

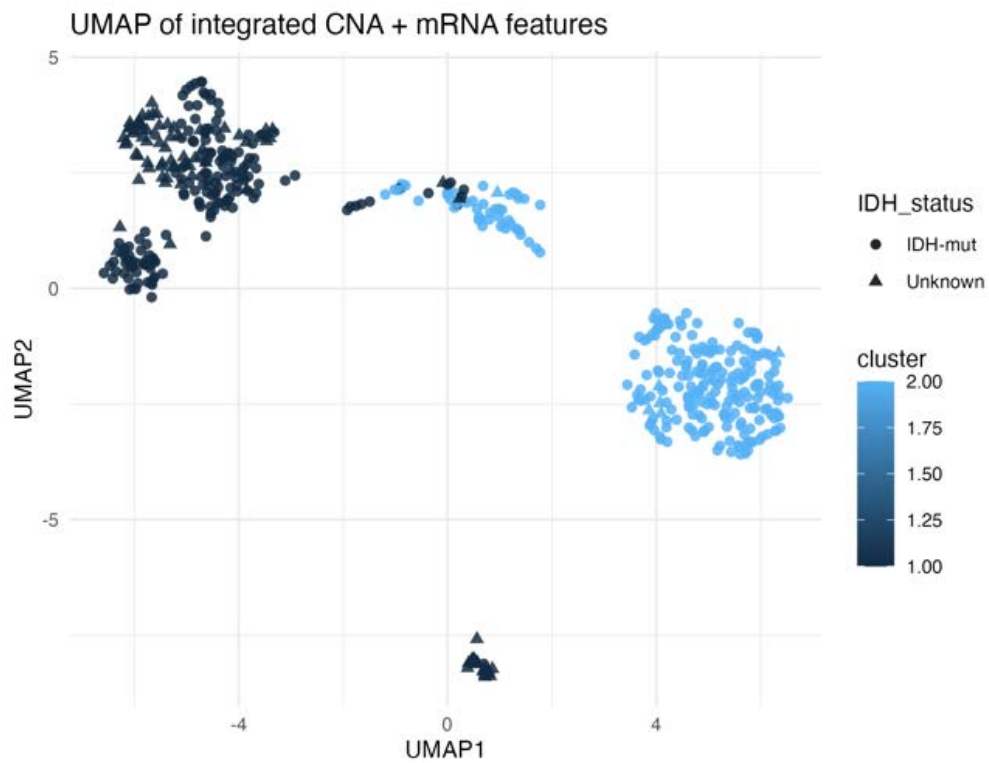


Figure 4.2.2 Uniform Manifold Approximation and Projection (UMAP) visualization of integrated CNA and mRNA data.

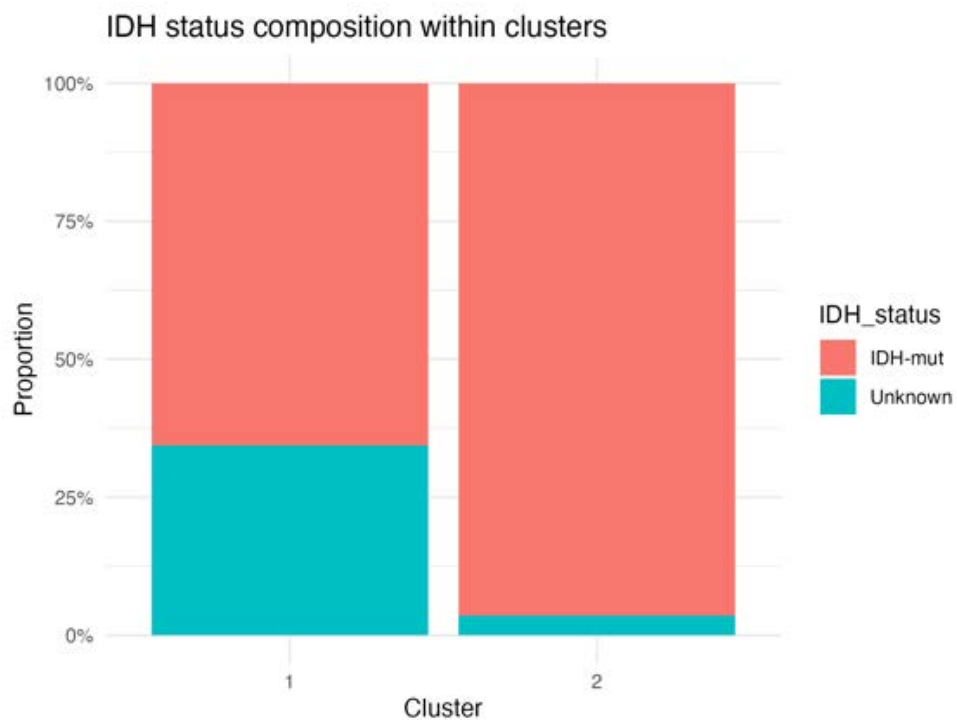


Figure 4.2.3 Distribution of IDH mutation of the two molecular clusters.

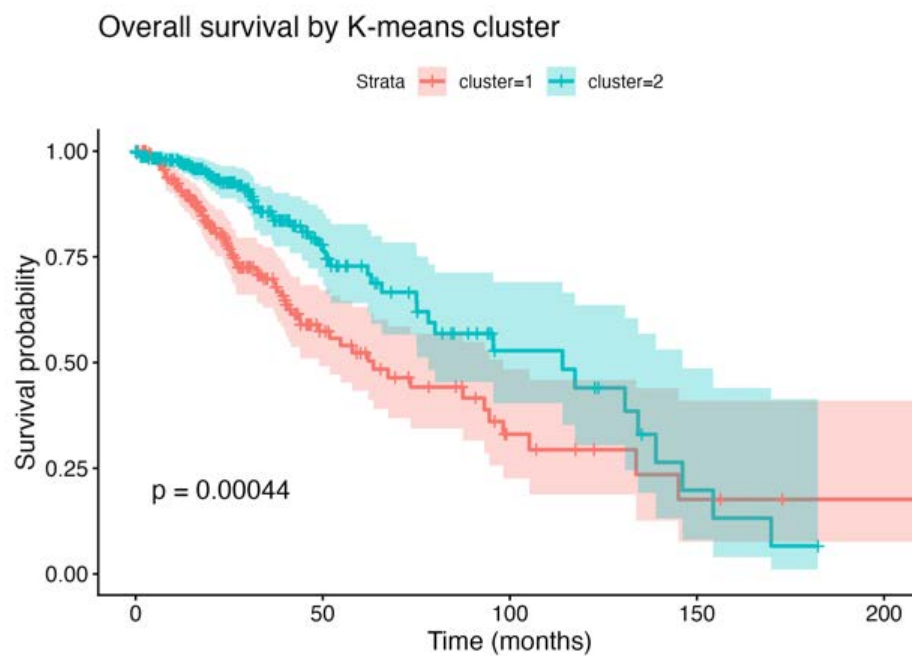


Figure 4.2.4 Kaplan–Meier overall survival curves based on the molecular clusters.

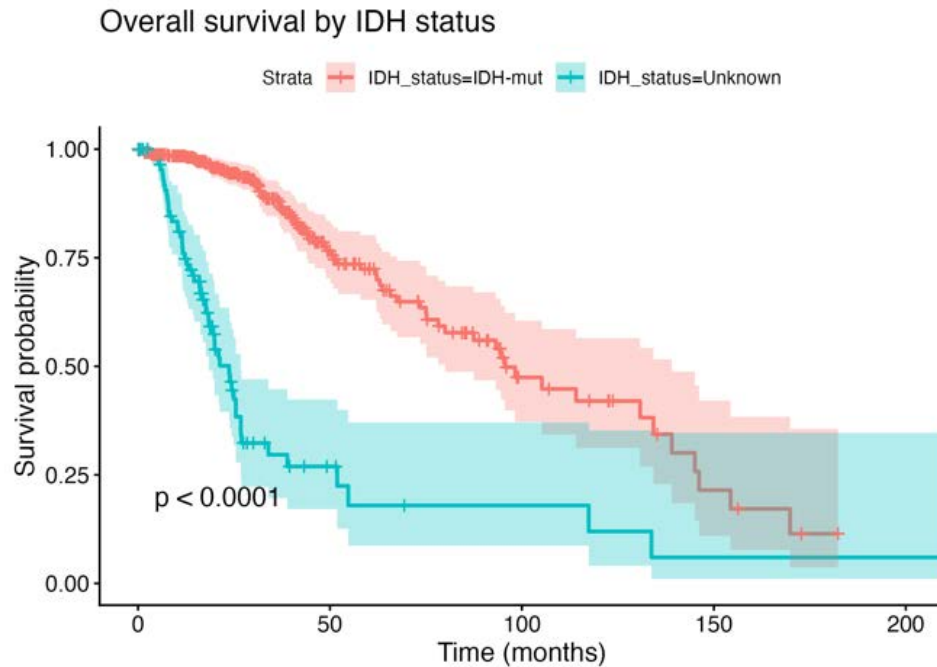


Figure 4.2.5 Kaplan–Meier overall survival curves based on the IDH mutation status.

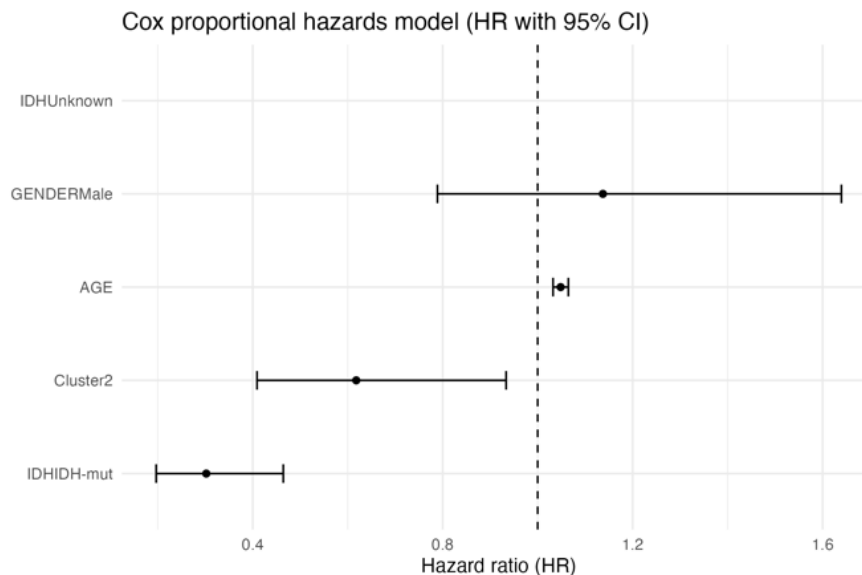


Figure 4.2.6 Multivariable Cox proportional hazards regression analysis of overall survival. Hazard ratios (HR) and 95% confidence intervals are shown for IDH mutation status and molecular clusters.

This research uses PCA and UMAP to reduce the dimensionality of the variables in the CNA and mRNA data, and to visualize the distribution of the samples. K-means algorithm is used to implement unsupervised clustering and to define clusters after identifying structural and transcriptional differences. The silhouette score(s) equation is applied to justify whether the number of clusters ($k=2$) is at optimal, to ensure the effectiveness of clustering. A wide numerical distribution is shown of samples along

the PC1 axis, in the PCA projection diagram (figure 4.2.1), which suggest obvious separations. Cluster 2 mainly locates along the positive PC1 direction and forms a compact structure. However, cluster 1 is more dispersed. According to the diagram, an aggregation of IDH unknown samples on the negative PC1 axis is observed, indicating a difference in their molecular features from the IDH-mutant group.

UMAP diagram (figure 4.2.2) reveals a clearer local

structure. A clear local structure is revealed in the UMAP diagram (figure 4.2.2). Both of the clusters are relatively compact, but cluster 1 involves fragmentations with scattered outliers comparing with the tightly-defined manifold of cluster 2. Cluster 1 shows small fragmentations with irregular scattered outliers comparing with the tightly defined manifold of cluster 2. However, overall, both clusters are relatively compact and shows a clear segregation. This pattern suggests that there might be a higher intra-heterogeneity in cluster 1.

The “IDH status composition within cluster” graph (figure 4.2.3) shows that cluster 2 is mainly composed of IDH-mut. Whereas in cluster 1, a higher percentage of unknown samples is found. This indicates that k-means clustering could reflect the IDH-driven molecular axis to an extent, but not perfectly replicate the IDH status.

Integrating the results from the three figures, a result could be drawn, that clusters with a higher proportion of IDH-mut have a higher intra-heterogeneity. This has also been validated in previous research, “IDH-mutant gliomas show prognostic heterogeneity despite favorable overall outcomes” [14].

The overall survival curve by cluster diagram (figure 4.2.4) shows a slight survival separation between clusters, $p=0.00044$, where cluster 2 showing improved long-term

survival, especially between 50-100 months. The overall survival curve by IDH status diagram (figure 4.2.5) shows an even stronger survival separation $p<0.0001$. This also correspond with the established prognostic value of IDH mutation. The difference between the two survival curves suggests that IDH is an essential and decisive biomarker for prognosis, which is consistent with the prior literature written by Yan et.al (2009), that IDH mutation is a dominant prognostic biomarker, and there is a stronger survival separation in IDH-based grouping which then supports WHO CNS5 integrated classification framework [5]. However, other factors should be considered to improve the accuracy.

In addition, the multivariable Cox modelling (figure 4.2.6) shows that IDH-mut is a strong protective factor, could significantly lower the death risk ($HR=0.35$). Similarly, cluster 2 also has a protective trend ($HR=0.5$), but less influential than IDH-mut, suggesting that clustering captured prognostic information is not fully explained by IDH alone. Likewise, in the research done by Zhirui et.al in 2022, a Cox analysis is carried out and demonstrates that “the risk score was an independent prognostic factor for overall survival in patients with IDH1-mutant glioma.”

4.3 Sub clustering within IDH-mut Tumors

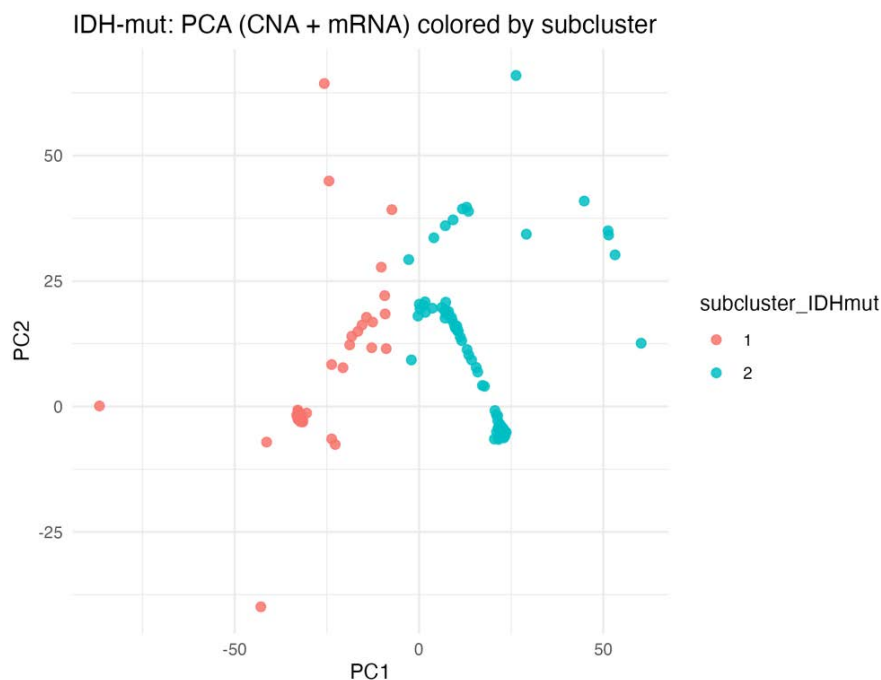


Figure 4.3.1 PCA projection of IDH-mutant tumors based on integrated CNA and mRNA features.

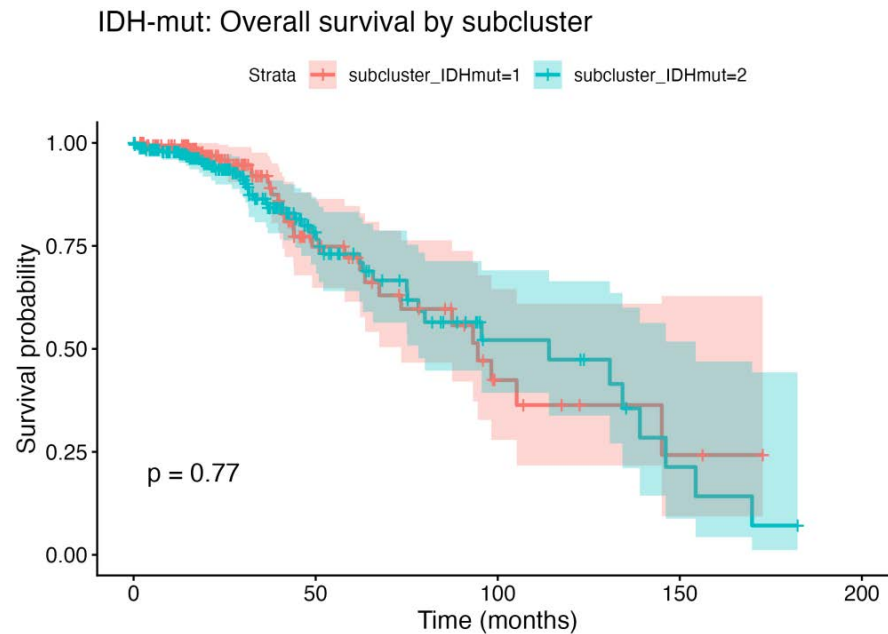


Figure 4.3.2 Kaplan–Meier overall survival curves based on the IDH-mutant subclusters.

Within the IDH-mut cohort, as visualized in the PCA projection, unsupervised clustering based on integrated CNA and mRNA profiles identified two distinct subclusters. (Figure 4.3.1) Subcluster 1 occupies a more dispersed region along the negative PC1 axis, whereas subcluster 2 forms a relatively condensed structure along the positive PC1 direction. This distribution indicates that even within the IDH-defined molecular class, multi-omics variation remains observable and is highly structured rather than random. This observation aligns with the results in the TCGA integrative analysis, which demonstrated that lower-grade gliomas are not biologically homogeneous despite sharing IDH mutation status. The TCGA Research Network reported that “LGGs segregate into biologically discrete subgroups defined by IDH mutation, 1p/19q codeletion, and additional genomic features” [8]. In particular, the study underlines that beyond the primary classification axis, IDH-mut tumors can still show internal genomic diversity.

However, there is no significant difference of the overall survival between the two clusters ($p=0.77$), by comparing trends on the KM curves. This result aligns with Eck-

el-Passow et al.’s findings [9]. It concluded that IDH mutation is the principal determinant of glioma cancer and prognosis, stating “IDH mutation status defines clinically relevant glioma groups more robustly than histology.” Meaning that internal molecular variation within the IDH-mut category does not have a definite translation into survival differences. Similarly, Ceccarelli et al. noted that although molecular profiling reveals additional substructures within IDH-mut gliomas, these differences often represent biological heterogeneity rather than prognostically distinct features unless accompanied by major genomic alterations such as copy-number instability [3]. This supports the present finding that clustering captures meaningful molecular diversity but does not independently stratify survival risk.

Together, these results suggest that although IDH-mut gliomas can be further subdivided at the molecular level, such subclusters may reflect intra-group heterogeneity, but clinically distinct prognostic states are not clearly identified.

4.4 Molecular Characteristics of IDH-mut Sub-clusters

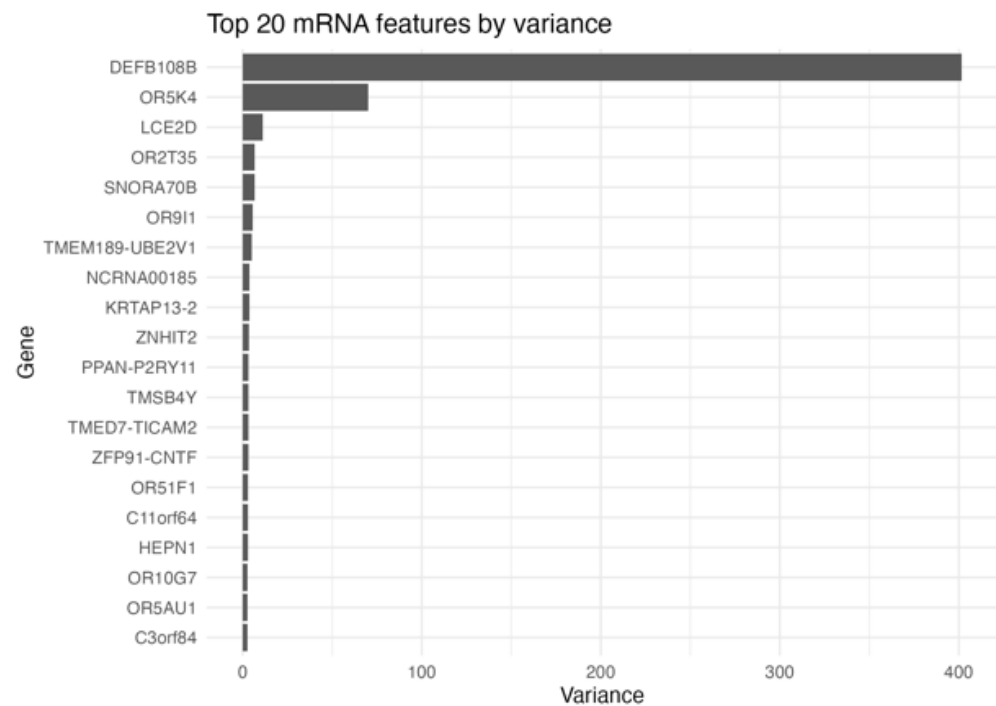


Figure 4.4.1 Variance of mRNA expression differences between IDH-mutant subclusters.

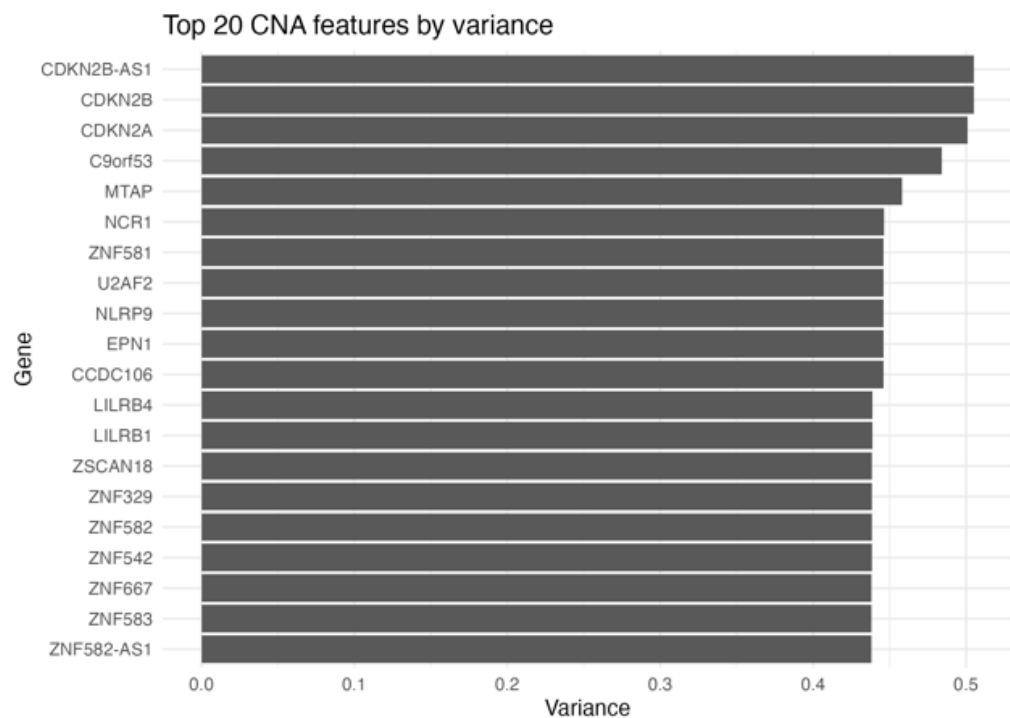


Figure 4.4.2 Variance of copy number alteration (CNA) differences between IDH-mutant subclusters.

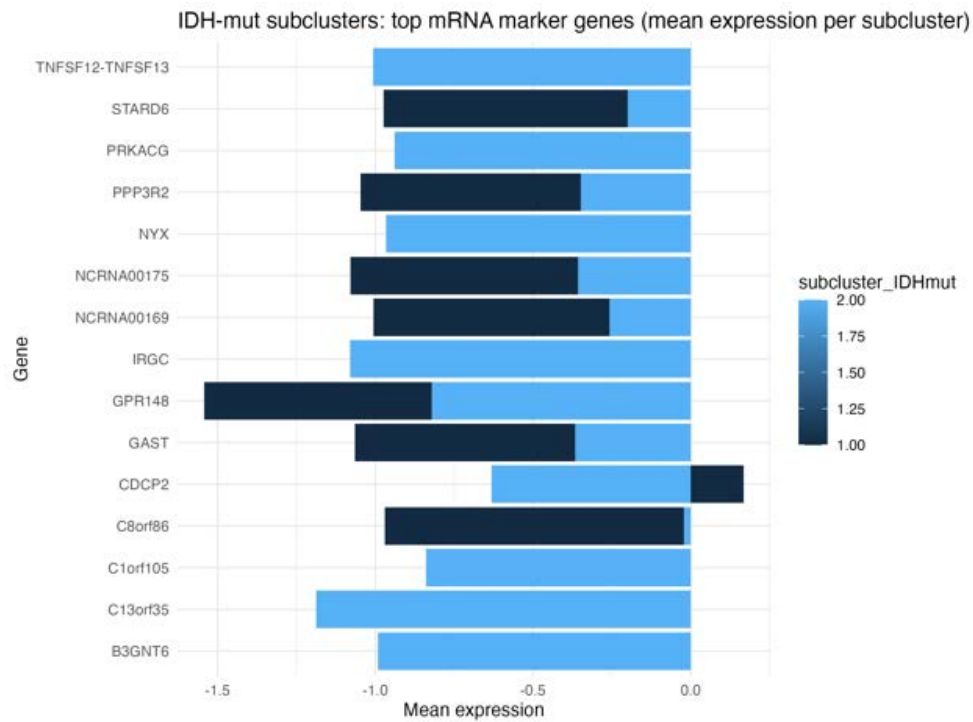


Figure 4.4.3 Comparison of gene expression levels between IDH-mutant subclusters for selected marker genes.

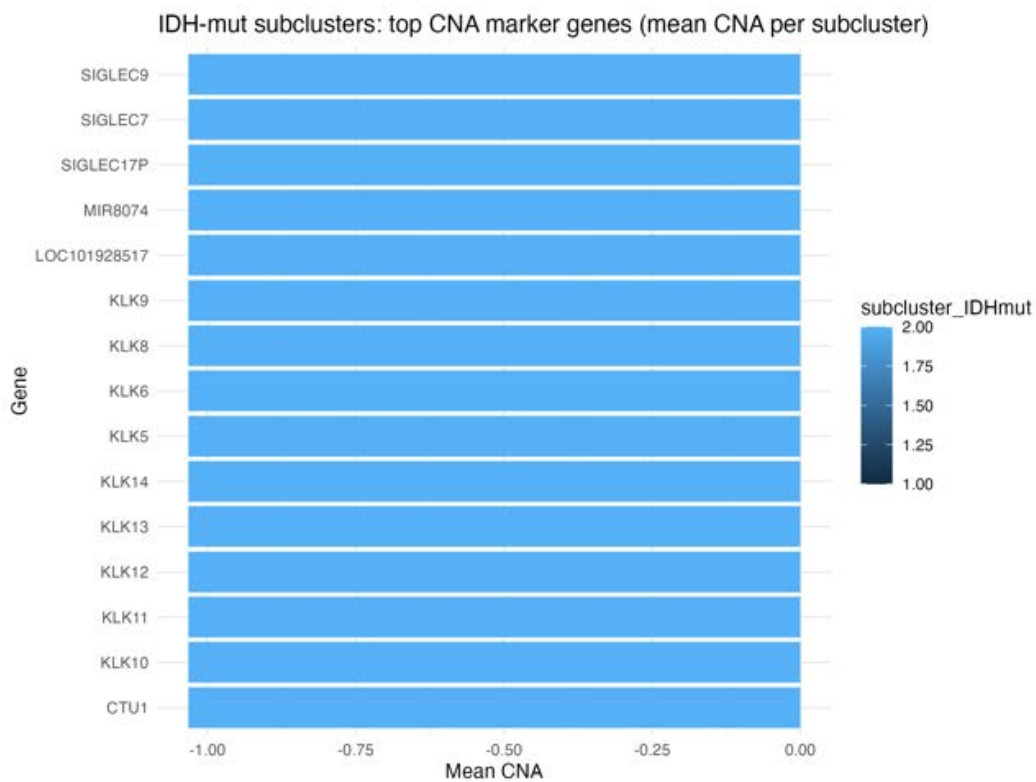


Figure 4.4.4 Comparison of CNA marker genes between IDH-mutant subclusters.

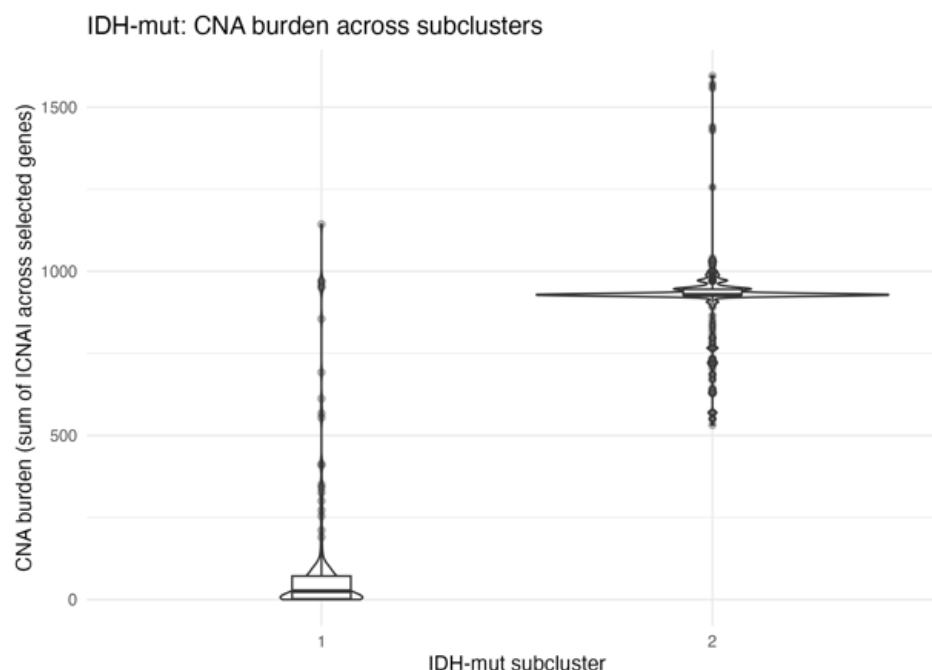


Figure 4.4.5 Distribution of overall CNA burden across IDH-mutant subclusters.

To further explore the intra-heterogeneity among IDH-mut samples, this research compares the differences between the mRNA and CNA data between the two IDH-mut clusters.

Variance analysis of mRNA expression comparisons (figure 4.4.1) shows that the differences in expression is caused by minority of the genes. DEFB108B has an observable higher variance (about >400), followed by OR5K4 (about 60-70), while majority of the genes display a minimal variability near the baseline. This indicates that the transcriptional differences within the IDH-mut groups are concentrated instead of broadly distributed. In contrast, the variance analysis of CNA (figure 4.4.2) shows a more even distribution among multiple genes. Genes including CDKN2A, CDKN2B, CDKN2B-AS1 and MTAP demonstrate a relatively higher variability (about 0.45-0.50). The difference in variation between mRNA data and CNA data indicates that structural alterations represent a more valid source of heterogeneity compared to transcriptional variability. Consistent with previous studies, they both suggest that copy number alterations provide additional evidence for heterogeneity beyond mutation status. For example, Ceccarelli et al. noted that “somatic copy-number alterations contribute substantially to the molecular diversity of lower-grade gliomas beyond IDH status.” [3].

On the gene expression level (figure 4.4.3), subtype-specific differences are shown by the distribution. GPR148, PPP3R2, GAST, and CDC2P, shows contrasting expression trends between the two subclusters. In subcluster 1,

these genes were markedly downregulated, but in subcluster 2, they remained closer to baseline levels. From this, a certain pattern of transcriptional activity variation is shown between the two subtypes instead of random. However, this result differs from the statements in previous literature. Cancer Genome Atlas Research Network emphasized the relative transcriptional stability of IDH-mut gliomas compared with IDH-wild tumors, which creates controversy [8].

On the structural level, consistent directional differences is revealed in CNA marker comparison diagram (figure 4.4.4). Members of the SIGLEC family and multiple KLK genes showed persistent deletion patterns in subcluster 2, with mean CNA values approaching -1. In contrast, subcluster 1 has a relatively neutral CNA value. This indicates that structural divergence is not only quantitative but also directional. Previous work has shown that IDH-mut tumors often maintain relatively stable genomic profiles [15]. However, the consistent deletion tendency observed here reveals that, a structurally distinct subtype may be presence within IDH-mut group. This finding contrasts with the traditional view that structural instability is primarily associated with IDH-wild tumors.

Comparison of overall CNA burden (figure 4.4.5) further demonstrates subtype separation. Subcluster 2 exhibits substantially higher CNA burden, with median values approximately 900–1000, and several samples exceeding 1500. However, subcluster 1 showed a markedly lower burden, with median values close to zero and only a limited number of high-value anomalies. Existing literature of-

ten associates CNA burden with poor prognosis in glioma [16], but this relationship has rarely been explored within IDH-mut tumors.

Taken together, these findings indicate that the two IDH-mut subclusters differ both at the transcriptional level and in the extent of genomic structural alteration. Subcluster 2 appears to represent a status characterized by higher

structural disruption accompanied by distinct expression patterns. While existing classification frameworks see IDH mutation as a dominant biomarker, the findings here suggest that, structural divergence within IDH-mut tumors may represent an additional prognostic perspective.

4.5 Clinical Relevance of IDH-mut Subclusters

TableA_subcluster_clinical_interpretation

subcluster	n	median_OS_months	event_rate	mean_age	grade3plus_rate	median_score	risk_label	dominant_regimen_level	dominant_regimen_short
1	166	26.2	0.175	37	NA	0	Lower-risk (better median OS)	De-escalation / surveillance-leaning	Surgery ± surveillance / delayed adjuvant
2	245	24.9	0.18	43.8	NA	2	Higher-risk (worse median OS)	Standard intensity	Surgery + RT + chemo (combined)

Figure 4.5.1 Risk evaluation of IDH-mutant subclusters based on overall survival and clinical indicators.

TableB_subcluster_regimen_distribution

subcluster	n	De-escalation / surveillance-leaning	Standard intensity
1	107	0.645	0
1	59	0	0.355
2	245	0	1

Figure 4.5.2 Inferred treatment stratification of IDH-mut subclusters

TableC_subcluster_top_regimen_short

subcluster	regimen_short	n	prop
1	Surgery ± surveillance / delayed adjuvant	107	0.645
1	Surgery + RT + chemo (combined)	59	0.355
2	Surgery + RT + chemo (combined)	245	1

Figure 4.5.3 Distribution of treatment strategies of IDH-mut subclusters

Subcluster-level clinical correlation reveals meaningful differences in inferred management strategies (figure 4.5.1). Subcluster 1 is classified as a lower-risk group because it demonstrates a slightly longer median overall survival (26.2 vs 24.9 months), whereas subcluster 2 is classified as a higher-risk group, based on poorer survival outcomes. This difference aligns with the different treatment intensity patterns. Subcluster 1 is mainly associated with de-escalation or surveillance-leaning approaches, while Subcluster 2 is consistently mapped to standard combined therapy strategies (figure 4.5.2). Distribution analysis is done to validate this separation (figure 4.5.3). Approximately 64.5% of patients in Subcluster 1 are assigned to surgery+ surveillance or delayed adjuvant therapy, whereas 100% of Subcluster 2 patients align with surgery+ radiotherapy and chemotherapy. These findings suggest that molecular heterogeneity within IDH-mut tumors may influence not only biological behavior but also

potential therapeutic strategies. In prior integrative genomic studies, similar results have also been demonstrated, suggesting that IDH-mut gliomas are not biologically homogeneous and different levels of treatment regimens should be implemented. The TCGA consortium reported that “integrated genomic analysis revealed clinically relevant subtypes within lower-grade gliomas” [8]. Likewise, Eckel-Passow et al. showed that combining IDH status with additional molecular features could stratify gliomas into groups with distinct prognostic outcomes [9]. The 2021 WHO classification also supports this paradigm shift by emphasizing that glioma diagnosis have to incorporate molecular markers beyond histology alone [5].

These results indicate that even within IDH-mut populations, risk stratification is present. This supports the argument that IDH mutation marks an entry point into stratification rather than a complete classification system. According to what Verhaak et al. previously noted, molec-

ular clustering can capture biologically meaningful characteristics that are not discovered by single-gene status alone [12]. Overall, those findings demonstrate that using integrated clustering method is very effective in proving a framework of treatment stratification within IDH-mut gliomas.

5. Conclusion

This study demonstrated that the integrative multi-omics clustering could effectively reveal the intra-molecular heterogeneity within the IDH-mut low grade glioma. Using CNA and mRNA data, among all the LGG samples, two distinct clusters are found, differ by the proportion of IDH mutation status. And the survival differences between clusters based on structural data and IDH-status data suggest a possible heterogeneity within IDH-mut group. Secondary sub clustering is done based on only the samples with IDH mutation, resulting with two sub clusters. They vary on CNA burden and RNA marker gene variation, which gives clear evidence for the molecular heterogeneity. In addition, structural variance appears to be a stronger source of heterogeneity than transcriptional variation. Although, there is no significant difference between the two clusters regarding overall survival, clinical risk could be effectively stratified. Thus, different levels of treatment regimen could be inferred for improving prognosis. In conclusion, the research suggests that solely focusing on IDH mutation status or single gene mutations cannot give a comprehensive explanation of the biological and clinical variability of LGG. Integrative clustering of structural and transcriptional features provides a more precise way of identifying intra-molecular heterogeneity while suggesting potential approaches for early diagnosis and treatment stratifications.

Summary

This study investigated intra-group molecular heterogeneity in IDH-mutant lower-grade gliomas (LGG) through integrative multi-omics clustering. Clinical, mutation, copy number alteration (CNA), and messenger RNA (mRNA) data from the TCGA Lower Grade Glioma dataset were analyzed using PCA, UMAP, and K-means clustering. Two molecular clusters were identified across the overall LGG cohort, with significant differences in IDH mutation proportion and overall survival. Further sub-clustering within the IDH-mutant group revealed two subtypes differing in CNA burden, mRNA expression, and gene markers. Although overall survival did not significantly differ between subclusters, risk-based treatment stratification was inferred. These findings demonstrated that integrative multi-omics clustering effectively revealed intra-group heterogeneity, with structural (CNA) variation represent-

ing a more important source than transcriptional variation. This research provided new guidance for precise glioma classification and personalized clinical treatment. From the aspect of data processing, the raw data are all extracted from TCGA and cBioPortal public retrospective datasets, which involves selection biases, sample acquisition determined by sequencing accessibility that could not fully represent real LGG patient population diversity, epigenetics and tumor microenvironment data were excluded due to missing data, incomplete treatment-related variables also block direct evaluation of subcluster therapeutic response and limit accurate clinical interpretation only to indirect survival associations. From the aspect of methodology, PCA only reserved two principal components and UMAP merely retains partial data structure, both causing partial loss of raw data information; K-means clustering assumes spherical cluster structure and requires manual preset cluster quantity, bringing analytical subjectivity, and alternative clustering algorithms like consensus clustering or hierarchical clustering may generate divergent grouping results. Future research can supplement epigenomic, tumor microenvironment and complete clinical treatment data, adopt multiple clustering methods for cross-verification, to further validate and optimize the molecular subtype stratification system of IDH-mutant LGG.

References

- [1] Schwartzbaum J A, Fisher J L, Aldape K D, Wrensch M. Epidemiology and molecular pathology of glioma. *Nat Clin Pract Neurol*, 2006, 2(9): 494-503.
- [2] Ostrom Q T, Patil N, Cioffi G, et al. CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the united states in 2013-2017. *Neuro Oncol*, 2020, 22(12 Suppl 2): iv1-iv96. Erratum in: *Neuro Oncol*, 2022, 24(7): 1214.
- [3] Ceccarelli M, Barthel F P. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell*, 2016, 164(3): 550-563.
- [4] van den Bent M J. Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective. *Acta Neuropathologica*, 2010, 120: 297-304.
- [5] Louis D N, Perry A, Wesseling P, et al. The 2021 WHO classification of tumors of the central nervous system: a summary. *Neuro-Oncology*, 2021, 23(8).
- [6] Miller J J. Targeting IDH-mutant glioma. *Neurotherapeutics*, 2022, 19(6): 1724-1732.
- [7] Reingold M, O'Reilly E M, Varghese A M, et al. Association of ablative radiation therapy with survival among patients with inoperable pancreatic cancer. *JAMA Oncol*, 2021, 7(5): 735-738.
- [8] Cancer Genome Atlas Research Network. Comprehensive,

- integrative genomic analysis of diffuse lower-grade gliomas. *New England Journal of Medicine*, 2015, 372(26): 2481-2498.
- [9] Eckel-Passow J E, Lachance D H. Glioma groups based on 1p/19q, IDH, and TERT promoter mutations in tumors. *New England Journal of Medicine*, 2015, 372(26): 2499-2508.
- [10] Yan H, Parsons D W, Jin G, et al. IDH1 and IDH2 mutations in gliomas. *New England Journal of Medicine*, 2009, 360(8): 765-773.
- [11] van den Bent M J, Brandes A A, Taphoorn M J, et al. Adjuvant procarbazine, lomustine, and vincristine chemotherapy in newly diagnosed anaplastic oligodendroglioma: long-term follow-up of EORTC brain tumor group study 26951. *Journal of clinical oncology*, 2013, 31(3): 344-350.
- [12] Verhaak R G W, Hoadley K A, Purdom E, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, 2010, 17(1): 98-110.
- [13] Monti S. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 2003, 52(1-2): 91-118.
- [14] Han Mingzhi, Sun Yanfei, Li Xingang. Epitranscriptomic control of glioblastoma malignancy: NAT10/ac4C-driven LPL mRNA stabilization reprograms lipid metabolism. *Neuro-Oncology*, 2025, 27(Supplement_5): v435.
- [15] Cancer Genome Atlas Research Network, Brat D J, Verhaak R G, Aldape K D, Yung W K, Salama S R, et al. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *The New England journal of medicine*, 2015, 372(26): 2481-2498.
- [16] Suzuki H, Aoki K, Chiba K, et al. Mutational landscape and clonal architecture in grade II and III gliomas. *Nature genetics*, 2015, 47(5): 458-468.
- [17] The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 2008, 455(7216): 1061-1068.
- [18] van Dellen E, Douw L, Hillebrand A, et al. MEG network differences between low- and high-grade glioma related to epilepsy and cognition. *PLOS ONE*, 2012, 7(11): e50122.
- [19] Zeng Z, Hu C, Ruan W, Zhang J, et al. A specific immune signature for predicting the prognosis of glioma patients with IDH1-mutation and guiding immune checkpoint blockade therapy. *Front Immunol*, 2022, 13: 1001381.
- [20] Meng L, Wu X, Su K, et al. Distinguishing IDH-mutant astrocytomas from IDH-wildtype glioblastomas using qualitative and quantitative MRI features: a WHO CNS5 study. *BMC Medical Imaging*, 2025, 25(1).
- [21] Miller J, Filipe L, McBrayer S K, et al. Isocitrate dehydrogenase (IDH) mutant gliomas: a society for neuro-oncology (SNO) consensus review on diagnosis, management, and future directions. *Neuro-Oncology*, 2022, 25(1): 4-25.
- [22] Evans L, Trinder S, Dodgshun A. IDH-mutant gliomas in children and adolescents - from biology to clinical trials. *Frontiers in Oncology*, 2024, 14.
- [23] Farahani S, Hejazi M, Tabassum M. Diagnostic performance of deep learning for predicting glioma isocitrate dehydrogenase and 1p/19q co-deletion in MRI: a systematic review and meta-analysis. *European Radiology*, 2025.
- [24] Han S, Liu Y, Cai S J. IDH mutation in glioma: molecular mechanisms and potential therapeutic targets. *British Journal of Cancer*, 2020, 122(11): 1580-1589.
- [25] Kobooldt D C, Fulton R S, McLellan M D, et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 2012, 490(7418): 61-70.
- [26] Li W, Zhou Y, Shafi S, et al. A rare IDH-mutant glioma case: insights from single-cell transcriptomic analysis of tumor recurrences. *World Journal of Surgical Oncology*, 2025, 23(1): 259.
- [27] Buda M, Saha A, Mazurowski M A. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Computers in Biology and Medicine*, 2019, 109: 218-225.