nSEA: n-Node Subnetwork Enumeration Algorithm Identifies Lower Grade Glioma Subtypes with Altered Subnetworks and Distinct Prognostics

Zhihan Zhang1, Christiana Wang1, Ziyin Zhao1, Ziyue Yi4, Arda Durmaz1,2, Jennifer S. Yu2, and Gurkan Bebek1,3†

1 Systems Biology and Bioinformatics Graduate Program, Case Western Reserve University, Cleveland OH 44106, USA
2 Center for Cancer Stem Cell Research, Cleveland Clinic Lerner Research Institute, Cleveland OH 44195, USA
3 Center for Proteomics and Bioinformatics, Department of Nutrition, Department of Computer and Data Sciences, Case Western Reserve University, Cleveland OH 44106, USA
ziihan.zhang@case.edu, christiana.wang@case.edu, ziyin.zhao@case.edu, ziyue.yi@case.edu, arda.durmaz@case.edu, yuj2@ccf.org,
† Corresponding Author: gurkan.bebek@case.edu

Abstract. Advances in molecular characterization have reshaped our understanding of low-grade glioma (LGG) subtypes, emphasizing the need for comprehensive classification beyond histology. Leveraging this, we present a novel approach, network-based Subnetwork Enumeration, and Analysis (nSEA), to identify distinct LGG patient groups based on dysregulated molecular pathways. Using gene expression profiles from 516 patients and a protein-protein interaction network we generated 25 million subnetworks. Through our unsupervised bottom-up approach, we selected 92 subnetworks that categorized LGG patients into five groups. Notably, a new LGG patient group with a lack of mutations in EGFR, NF1, and PTEN emerged as a previously unidentified patient subgroup with unique clinical features and subnetwork states. Validation of the patient groups on an independent dataset demonstrated the robustness of our approach and revealed consistent survival traits across different patient populations. This study offers a comprehensive molecular classification of LGG, providing insights beyond traditional genetic markers. By integrating network analysis with patient clustering, we unveil a previously overlooked patient subgroup with potential implications for prognosis and treatment strategies. Our approach sheds light on the synergistic nature of driver genes and highlights the biological relevance of the identified subnetworks. With broad implications for glioma research, our findings pave the way for further investigations into the mechanistic underpinnings of LGG subtypes and their clinical relevance.

Availability: Source code and supplementary data are available at https://github.com/bebeklab/nSEA

Keywords: Cancer Systems Biology · Network Analysis · Protein-protein Interaction Networks.

1 Introduction

Lower-grade gliomas (LGG) are brain neoplasms classified into 3 grades by the World Health Organization (WHO), where grades 2 and 3 present an infiltrative phenotype. While some LGGs remain stable, others progress to grade 4 gliomas (grade 4 astrocytoma [IDH-mutant tumors] and glioblastoma [IDH-wildtype tumors]), resulting in survival ranges between 1 and 15 years. Common treatment options include resection, chemotherapy, and radiation therapy. Based on the origin of glial cells, LGG can be classified into two subtypes: astrocytomas and oligodendrogliomas. Molecular features are also associated with clinical outcomes; for example, LGG with both an IDH mutation (IDH1 or IDH2) and deletion of chromosome arms 1p and 19q (1p/19q codeletion) show a better response to radiochemotherapy and are associated with longer...
survival. However, neither grade-based stratification nor molecular features can fully capture the complex architecture of LGG.

Gliomas are histopathologically classified into four grades associated with a worse prognosis. While this classification has prognostic value, investigating the complex molecular alterations within gliomas can lead to a better understanding of the biology behind the tumor types. For instance, some low-grade gliomas behave like malignant glioblastoma, while others have a favorable outcome similar to low-grade gliomas. Identifying genetic and epigenetic alterations in these tumors can reveal biomarkers with both prognostic value and the potential to guide therapeutic decisions [1].

Recently, studies by The Cancer Genome Atlas (TCGA) on lower-grade diffuse gliomas defined disease classification based on genetic and epigenetic alterations, providing biological justification for the utility of these features over histologic ones. Integrated genome-wide data analysis from multiple platforms delineated three molecular classes of lower-grade gliomas that were more concordant with IDH, 1p/19q, and TP53 status than with histologic classes [2].

In recent years, various approaches have been proposed for finding disease-related sub-networks [3–7] or predicting disease-causing genes [8–11] from large knowledge bases, such as protein-protein interaction (PPI) networks or signaling pathway databases. Most of these methods integrate systems-level measurements of gene and/or protein expression to prioritize networks [12–17]. A scoring function is combined with a search strategy to evaluate identified sub-networks. However, since finding sub-networks is an NP-hard problem [12], long run times and sub-optimal solutions are major drawbacks of these applications. Among all applicable methods, Kernel clustering, modularity optimization, random-walk-based, and local network search methods outperform others [6]. While some of these approaches can identify prognostic modules or disease-relevant pathways [12, 18, 6], they lack the ability to prioritize modules for disease subtype identification and subsequent survival analyses.

Enrichment-based pathway analyses are also commonly used to identify biological functions related to biomarkers and study disease subtypes in cancer [19–21]. However, since such approaches depend on previously selected genes, these analyses may lead to biased results. For instance, Sanchez-Vega et al. [22] analyzed the mechanisms and patterns of somatic alterations in ten canonical pathways and mapped them to multiple tumor types to discover pan-cancer subtypes and link them to possible drug targets. This supervised approach easily captured known subtypes with known disease pathways. In contrast, Durma et al. [23] reported an unsupervised approach that repeated this identification process using frequent subgraph mining with sampling and identified 106 clusters from 43K sub-networks mined from patient-specific networks. However, the former approach lacks the freedom to discover new subtypes, while the latter randomized approach requires careful filtering and repeated trials to arrive at robust discoveries.

In this paper, we introduce a novel network analysis algorithm known as the n-Node Subnetwork Enumeration Algorithm (nSEA). Our aim is to address challenges encountered by disease classification methods, which often rely on disease-associated genes or subnetworks for patient characterization and prognostics. Here, we discern robust patient subtypes based on functional variations in gene/protein expression within each sample and their interactions. This approach enables us to establish a patient classification framework that not only enhances prognostic accuracy but also elucidates the distinct pathway-level differences among patient subgroups. Such an approach holds the promise of improved prognostication for future patients, along with opportunities for enhanced treatment strategies and personalized interventions.

The (nSEA) algorithm takes a protein-protein interaction (PPI) network and system-level measurements of gene expression profiles as inputs. The goal of nSEA is to identify differentiating patterns among disease samples in an unsupervised manner. The algorithm is based on a bottom-up methodology in which a large sparse biological network (a PPI network filtered by patient gene expression profiles) is exhaustively enumerated and decomposed into n-node sub-networks (Figure 1A and 1B). These sub-networks are then evaluated, ranked, and filtered based on their inner-pattern consistency and network topology (Figure 1C).

In simple terms, the presented method aims to exhaustively identify n-node sub-networks that exhibit consistent expression patterns of network edges, quantified by the delta of gene expressions. The selected n-node sub-networks are expanded to include their neighboring nodes, forming more stable network structures (Figure 1D). By applying principal component analysis to network states, we identified sub-networks capable of discriminating disease states (Figure 2A-E) [24, 25]. The final set of sub-networks represents the major dynamics in the PPI network and provides a global picture of pathway dysfunction across cancer subtypes.
Fig. 1: Diagram of the nSEA algorithm. The algorithm takes a protein-protein interaction (PPI) network and gene expression profiles of samples as inputs. (A) The PPI network is converted into a sparse network. Edges are filtered based on the expression difference of their corresponding node pairs. (B) Network enumeration concept: All possible 4-node sub-networks are extracted from the original network, forming a list. Letters represent proteins. Three 4-node sub-networks and their positions in the list are annotated in colors as examples. (C) Feature selection based on the sub-network list. Sub-networks are ranked according to their inner-pattern consistency in a decreasing manner. They are then scanned and tested for topology (not shown in the diagram) from top to bottom. If a sub-network is selected into the feature set, it will exclude other sub-networks that share any node with it. (D) Selected sub-networks are expanded to neighboring nodes that share similar patterns, forming larger sub-networks. Solid lines represent edges at the current step, while dashed lines represent potential edges that can be added during expansion. Non-significant edges are omitted in this figure. (E) Specific application of nSEA to Lower grade gliomas (this study). Data is represented by a square and the process is represented by a "squircle." The basic properties of the data between each step were also annotated.

We applied nSEA to LGG samples and identified 5 latent groups/subtypes. We compared our subtypes with the current classification and identified significant sub-networks related to our clustering. We also explored the mutation, copy-number variation, and methylation features driving the force behind this classification and discussed several hypotheses based on these results. Furthermore, we compared our method with existing disease classification methods and validated our classification using an independent LGG cohort.

2 Methods

2.1 nSEA algorithm

The nSEA algorithm is based on a bottom-up methodology with which a large sparse biological network, \( G(V, E) \), is enumerated and decomposed into \( n \)-node subnetworks exhaustively. The goal of the algorithm is to identify subnetworks that can classify patients into subgroups and also provide distinctive biological states for each patient group based on these subnetworks. The first step is to create a network that is sparse enough for further processing. The PPI networks available today are too large for any enumeration algorithm to complete in a reasonable time. We create a sparse network to speed up the process while preserving relevance to disease classification by utilizing gene expression profiles. This is accomplished by using a protein-protein interaction (PPI) network and system-level measurements of gene expression profiles as inputs. Since the subnetwork vector we will calculate in the next steps represents the first principal component or the largest variance of the expression values within the subnetwork, edge filtration should also
facilitate achieving this (See a toy example of how this vector is generated in Section S1.1). Let \( e \in E \) and \( v \in V \) of the PPI network \( G \sim (V,E) \). We define an edge score \( S_{ae} \) between nodes (genes/proteins) \( v_i \) and \( v_j \) as:

\[
S_{ae} = \sigma(g_i - g_j), \quad e_k = (v_i, v_j), \quad i > j
\]

where \( \sigma \) is the standard deviation and \( g \) is the expression vector of the gene (Figure 2). Edge filtration was done by selecting the top 5\% edges ranked by the edge score \( S_{ae} \).

Enumeration was done by generating up to 4-node connected subnetworks from the filtered dataset. While larger \( n \) is possible to use, due to exponential increase in size, we only generated up to 4-node subnetworks only (See Section S1.2). Enumeration of all possible subnetworks was done to exhaustively identify and rank all possible subnetworks. To filter out insignificant subnetworks, the subnetwork score (inner-pattern consistency) of each \( n \)-node subnetwork was calculated:

\[
\Delta \phi_k = g_i - g_j, \quad e_k = (v_i, v_j), \quad i > j
\]

\[
S_{sub} = \sum_{x>y} \left| \frac{\text{cor}(\Delta \phi_x, \Delta \phi_y)}{\bar{e}} \right| \quad x \geq y
\]

where \( g \) denotes expression vector of node (gene) \( v_i \) and \( \Delta g \) denotes edge vector of edge \( e_k \). \text{cor} denotes Pearson correlation. \( \bar{e} \) denotes the total edge count in the subnetwork. \( S_{sub} \) denotes score for subnetwork. To avoid extreme cases when only one node has a degree larger than 1, 4-node subnetworks with an average degree less or equal to 0.75 were discarded. A threshold of the subnetwork score was set and all subnetworks with a score below the threshold were discarded.

Feature selection for the subnetwork list \( L \) was done using Algorithm 1. First, all subnetworks are ranked in descending order and placed in an array. While there are subnetworks in this array, the top network is saved as a feature and removed from the array. The feature network is then compared against the other subnetworks in the array. If any subnetwork has shared genes with the selected feature, it is removed from the array. The final set of subnetwork features is returned.

**Algorithm 1: Feature selection for \( n \)-node subnetworks**

**Data:** Set of subnetworks \( L \), scoring function \( S \)

**Result:** Feature Set \( F \), a set of subnetworks with unique nodes

\[
S \leftarrow \text{rank}(L, S) \quad // \text{rank subnetworks with score function } S \text{ from Eq. 3}
\]

\[
F \leftarrow \emptyset \quad // \text{Feature set is empty}
\]

while \( S \neq \emptyset \) do

\[
t \leftarrow \max(S) \quad // \text{first subnetwork in the ranked list is } t
\]

\[
S \leftarrow S - t
\]

foreach \( u \in S \) do

\[
// \text{check if any nodes (genes) are shared}
\]

if \( V(u) \cap V(t) \neq \emptyset \) then

\[
S \leftarrow S - u
\]

end

end

\[
F \leftarrow F \cup t \quad // \text{add } t \text{ to Feature set}
\]

end

For subnetwork expansion, nodes (genes) neighboring the subnetwork \( u \) were added to the subnetwork one by one (Algorithm 2). At each iteration for each neighboring node, we test:

\[
S(u) \geq S_T, \quad S(u) - S(j) \geq T, \quad \text{and} \quad |E(j)| - |E(u)| \geq a|E(u)|
\]

where \( S(u) \) denotes the subnetwork score at the expansion step. \( S_T \) denotes the minimum threshold for subnetwork score expansion, which is set to be 0.87. \( T \) is a threshold for the tolerance of score decrease. \( |E(u)| \) denotes the total number of edges in the subnetwork. \( a \) is a constant coefficient, where the set of
nodes in the network will not grow in size more than this ratio. $j$ is the network state assuming the node being considered is added to the subnetwork. The purpose of these two rules is to prevent the subnetwork from infinite expansion. If the rules are not satisfied, the expansion will stop. In this study, we set $T$ to 0.65 and $a$ to 0.25. We then select the neighboring node (gene) which has the largest score and add that node to the subnetwork. This process is repeated until no node can be added due to constraints.

### Algorithm 2: The subnetwork expansion algorithm

**Data:** Set of feature subnetworks $F$, where $u \in F$, and networks are scored by function $S$

$T_s$ denotes the minimum threshold constant for subnetwork score expansion (see Section 2.2)

$G$ is the protein-protein interaction network.

**Result:** Expanded subnetwork $u$

**foreach** $u \in F$ do

**foreach** $v' \in V(G), v \in V(u), (v, v') \in E(G)$ do

if $S(u) > S_T, S(u) - S(j) > T, |E(u)| - |E(u)| > a|E(u)|$ then

maxj < $S(j)$ then

if $S(u) > S_T, S(u) - S(j) > T, |E(u)| - |E(u)| > a|E(u)|$ then

$S_j$ = $S(u)$;

end

end

break;

end

$u = u \cup \{v'\}$

until $S(u) > S_T, S(u) - S(j) > T, |E(u)| - |E(u)| > a|E(u)|$

end

### 2.2 Parameter Tuning

The aforementioned values of parameters were determined by parameter tuning. These include the edge selection proportion ($a$), the low threshold of subnetwork score ($S_T$), and the number of clusters for patient clustering ($N_C$). First, $S_T$ and $N_C$ were tuned while $a$ was fixed to 5%. Two indicators were used to optimize $S_T$ and $N_C$. One was the clustering stability ($C_B$), and the other one was the distance from the background ($D_B$). $C_B$ is the mean of cluster-consensus values calculated by the ConsensusClusterPlus package. $D_B$ is defined as the distance from background clustering, the clustering result generated by setting $S_T$ to 0. Specifically, the distance is defined as:

$$D_B = 1 - F_M\text{index}(C_B, C_0)$$

where $C_B$ is the clustering labels from threshold $S_T$ and $C_0$ is the clustering labels when $S_T = 0$. Forden-Mallows index ($F_M\text{index}$) is a measurement of similarity between two clustering results [26]. By gradually increasing $S_T$, for each number of clusters ($k$), the relationship between $S_T$ and two indicators, $C_B$ and $D_B$, was explored (Figure S1A and S1B). Noticeably, $D_B$ increases with $S_T$, which indicates that the feature selection step is necessary in order to generate different clustering results from the background. For $C_B$, it is interesting that $C_B$ reaches its maximum value when $N_C$ is 5. We then further explored the relationship between $C_B$ and $D_B$ (Figure S1C). By considering both indicators, three $S_T$ values from $N_C = 5$ were very prominent. Among 0.83, 0.85, and 0.87, we chose 0.87 as the final $S_T$ value since when both $D_B$ and $C_B$ are similar, $C_B$ is a more important parameter than $D_B$.

Second, the proportion of edge selection ($a$) was evaluated. Due to the limitation of computation power, 5% is almost the maximum percentage of edges we can keep. We then gradually decreased $a$ to inspect its influence on patient clustering. By fixing $D_B$ and $C_B$ as mentioned above, $F_M$ indices between each clustering result caused by different $a$ values were calculated. In addition, we fixed $a$ to 5% but sampled its
Fig. 2: Subnetwork variables and their relationships A subnetwork consisting of 6 nodes and 8 edges. The subnetwork state, which represents the expression pattern of this subnetwork in sample 1, is colored according to gene expression levels. Expression matrix of the subnetwork in (A) with 10 samples. Expression values are centered and scaled. Edge vector is defined as the difference between expression vectors of the corresponding node pair. Edge A-D is used here as an example. The edge matrix combines all edge vectors from the subnetwork. The edge correlation matrix is calculated from the edge matrix. The lower triangle (diagonal excluded) of the matrix is used to calculate the Pattern Consistency score which is defined as the mean of the absolute values of the correlations. The subnetwork vector is defined as the first principal component of the expression matrix. It is used as the summary of the patterns of this subnetwork across all the samples. It is also used to cluster samples in the following steps.

2.3 Clustering of LGG patients and subnetworks
Subnetwork vector was calculated by the pcomp function from R package stats. Consensus clustering of patients and subnetworks were done with R package consensusplus. Clustering stability was defined as the mean of cluster-consensus values. Fowlkes-Mallows index was used to measure the distance of current clustering from the background. Consensus clustering of patients and subnetworks was done for 10,000 iterations with sampling proportion set to 0.75 and hierarchical clustering (Ward’s method). The self-organizing map was done using R som.

2.4 Clinical analysis and tree models
Survival difference (including p-value) was calculated by survdiff function from R package survival. Distances between patient groups and previous subtypes were defined as the mean Euclidean distance of all possible patient pairs from the two clusters. Correlation between subnetwork cluster vectors and telomere...
length or Karnofsky score was calculated with `cor.test` function with Spearman’s method and exact set to false. GO term (biological process) of subnetwork groups were annotated with `enrichgo` function from R package `clusterProfiler`. Mutation fold change was defined as the actual mutation count divided by the expected count.

Tree models were trained with `rpart` function from R package `rpart`. For binary classification of LG3, the parameter `minbucket` was set to 10, and parameter `maxdepth` was set to 2. For multi-label classification, `minbucket` was set to 22 to simplify the model and `maxdepth` was left as default (30).

Random forest model is trained with TCGA data using the subset function in R. The training process used 1000 trees and tried 8 variables at each split, while the importance of the predictor is set to be true.

Oncogenes and driver genes within each group were identified according to CCGD [27] and Uniprot [28] (Supplementary Table S4). Each subnetwork group was annotated by its corresponding activated oncogenes as well as the signs of the subnetwork vectors.

### 2.5 Comparison with existing methods

Clustering without gene selection and also nearest shrunken centroid-based gene selection [29] followed by network integration was used to compare with the nSEA approach. First, utilizing Consensus clustering, hierarchical clustering, principle component analysis, and k-means clustering we grouped patients and investigated the patient groups by running survival analysis and investigating clinical variables. Secondly, we trained a nearest shrunken centroid classifier. This widely used approach [30–33] is used to identify genes that stratify LGG samples. Subsequently, a protein-protein interaction (PPI) subnetwork was generated by overlaying the gene expression profiles with a network downloaded from STRING (Section 2.6), followed by node pruning and edge filtration. Networks were scored similar to nSEA approach as described in Section 2.1. PCA scores were subjected to various clustering techniques, including consensus clustering, K-means clustering, hierarchical clustering, and PCA, to classify individuals into multiple distinct classes. The Kaplan–Meier plots are generated based on the clustering results.

### 2.6 Data preparation

Gene expression data were downloaded from previously published studies by TCGA [34] and CGGA [35–37]. The TCGA datasets were generated by Illumina HiSeq 2000 platform. The level-3 expression data was obtained from UCSC Xena Portal [38]. Non-tumor samples were removed from the data resulting in data for 516 patients. Gene expression matrix was already log transformed. Genes were normalized using z-score normalization across all patients. Outliers were identified by `adjboxStats` from `robustbase` R package. The CGGA datasets were generated by Illumina HiSeq platform. The raw gene counts were downloaded from CGGA portal from the ‘mRNAseq 693’ dataset. CGGA data is log-transformed and normalized similar to the TCGA dataset. PPI data were downloaded from String PPI Database [39]. PPI network was filtered by eliminating edges with a combined evidence score of less than 0.7. The PPI network we downloaded had 13,562 nodes and 277,172 edges.

### 3 Results

#### 3.1 Subnetworks Classify LGG Samples into 5 Groups

We employed the n-Node Subnetwork Enumeration Algorithm (nSEA) to analyze LGG gene expression profiles [40], comprising 516 patients categorized as astrocytoma (33%), oligodendroglioma (34%), and oligoastrocytoma (22%). A protein-protein interaction (PPI) network was derived from the STRING database using a threshold of combined evidence score set to 0.7 [39], resulting in an undirected PPI network with 13,562 nodes and 277,172 edges (Figure 1E). A sparse network was constructed by retaining the top 5% edges based on edge vector deviation (Figure 1A, Figure 2C), yielding 5,681 nodes and 13,643 edges. The subnetwork size (\(n\)) was set to 4 for balance between robustness and computational efficiency, generating a total of 25,413,392 4-node subnetworks through subnetwork enumeration.

We investigated diverse properties of subnetwork feature sets to determine the optimal threshold for inner-pattern consistency in subnetwork selection. Decreasing the threshold led to an incremental rise in...
Patient Group
Tissue Source
LG1 Patient
Subnetwork Group
Subnetworks
LG1
LG2
LG3
LG4
LG5

Legend
B
LG1
LG2
LG3
LG4
LG5

C
SNG1: HCK/ABL1/RAF1
SNG2: VAV1/SPI1/CSF1R
SNG3: TPR/PCM1
SNG4: BRAF/PIK3CA/KRAS/NCOA1
SNG5: TCF3
SNG6: NFATC1/HDAC1/GSK3A
SNG7: AURKA
SNG8: TPM3

Fig. 3: Patient Groups and Subnetwork Clusters (A) Distance from background versus clustering stability from different inner-pattern consistency thresholds. 0.87 is highlighted in red. (B) Self-organizing map with 100 units. Patients were mapped to the units, with different shapes representing different patient groups. Units were also annotated with groups by majority voting. (C) Heatmap of subnetwork versus patients. LG1 patients were clustered into 5 groups (LG1~5) by consensus clustering using Euclidean distance. Subnetworks were clustered into 8 clusters by consensus clustering using absolute Pearson correlation distance. The sign of each subnetwork vector was adjusted to positively correlate with selected oncogenes or driver genes.

subnetwork inclusion in each feature set until saturation (Figure 3A). Clustering, based on subnetwork state matrices formed from the first principal component of subnetwork expression (Figure 2F), was then assessed for stability across thresholds. Interestingly, clustering stability peaked at both ends of the threshold curve for cluster numbers between 4 and 7 (Figure S1B), indicating distinct clustering patterns between high and low-threshold feature sets. Employing stability curves, we selected 5 clusters based on the relative change of cumulative distribution function (CDF) area (Figure S2E) [41].

Upon fixing the cluster number at 5, we applied the selection algorithm without a threshold to create a background for comparison against feature-based clustering (Figure S2C). The transition from background to high-threshold clustering was evident by a sharp increase around threshold 0.8. Examining the relationship between clustering stability and distance from the background revealed optimal thresholds (0.80 to 0.87) with high stability and separation (Figure 2A). Opting for 0.87 over 0.83 and 0.85, we selected a threshold conducive to subsequent steps.

Patient samples were clustered based on subnetwork state matrices derived from a final feature set of 92 subnetworks. Subnetwork sizes ranged from 6 to 11 nodes, predominantly comprising 6-node subnetworks (57%). Consensus clustering with Ward’s method (10,000

Fig. 4: The Kaplan Meier Plot shows the survival analysis for the TGGA patient groups based on TCGA prognostic networks. The $p-value < 4.1 \times 10^{-15}$ show that groups have distinct survival patterns.
iterations) generated a heatmap ordered by clustering dendrogram, revealing 5 patient groups exhibiting distinct subnetwork state patterns (Figure 3). Validation of the consensus clustering approach using unsupervised self-organizing map affirmed unbiased clustering (Figure 3B).

To annotate subnetworks, we performed consensus clustering on subnetwork vectors, identifying 8 subnetwork groups (SNG1–8). Genes within each group were divided into 2 clusters by correlations. Notably, SNG3 and SNG4 were enriched in cancer driver genes, with SNG4 housing 4 oncogenes associated with the p53 pathway. Protein classes and biological processes analysis further revealed significant associations with specific subnetwork groups, illuminating potential biological implications (Supplementary Table S2-S3).

Additionally, we explored the correlation between subnetwork vectors and clinical attributes like Karnofsky performance score and telomere length (Supplementary Table S6). Remarkably, SNG5 and SNG8 were significantly correlated with Karnofsky scores ($p$-value $< 8.5e-06$ and $p$-value $< 5.0e-03$, respectively). Further, gene cluster 2 of SNG5 contained driver genes linked to mental illnesses (Supplementary Table S7). Telomere length showed significant association with SNG3, SNG6, and SNG8 ($p$-value $< 0.021$), reinforcing links between chromatin remodeling and telomere regulation. Notably, NIPBL and KALRN emerged as promising gene candidates correlated with distinct patient subgroups, emphasizing their potential roles in promoter regulation and neuropathological disorders.

3.2 LG3: A Previously Unidentified Patient Group with Distinct Features

A comparison of our patient groups with TCGA subtypes and clusters demonstrated LG1-3’s alignment with known LGG subtypes. However, LG3 defied such classification, signifying a novel patient group unnoticed in prior TCGA studies (Table S5). Intriguingly, LG3 exhibited a unique clinical profile and subnetwork state pattern.

LG4 exhibited the highest proportion of grade-3 tumors and the oldest mean age (Figure S3A-B). LW2 displayed the worst Karnofsky performance score (Table S6). LG2 included relatively younger patients compared to LG1, LG3, and LG5. Telomere length analysis showcased pronounced shortening in LG4, consistent with previous research (Figure S3C) [42]. Notably, LG3 displayed a distinct advantage with the highest proportion of patients exhibiting high Karnofsky scores ($\geq 80$).
Survival analysis further underscored the significance of LG1, presenting improved survival compared to other groups, including LG1, LG2, and LG4, which mirrored IDHmut-codel, IDHmut-non-codel, and IDHwt subtypes (Figure 4). Decision tree modeling unveiled key subnetworks (SNG4 and SNG5) driving LG3’s unique clinical outcome (Figure S4).

Methylation analysis elucidated distinct genomic characteristics of LG3, marked by a scarcity of EGFR, NF1, and PTEN mutations, which could potentially contribute to its favorable prognosis. Additionally, supervised learning revealed methylation of NIPBL and KALRN as distinguishing features of LG3, offering novel insights into regulatory mechanisms and neuropathological associations.

3.3 Comparison with existing methods
First, we employed K-means clustering, hierarchical clustering, Principle Component Analysis and Consensus Clustering to determine subtypes of diseases based on mRNA gene expression profiles alone. While the groups had significant survival differences, the clusters did not follow any particular pattern and the number of genes was extremely high to discover any particular pattern from these analysis (Figure S5).

We also compared our method to sample classification from gene expression data by the method of nearest shrunken centroids [29]. We were able to stratify the samples into four distinct classes by utilizing sample differences based on correlation analysis. This classification informed the selection of an optimal gene inclusion threshold through a rigorous cross-validation procedure (PAMR package in R). Subsequently, we refined our original genomic matrix to incorporate only these curated genes. A tailored Protein-Protein Interaction (PPI) subnetwork was generated. This started with integrating the genomic expression matrix with the PPI network, followed by node pruning and edge filtration. High-correlation edges were selected using a stringent threshold to create subnetworks, revealing gene pairs with potential interconnected functionalities. While consensus-based clustering for both the PAMR-refined matrix and the PPI subnetwork yielded Kaplan Meier Plots with statistically significant survival differences, (Figure S6), the clusters had no discernible feature to study (Figure S7).

3.4 Validation of LGG Patient Groups
To ascertain the robustness of our patient groups, we validated our findings using an independent dataset, CGGA693. Through this validation, we verified the consistent clustering of LGG patients into LG1-5, confirming the existence and preservation of distinct subnetwork-based patient groups across different datasets and platforms. Further survival analysis validated the prognostic significance of these patient groups (Figure 6).

The subnetwork feature vectors from the TCGA dataset retained their ability to characterize the CGGA693 dataset (Figure 7), solidifying the robustness and generalizability of our approach. The relationship between TCGA groups (LG1-5) and CGGA groups further confirmed the concordance between these datasets. Importantly, the conserved survival traits of LG1-5 across datasets validated the clinical relevance of our patient groups, offering a promising avenue for refined LGG prognosis and treatment strategies.

4 Discussion
Many researchers have proposed subtypes of LGG over the last decade. Classification based on genetic features rather than histological features has been demonstrated to be more biologically relevant. The most widely accepted classification is based on molecular subtypes, which classify LGG patients into three clusters.
based on IDH mutation and chromosome 1p/19q co-deletion. However, recent studies have challenged this classification by suggesting that TERT may play an important role in glioma development. Despite the increasing specificity of LGG classification, the underlying mechanisms of these biomarkers remain unclear. For instance, patients with IDH wildtype genotype experience the worst survival outcomes. However, if they have both TERT and IDH mutations, their survival length is significantly extended, forming the best survival group. This suggests the existence of synergistic relationships among driver genes in LGG.

In this context, our developed algorithm, nSEA, offers insight into characterizing these tumors by capturing dysregulation within pathways. Unlike common bioinformatics approaches that focus on mutations, methylation, and copy-number variation, our approach exploits a different methodology. By scanning over nearly thirty million 4-node subnetworks, we provide a comprehensive view of subnetwork states within LGG. Through feature selection based on clustering statistics, we identify 92 subnetworks that categorize LGG patients into 5 groups. Three of these groups can be mapped to the general subtypes, demonstrating the ability of our algorithm to capture biologically significant signals. Additionally, we uncover one patient group, LG3, which not only exhibits distinct subnetwork states but also holds clinical significance. We further validate these patient subtype groups using a second cohort, showing that survival traits are conserved even across different patient populations.

Further analysis reveals that compared to other groups, LG3 demonstrates the best survival and Karnofsky performance score. The decision tree model trained on LG3 suggests that SNG4 and SNG5, enriched with oncogenes and associated with mental disorders respectively, can effectively distinguish LG3 from other patients with high accuracy. Mutation analysis indicates that LG3’s improved clinical performance may be attributed to the absence of mutations in EGFR, NFI, and PTEN. Moreover, a tree model based on methylation data highlights NIPBL and KALRN as two genes responsible for the primary and secondary splits of the tree respectively. Apart from their roles in transcription regulation through promoters, NIPBL has been linked to various types of cancers [43], suggesting its potential importance in gliomagenesis. The protein encoded by KALRN, Kalrin, belongs to the RhoGEP protein family, several members of which have been identified as cancer driver genes [44]. The Dbl-homologous domain of this protein could potentially become a target for future drug development [45].

The unsupervised nSEA approach also identified high percentages of cancer driver genes in each subnetwork group. These networks underscored the biological significance of the subnetworks captured by nSEA. The synergistic nature of driver genes has been extensively studied in the past, and nSEA networks provide insights into how driver genes synergistically contribute to tumor progression. Our findings offer valuable insights based on correlation analysis. However, it is imperative to establish causative relationships in order to gain a deeper understanding of each subtype. Driver mutations and epigenetic events warrant further investigation to delineate these causative relationships. While our approach involved feature selection to categorize patients into groups, numerous driver genes that could differentiate patient groups were identified. Any drivers not included could be further explored using nSEA networks to better understand their roles in gliomagenesis.
References


