# Machine learning for integrative multi-omics clustering and feature gene identification

Yang Zhang[1,2], Lin Liu[1,2,§], Lina Ma[1,2,3,*] and Zhang Zhang[1,2,3,*]

[1] *National Genomics Data Center, China National Center for Bioinformation, Beijing 100101, China*
[2] *Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China*
[3] *University of Chinese Academy of Sciences, Beijing 101408, China*
[§] *Present address: Beijing Forestry University, Beijing 100083, China*

[*]Corresponding authors: malina@big.ac.cn, zhangzhang@big.ac.cn

## Abstract

Multi-omics integrative analysis is pivotal for elucidating complex molecular mechanisms and biological processes, yet remains challenging in multi-omics data integration and feature selection. Here we present MIA, a machine learning framework for multi-omics integrative analysis that features unsupervised sample clustering via tensor decomposition and Fuzzy C-Means and supervised feature selection through an enhanced random forest. Benchmarking on simulated datasets demonstrates that MIA achieves higher accuracy in clustering and feature identification by comparison with extant algorithms. Furthermore, in TCGA empirical datasets, MIA is more effective to cluster samples and identify feature genes with significantly clinical outcomes. Particularly, as applied to glioblastoma, MIA is capable to identify three previously uncharacterized subtypes that exhibit more pronounced differences in survival patterns and powerful to uncover 167 feature genes that contribute to glioblastoma subtyping and associate closely with sensitivity of Temozolomide treatment. Collectively, these results establish MIA as a generalizable framework for multi-omics integrative analysis, enabling systematic molecular subtyping and significant feature discovery across complex biological systems.

**Keywords:** multi-omics integrative analysis; feature gene identification; tensor decomposition; FCM clustering; random forest

## Introduction

Powered by higher-throughput sequencing technologies, data from multiple omics layers, such as genomics, transcriptomics, proteomics, and epigenomics, has been generated at exponential rates and scales[1]. Integrative analysis of multi-omics data holds great promise for elucidating complex molecular mechanisms and biological processes[2, 3], particularly in the context of precision medicine[4], through precise sample clustering[5-7] and accurate identification of feature genes as potential molecular biomarkers[8, 9]. In machine learning and artificial intelligence, a feature is an individual characteristic of a data set and selection of informative and discriminative features is crucial for effective sample clustering. Hence, in biomedical research, a feature gene reflects a specific biological process and a set of feature genes is defined as those that contribute significantly to clustering. Therefore, multi-omics integrative analysis represents a powerful approach for systematically delineating feature genes from diverse omics data and thereby enabling a comprehensive understanding of complex physiopathological processes and mechanisms[10, 11].

Over the past several years, several algorithms have been developed for multi-omics data integrative analysis, which roughly fall into two groups in light of data integration strategy[12]: similarity network and consensus factor. For the former, representative algorithms include SNF[13], PINS[14], NEMO[15], and MANAclust[16]. They calculate sample similarities and fuse network across different omics types, but are sensitive to data perturbation easily affected by outliers and noise, accordingly reducing their robustness and increasing computational complexity. For the latter, representative algorithms are MOFA[17], iClusterPlus[18], and intNMF[19]. They derive shared factor matrices across multi-omics types but often rely on specific statistical assumptions that may not apply universally across diverse biological datasets, thus susceptible to model misspecification and noise influence. Importantly, among these existing algorithms, only SNF, MANAclust, and iClusterPlus are capable for feature gene identification as well as clustering. Yet, these three algorithms filter genes at individual omics level and are unable to systematically identify feature genes from multi-omics levels. Overall, the two strategies transform multi-omics data into two-dimensional representations, but fail to effectively integrate complementary information between different omics types, resulting in inaccurate multi-omics data analysis.

Nowadays, a promising strategy is tensor integration based on machine learning, which provides a powerful framework to effectively capture high-dimensional features inherent in multi-omics data[20]. Specifically, tensor integration allows for the representation of data in multi-dimensional arrays, enabling the simultaneous encapsulation of sample, gene and omics type, without any prior assumption[21-23]. In addition, tensor exhibits unique advantages in handling high-dimensional heterogeneity of biomedical data, flexibly adapting to variable sample sizes ranging from dozens to over thousands, and accommodating ultra-high feature dimensionality greater than 10,000 features per omics type[24]. Recently, tensor integration has been applied in sample classification[20, 25], survival prediction[26], and gene interaction prediction[27]. However, utilizing tensor integration for multi-omics data analysis remains unexplored.

Toward this end, here we present MIA, a **M**ulti-omics **I**ntegrative **A**nalysis algorithm for machine learning clustering and feature gene identification. Unlike extant algorithms that rely on the two-

dimensional data representation, MIA adopts a tensor framework for multi-omics data integration, structuring into a three-dimensional matrix comprising samples, genes and omics types, and utilizes tensor decomposition to reduce the matrix. It further uses Fuzzy C-means (FCM) for clustering to accurately capture sample similarity while minimizing the influence of initializing cluster centers. For feature gene identification, we develop a new enhanced random forest model (ERFM) for three-dimensional matrix, allowing for a comprehensive assessment of gene importance at multi-omics level. To assess the power of MIA, we conduct extensive comparative analysis using both simulated and empirical datasets, and apply in glioblastoma (GBM) subtyping and feature gene identification.

# Results

## MIA framework

We modularize the framework of MIA into three modules, namely, multi-omics integration, tensor decomposition & clustering, and feature gene identification (**Fig. 1**). For multi-omics integration, MIA utilizes a tensor framework to organize multi-omics data into a three-dimensional matrix comprising samples, genes, and omics types (**Fig. 1a**). It preserves critical biological information by maintaining data topology and facilitates extraction of joint features across omics types.

As biomedical samples are often unlabeled, MIA is equipped with a useful functionality for analyzing unlabeled data using tensor decomposition for dimensionality reduction. Among available techniques, Tucker decomposition is widely adopted, owing to its flexibility and accuracy in capturing complex data structures and interdependencies[22, 23]. Therefore, MIA adopts Tucker decomposition to transform the tensor into three matrices: sample-feature matrix, gene-feature matrix, and omics type-feature matrix, and employs FCM to cluster the sample-feature matrix with labeled information (**Fig. 1b** and **Supplementary Fig. 1**). Unlike *K*-means, FCM allows samples to belong to multiple clusters, providing more accurate characterization of ambiguous samples. As noted, there are two critical parameters for FCM—uncertain fuzzy index ($m$) and cluster count ($K$)—that are typically pre-defined. In MIA, they are optimally determined for each dataset; $m$ is estimated from sample and feature counts[28] and $K$ is estimated by maximizing the silhouette coefficient, thereby providing data-driven parameter estimation tailored to each dataset (see Methods for details; **Supplementary Figs. 2 and 3**).

MIA further utilizes Gini index to perform initial gene filtering and employs ERFM for feature gene identification. Unlike traditional random forest model (RFM) that handle only a single data type, ERFM is capable to take good account of multi-dimensional data types to quantify the contribution of each gene in clustering and thus to yield a set of feature genes (**Fig. 1c**). There are two significant improvements in ERFM: an enhanced tree structure and an iterative computational procedure (**Supplementary Fig. 4**). First, the enhanced tree structure accounts for multi-dimensional data types by creating multiple tree nodes based on different omics types for each gene, rather than a single tree node in RFM. The importance score of each gene is determined by the increase of classification error after simultaneous randomization of all corresponding nodes, which allows for

an effective assessment of the gene contribution within overall influence of multi-omics data. Second, instead of a single forest generation in RFM, ERFM employs an iterative computational procedure that repeatedly constructs and refines the forest, thus enhancing the consistency of the identified feature gene set. Specifically, after each iteration, genes with negative importance scores are eliminated, and the iterative refinement continues until there is no remaining gene with negative importance scores, ensuring that only the most relevant feature genes are retained for further analysis (**Supplementary Fig. 4c,d**; see details in Methods).

## Benchmarking of MIA on simulated datasets

To evaluate the performance of MIA, we consider multiple parameters, including cluster count ($K$), omics type count ($T$), sample count ($N$), and the numbers of real (true) features ($M_t$) and noise (false) features ($M_f$), and generate simulated datasets with different parameters' combinations in terms of possible scenarios of real data (see details in Methods). We examine the clustering accuracy of MIA by comparison against existing six popular algorithms—SNF, MANAclust, iClusterPlus, PINS, NEMO and MOFA—among which only the former three support feature gene identification as well.

When varying $K$ and $T$ with fixed $N$, $M_t$ and $M_f$, MIA overall achieves the stable higher accuracy across all parameters' combinations (**Fig. 2a**). Among these competing algorithms, NEMO and SNF present comparable performance, while iClusterPlus exhibits marked instability when varying $K$. Consistently, MIA performs much better than other algorithms at varied values of $N$, $M_t$, $M_f$ with fixed $K$ and $T$ (**Fig. 2b**). In contrast, MANAclust shows significant deviations from the true labels. Considering the significance of $K$ in data analysis (that is often arbitrarily set), we further evaluate their accuracy in optimal $K$ selection and reveal that MIA achieves an average accuracy of 0.987 across all datasets, outperforming other algorithms (**Fig. 2c**). Since MANAclust often overestimates the value of $K$, we exclude it from the following analysis in real datasets. Regarding feature gene identification, it is clearly shown that feature genes identified by MIA yield higher significance than those by the other algorithms (**Fig. 2d and Supplementary Table 2**).

## Benchmarking of MIA on TCGA datasets

We evaluate MIA in ten empirical real datasets from The Cancer Genome Atlas (TCGA), covering different omics data types (copy number variation (CNV), gene expression, and methylation) and involving ten cancer types: Bladder Cancer (BLCA), Cervical Squamous Cell Carcinoma (CESC), Breast Cancer (BRCA), Colon Cancer (COAD), GBM, Head and Neck Squamous Cell Carcinoma (HNSC), Lower Grade Glioma (LGG), Liver Cancer (LIHC), Lung Adenocarcinoma (LUAD), and Prostate Cancer (PRAD). To facilitate the comparison, we run MIA as well as existing related algorithms to perform clustering and survival analyses on each dataset.

In these ten datasets, MIA overall outperforms existing algorithms by achieving more significant survival outcomes across different omics data combinations (**Fig. 3a-c** and **Supplementary Table 3**), particularly for COAD and LUAD (**Supplementary Figs. 5 and 6**). Compared to all possible combinations of single-omics and two-omics data, clustering by MIA on three-omics data generally

146 demonstrates higher survival significance ($p$<0.05, Log-rank test) in nearly all cancer types, which
147 is more strikingly than other algorithms. Particularly, MIA achieves significantly different survivals
148 ($p$<0.0001, Log-rank test) in LIHC and GBM based on three-omics data, in contrast to single-omics
149 or two-omics combinations (**Supplementary Figs. 7 and 8**). Additionally, we observe that although
150 nearly all algorithms exhibit comparable survival differences in LGG on multi-omics combinations,
151 clusters determined by MIA show more divergences (**Supplementary Fig. 9**). Detailed survival
152 analyses for all these ten cancer types (**Supplementary Figs. 5-14**) show that multi-omics data
153 integrative analysis by MIA is superior to other algorithms in precise clustering.

154 Following a previous study[12], we adopt six parameters to clinically evaluate the performance of
155 MIA, viz., gender, age at diagnosis, progression of the tumor (pathologic T), metastases (pathologic
156 M) and cancer in lymph nodes (pathologic N), and the total progression (pathologic stage), which
157 can be used to indicate demographic characteristics, disease severity and progression. We
158 investigate the enrichment of these clinical parameters in all clusters determined by different
159 algorithms and find that MIA clustering achieves more significantly clinical outcomes across ten
160 cancer types (**Fig. 3d**, **Supplementary Fig. 15** and **Supplementary Table 5**). In BLCA, all the
161 algorithms reveal more enriched parameters in a consistent manner, possibly due to its distinct
162 clinical stages and molecular differences between non-muscle invasive and muscle-invasive types[29].
163 Notably, no clinical parameters are enriched in GBM across all algorithms, most likely owing to the
164 missing clinical data as well as high malignancy of GBM that is difficult to distinguish subtypes
165 based on tumor morphology and pathology. Given the clinical significance of these six parameters,
166 our results show that MIA clustering has a stronger correlation with disease progression than other
167 algorithms.

168 We further assess the MIA performance of feature gene identification in seven cancer types with
169 sufficient samples. In contrast to existing algorithms, feature genes identified by MIA present higher
170 significant difference between cancer and normal samples across nearly all examined cancer types
171 (**Fig. 3e**). Clearly, this result demonstrates the superior capability of MIA in feature gene
172 identification from multi-omics data integrative analysis.

## 173 GBM subtyping and feature gene identification

174 GBM is the most aggressive primary brain cancer, with a five-year survival rate of only ~5%[30].
175 Therefore, it is crucial to identify feature genes that can be used as potential biomarkers for GBM
176 diagnosis and prognosis. We apply MIA to the GBM dataset described above, covering 15,372
177 genes with multi-omics profiles (CNV, gene expression and methylation). For clustering, MIA
178 determines the optimal value of $K$ at 3 (**Fig. 4a**), as further supported by the well-separated clusters
179 (**Fig. 4b**). Most important, clinically, the clustering of $K$=3 yields more significant survival
180 differences ($p$-value<0.0001, Log-rank test; **Fig. 4c**), namely, clusters 1, 2 and 3 at high, medium
181 and low survivals, respectively. In contrast, classical subtypes based solely on gene expression[31]
182 show no significant survival differences ($p$-value=0.59, Log-rank test; **Supplementary Fig. 16**).
183 Therefore, our results show that MIA is more powerful in multi-omics integrative clustering.

184 It has been reported that the fraction of altered genome reflects the extent of genetic alterations in
185 cancer[32] and stemness indices (mRNA expression-based stemness index, mRNAsi and DNA
186 methylation-based stemness index, mDNAsi) signify the stem cell-like characteristics of cancer
187 cells[33], together indicating tumor aggressiveness and prognosis. Thus, we further evaluate the MIA
188 clustering in GBM based on these metrics. We find that the fraction of altered genome decreases
189 significantly from cluster 1 to cluster 3 ($p<0.05$; **Fig. 4d**), corresponding well with the declining
190 survival rates. Consistently, the stemness indices (mRNAsi and mDNAsi) gradually reduce, namely,
191 higher in cluster 1, medium in cluster 2, and lower in cluster 3 (**Fig. 4e** and **Supplementary Table**
192 **6**). These results indicate that GBM with fewer genomic alterations may exhibit more aggressive
193 behavior and lower survival rates, and GBM with poor prognosis has lower stemness and higher
194 cancer cell differentiation[33, 34]. Overall, the clinical analysis demonstrates that MIA is effective to
195 identify GBM subtypes with distinct aggressiveness and prognosis.

196 In addition, MIA is capable to identify a set of 167 feature genes from the GBM three-omics profiles
197 that contribute to the clustering (**Fig. 4f and Supplementary Table 7**). Among them, 87 (52.1%)
198 genes have already been documented to be associated with GBM, in which 70 are recognized as
199 GBM markers (**Fig. 5a**). Clusters 1, 2, and 3, built on these 167 feature genes, exhibit distinct
200 molecular characteristics, with representative reported markers as listed below; cluster 1 shows copy
201 number amplification in *LTBR* and *MICALCL* and increased expression in *FAH*, *SENP8*, and
202 *DIABLO*, cluster 2 exhibits increased expression in *HDAC9*, *LSAMP*, and *CHD6*, and cluster 3
203 presents increased expression in *VGF*, *TRPV4*, and *CCR2* and increased methylation in *PCDH8*,
204 *GPD1*, and *C1QB* (**Fig. 4f**). Furthermore, tissue enrichment and functional annotation analyses
205 reveal that the 167 feature genes are closely associated with brain tissues and crucial for
206 pathobiological processes, such as neuroblast proliferation, immune microenvironment dynamics,
207 and tumor-related pathways (**Supplementary Fig. 17a,b**).

208 We also use MIA to identify feature genes based on single-omics and two-omics data. Among these
209 167 feature genes, 96 (57.5%) and 43 (25.7%) genes are specifically identified based on three-omics
210 data, by comparison against single-omics and two-omics data, respectively (**Fig. 5b,c**). Functional
211 enrichment analysis shows that these feature genes specifically identified by three-omics data
212 exhibit stronger associations with brain tissue and tumor development (**Supplementary Fig. 17c**).
213 We further analyze the clustering power of feature genes specifically identified by different omics
214 combinations. We observe that feature genes from three-omics always show higher clustering scores,
215 clearly indicating that multi-omics integrative analysis achieves more enhanced clustering power
216 (**Fig. 5d,e**). Notably, several literature-reported GBM marker genes, such as *VGF* (promoting GBM
217 stem cell survival and tumor growth[35]) and *BRD8* (hindering *p53* transcriptional activation and
218 promoting GBM cell proliferation[36]), are successfully identified by three-omics data integrative
219 analysis, yet absent in single-omics or two-omics data analysis (**Supplementary Table 8**). It is
220 worth mentioning that based on three-omics integrative analysis, *LTBR* (activating anti-tumor
221 immune response in glioma[37]) and *GPD1* (whose inhibition prolongs GBM mouse survival[38]) are
222 computationally identified as feature genes with markedly enhanced priority (**Supplementary Fig.**
223 **18**), thus highlighting the great potential of MIA to uncover biologically significant features from

multi-omics data and thereby guiding in-depth experimental validation and mechanism investigation.

## Characterization of Temozolomide treatment sensitivity through GBM feature genes

Temozolomide (TMZ) is one of the most effective chemotherapy agents for GBM treatment[29]. Based on these 167 feature genes identified by MIA, we further explore whether these genes can be used to predict TMZ chemotherapy response. Toward this end, firstly, we collect a publicly available dataset (expression data only) with GBM TMZ treatment, containing samples of control, TMZ + radiotherapy, and TMZ + radiotherapy + inhibitor[39]. We use MIA to cluster these samples based on the 167 genes' expression profiles, yielding three clusters with distinct expression patterns. We find that the control samples are most enriched in cluster 1, the TMZ + radiotherapy samples are most in cluster 2, and the TMZ + radiotherapy + inhibitor samples are exclusively included in cluster 3 (**Fig. 6a** and **Supplementary Table 9**). Multi-class ROC analysis demonstrates the accuracy of these 167 genes to predict TMZ treatment, with an average AUC value of 0.85 (**Fig. 6b**). To validate this result, we collect three independent datasets containing TMZ-resistant and TMZ-sensitive samples. For these three datasets, clustering results align perfectly with the outcomes of TMZ treatment (**Fig. 6c-e** and **Supplementary Table 9**), validating the effectiveness of these 167 feature genes in TMZ chemotherapy response.

To systematically identify genes associated with TMZ chemotherapy sensitivity, we discover that 64 genes display consistent expression patterns in the above datasets. We collect these 64 genes' associations with chemotherapy drugs across different cancers (**Supplementary Table 10)**. Notably, ten genes (*GPD1*, *MSI1*, *IFFO1*, *NCBP2*, *SGTA*, *BCL9L*, *FOXP4*, *KCP*, *MAPK13*, and *PAFAH1B1*) have already been reported to be related to chemotherapy sensitivity (**Fig. 6f**). Specifically, *GPD1* is specifically expressed in brain tumor stem cells and involved in tumor initiation, metastasis, and recurrence; increased expression of *GPD1* correlates with TMZ resistance in GBM[38]. *MSI1* is a stem cell gene that controls the balance between self-renewal and differentiation; GBM lines with *MSI1* knockout displays increased sensitivity to cell cycle and DNA replication inhibitors[40]. The rest eight genes, albeit unexplored with GBM chemotherapy, associate closely with gynecological cancers[41-43], LIHC[44], gastrointestinal cancers[45, 46], hematological malignancies[47], and other cancers[48], presumably indicating their potential relevance in GBM treatment.

## Discussion

In this study, we presented MIA, a powerful framework for multi-omics integrative analysis capable of accurate sample clustering and feature gene identification. MIA utilizes various machine learning techniques, including Tucker decomposition, FCM, and ERFM, to effectively integrate and analyze multi-omics data and identify biologically important features contributing to clustering. Comparative analyses on simulated and real datasets revealed that MIA outperforms existing algorithms in both integrative clustering and feature determination. As testified in GBM multi-

omics data analysis, MIA is powerful to cluster samples into clinically relevant subtypes with distinct survival outcomes and identify feature genes associated with TMZ treatment sensitivity.

MIA adopts a machine learning-based strategy to achieve powerful multi-omics integrative analysis. Although deep learning has achieved remarkable progress in biomedical research, its reliance on large, well-annotated datasets limits its applicability to omics data, which are typically high-dimensional, unlabeled, and sample-limited. In contrast, the machine learning models used in MIA, such as random forest, are more robust to small sample size, less prone to overfitting, and offer interpretable results with biological insights. Despite this, technically, future explorations and improvements for MIA include the following. First, the tensor framework requires both matched samples and genes across omics types, limiting its applicability to data with heterogeneous features. Second, after Tucker decomposition, only sample-feature matrix is used for clustering, whereas gene-feature and omics type-feature matrices remain underutilized. Notably, clustering heatmaps from gene-feature matrix in simulated data clear separated real from noise features (**Supplementary Fig. 1b**), suggesting potential for feature filtering. Additionally, omics type-feature matrix may quantify contributions from different omics layers; in GBM multi-omics, for example, eigenvalues for CNV, gene expression, and methylation were 0.446, 0.674, and 0.589, respectively (**Supplementary Table 7**), indicating a stronger influence of gene expression, which aligns well with single-omics analyses showing its most relevance to survival (**Fig. 3c**). Thus, omics type-feature matrix contains substantial biological relevance and further exploration of its application for optimizing feature gene identification would be desirable.

Across the ten cancer types, clusters derived from CNV and methylation data generally showed smaller survival differences than those from gene expression (**Fig. 3c**). From the data perspective, this may stem from limited or/and uneven distribution of samples exhibiting heterogeneous copy number amplifications/deletions, and signal attenuation introduced during the conversion of methylation measurements from the site level to the gene level. From the biological perspective, gene expression is closely correlated with CNV and promoter methylation status[49], together exerting an influence on protein production, cell function, and thereby, phenotypic and clinical outcomes[50]. Furthermore, we observe that nearly all algorithms exhibit comparable survival differences in LGG on multiple omics combinations, primarily due to substantial molecular variations between grade II and grade III LGG, with the latter associated with a worse prognosis[51].

In GBM, we observed that the fraction of genome altered, a measure of the extent of copy number changes in the genome, decreased with declining survival rates (**Fig. 4d**). In early-stage tumors, studies have demonstrated that a higher fraction of genome altered is typically associated with poorer pathological characteristics and survival rates[52, 53], and is also a sign of poor prognosis[54, 55]. On the contrary, in advanced tumors, several studies have also found that tumors with lower genomic alterations can exhibit more aggressive behavior and poorer survival rates[32]. For example, in metastases compared to primary lung adenocarcinoma, the fraction of genome altered was significantly lower[56]. Additionally, excessive genomic instability may cumulate damage, exceeding the threshold of tumor cell activity, thereby inhibiting advanced tumor growth and metastasis[57]. Our observations in GBM are consistent with these findings of advanced tumors, but more research is needed to explore the association between this indicator and the prognosis of advanced tumors.

To sum up, with the vast volume of multi-omics data generated by higher-throughput sequencing technologies, multi-omics integrative analysis becomes more increasingly routine and popular. Thus, MIA bears great utility for providing multi-layered insights into complex biological and physiopathological processes and promoting precision diagnostics and targeted therapies.

## Methods

### Simulated datasets

Simulated multi-omics datasets were generated by a Python script. Several parameters—number of clusters ($K$), number of omics types ($T$), sample size ($N$), and number of real features ($M_t$), and number of noise features ($M_f$)—were considered and five replicates were produced for each parameter combination. Samples were assigned to $K$ clusters using a Gaussian probability distribution. Each omics layer was represented as an $N \times (M_t + M_f)$ matrix, where $M_t$ features were differentially expressed across clusters with group-specific means and Gaussian noise and $M_f$ features were randomly generated. The code for dataset simulation is provided in *simulate_datasets.py* within the MIA package (available at https://ngdc.cncb.ac.cn/mia/download).

For clustering evaluation, we considered the effects of $K$ and $T$ ($K$ varies from 2~6 and $T$ varies from 1~3 with fixed $N$=500, $M_t$=5,000 and $M_f$=5,000) and then the effects of $N$, $M_t$, and $M_f$ ($N$=50, 500, 5,000, $M_t$=1,000, 5,000, 10,000, and $M_f$ =1,000, 5,000, 10,000, with fixed $K$=4 and $T$=2). The Adjusted Rand Index (ARI) was used to quantify clustering accuracy by measuring the similarity between predicted and true labels. The ARI ranges from -1 to 1, with higher values indicating higher accuracy. For feature gene identification, datasets were simulated with $K$=4 and 5, $N$=500, $M_t$=1,000, and $M_f$=10,000. To evaluate the average significance of each feature across different omics layers, identified features were assessed using the Kruskal-Wallis test.

### The TCGA datasets

Multi-omics datasets for ten TCGA cancer types, including CNV (gene-level), gene expression (FPKM), and DNA methylation data, were downloaded from UCSC Xena (https://xenabrowser.net/). For DNA methylation data, the methylation level of each gene was determined by averaging methylation levels of all sites within the first 1,500bp upstream of the gene. For integrative analysis, only samples and genes shared across all three omics types were retained for each cancer type. Survival data and clinical information were obtained from cBioPortal (https://www.cbioportal.org/). Chi-square and Kruskal-Wallis tests were used to assess the significance of six clinical parameters for each identified cluster, and a permutation test was applied to estimate confidence intervals. Tumor stemness indices (mRNAsi and mDNAsi) were adopted by referencing a previous study[33]. Feature genes identified for GBM were annotated by PubMed (https://pubmed.ncbi.nlm.nih.gov/), BrainBase[58], MSigDB[59], and EWAS[60].

### GBM temozolomide treatment datasets

Gene expression datasets related to GBM temozolomide (TMZ) treatment were retrieved from the Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/gds/). Specifically, GSE211272[39] contained 72 samples from 12 GBM cell lines, including 24 control samples, 24 samples treated with TMZ and radiotherapy, and 24 samples treated with TMZ, radiotherapy and molecular inhibitor of Ataxia-telangiectasia and Rad3-Related protein. GSE113510[61] comprised 3 control samples and 3 TMZ-resistant samples cultured from GBM cells. GSE100736 included 3 TMZ-resistant GBM cell lines and 3 TMZ-sensitive cell lines. GSE68029[62] included 3 samples that survived one cycle treatment with 500uM TMZ, and 3 samples that survived two cycle treatment with 500uM TMZ, representing low to high TMZ resistant. The datasets from GSE211272 and GSE68029 were corrected for batch effects using the Combat[63] method. Feature gene expression matrices were then constructed by mapping the predefined set of 167 GBM feature genes to each dataset, with 165, 161, 165, and 162 matched genes in GSE211272, GSE113510, GSE100736, and GSE68029, respectively.

## Multi-omics integration by tensor

Multi-omics data are integrated into a three-dimensional tensor: sample, gene, data type, denoted as $A \in R^{N\times M\times T}$, where $N$ is the number of samples, $M$ is the number of genes, $T$ is the number of omics types, and $R$ is the set of real numbers. For unlabeled data, Tucker decomposition is applied to reduce the tensor $A$ into factor matrices, as defined in Equation 1.

$$A = \mathcal{G} \times S_p \times G_q \times O_r \qquad \text{Eq. (1)}$$

Here, $S \in R^{N\times p}$, $G \in R^{M\times q}$ and $O \in R^{T\times r}$ are the factor matrices of sample-feature, gene-feature, and type-feature, respectively, where $p \in [1, N]$, $q \in [1, M]$, and $r \in [1, T]$ are the number of columns in each factor matrices. $\mathcal{G} \in R^{p\times q\times r}$ is the core tensor, whose entries show interaction between the different components. By default, $p$, $q$ and $r$ are set to $log_2(N)$, $log_2(M)$ and $log_2(W)$, respectively, but they can be specified by users. The sample-feature matrix $S$ is subsequently used for clustering analysis. Tucker decomposition was performed using the Python package "TensorLy"[64] (http://tensorly.org/stable/index.html).

## Fuzzy C-means clustering for unlabeled data

The sample-feature matrix $S$ obtained from Tucker decomposition is used for FCM clustering, which attempts to partition the $N$ samples, denoted as $x_i$ ($i$=1, 2, …, $N$), into $K$ ($K$>1) fuzzy clusters. FCM initializes a set of cluster centers $C_k$ ($k$=1, 2, ⋯, $K$) and a membership matrix $U = u_{ik} \in [0,1]$ ($i$=1, 2, …, $N$ and $k$=1, 2, ⋯, $K$), where each element $u_{ik}$ tells the degree to which sample $x_i$ belongs to cluster $C_k$, satisfying $\sum_{k=1}^{K} u_{ik} = 1$. Given a fuzzy index $m \in (1, \infty)$ and the number of clusters $K$, FCM aims to minimize a loss function: $J(U, C) = \sum_{i=1}^{N}\sum_{k=1}^{K} u_{ik}^m ||x_i - C_k||^2$, where $u_{ik} = \frac{1}{\sum_{q=1}^{K}\left(\frac{||x_i - C_k||}{||x_i - C_q||}\right)^{\frac{2}{m-1}}}$ and $C_k = \frac{\sum_{k=1}^{K} u_{ik}^m x_i}{\sum_{k=1}^{K} u_{ik}^m}$. When $J_n$-$J_{n-1}$<$e^{-15}$, where $n$ represents the number of iterations,

it means the function convergence as indicated by high inter-cluster dispersion and high intra-cluster compactness.

FCM involves two key parameters: the fuzzy index ($m$) and the number of clusters ($K$). For the fuzzy index, Schwammle's formula[28] is employed to estimate a potential optimal value, namely, $m$=1+(1418/$N$+22.05)$M$^(-2)+(12.33/$N$+0.243)$M$^(-0.0406$ln$($N$)-0.1134), where $N$ is the number of samples and $M$ is the number of features. Additionally, an automatic procedure is implemented to determine $K$ when it is not provided. Five clustering evaluation indicators—Silhouette Coefficient[65], Davies-Bouldin Index[66], Elbow method[67], Separation index, and Compactness index—were adopted to compare the accuracy of optimal $K$ estimation. Among them, the Silhouette Coefficient consistently achieved the highest accuracy (**Supplementary Fig. 3**), and therefore was used to estimate the optimal $K$ value.

## Gene filtering by Gini index

The Gini index was used to filter genes (**Supplementary Fig. 4a,b**), defined as $gini(split) = 1 - \sum_{k=1}^{K} p_k^2$, where $p$ represents the proportion of samples belonging to the $k$-th category. For the $t$-th omics type, the Gini index of the $j$-th gene under a partition threshold $X$ is calculated as:

$$Gini(j,t,X) = gini(K) - \frac{N_{value \le X}}{N} gini(value \le X) - \frac{N_{value > X}}{N} gini(value > X), \quad \text{Eq. (2)}$$

where $N$ represents the number of samples, and $value \le X$ (or $value > X$) denotes the subset of samples with values less (or greater) than $X$. The objective is to identify the optimal threshold $X$ that maximizes $Gini(j,t,X)$:$Gini(j,t)=max(Gini(j,t,X_1), Gini(j,t,X_2), \ldots, Gini(j,t,X_n))$. For each omics type, the mean Gini index of $M$ genes is denoted as $meanGini(t) = \frac{1}{M}(\sum_{j=1}^{M} Gini(j,t))$. Genes satisfying the condition $Gini(j,t) > meanGini(t)$ across all omics types are retained. These retained genes, denoted as $M'$, are considered as potential features for subsequent ERFM analysis.

## Featured gene identification based on multi-omics using ERFM

In this study, ERFM, an enhanced RFM, was developed to identify feature genes that play a significant role in clustering on multi-omics data. After gene filtering, ERFM is constructed using $N$ samples, $M'$ filtered genes, and $T$ omics types, and the corresponding sample labels or clustering results. ERFM consists of multiple decision trees, and the workflow is as follows (**Supplementary Fig. 4c**):

**(a) Data preparation**: Using bootstrap sampling, 70% of $N$ samples are randomly selected for training each decision tree, while the remaining 30% samples are used for testing. Each decision tree has different training and testing samples. $M_s$ ($M_s \le sqrt(M')$) genes are randomly selected from the $M'$ genes, and the Gini index is used to select the best combination among the $M_s$ genes for constructing the decision tree.

**(b) Enhanced tree structure**: In a traditional decision tree, each node contains only one-dimensional feature information. To handle multi-omics data, the tree structure is modified and enhanced to integrate multi-omics features of each gene. When a gene $M_i$ is selected, data from all $T$ omics layers of that gene are used to construct tree nodes, yielding $T$ nodes per gene. The optimal splits for each node are determined using $Gini(j,1)$, $Gini(j,2)$, ..., and $Gini(j, T)$. For example, in the case of three omics types, three nodes are built for a single gene. The trees are grown until either the depth reaches $log2(M')$-1 or the number of features reaches $sqrt(M')$.

**(c) Gene importance in a decision tree**: Once a decision tree $d$ is trained, the remaining 30% of samples are used for classification prediction to calculate out-of-bag error rate ($obb_error_d$). To evaluate the importance of a gene $M_i$, the values of all omics layers corresponding to $M_i$ in the test set are simultaneously randomized, and the out-of-bag error rate ($obb_error_{\mathrm{random}(Mi)}$) is calculated. The importance of gene $M_i$ in the decision tree $d$ is then defined as: $importance_d(M_i,d)=obb_error_{\mathrm{random}(Mi)} - obb_error_d$.

**(d) Gene importance in a random forest**: A random forest $r$ creates 100 decision trees $d\in\{1, 2, \ldots,100\}$ by default, repeating **steps (a-c)**. The importance of gene $M_i$ in forest $r$ is calculated as: $importance_r(M_i,r) = \frac{1}{100}\sum_{d=1}^{100} importance_d(M_i,d)$ . The number of genes whose $importance_r(M_i,r)\leqslant 0$ is recorded as $M_{\leq 0}$.

**(e) Iterative computational procedure**: To enhance consistency of the identified feature gene set, an iterative procedure is applied with a termination condition: if $M_{\leq 0}>0$, these $M_{\leq 0}$ genes are removed, and $M'=M'-M_{\leq 0}$ is updated to repeat **steps (a-d)**; if $M_{\leq 0}=0$, the iterative procedure stops, and the remaining $M'$ genes are considered as the final feature gene set. The overall importance of gene $M_i$ is computed as: $importance(M_i) = \frac{1}{R}\sum_{r=1}^{R} importance_r(M_i,r)$, where $R$ represents the number of iterations.

Finally, MIA normalizes and ranks the importance scores of the remaining $M'$ genes, and the Wilcoxon rank-sum test is used to identify genes showing significant differences among clusters (**Supplementary Fig. 4d**).

## Feature gene evaluation

The performance of feature genes was assessed using a unified statistical framework applied to both simulated and empirical datasets. For each gene, the Kruskal–Wallis test was used to quantify its ability to distinguish groups—true labels in simulated datasets or identified clusters in empirical datasets. This test was applied separately to each omics layer (e.g., CNV, expression, methylation), and the mean significance across layers was calculated as the gene's overall significance score. For GBM, the clustering power of the feature-gene set was further assessed by the average Kruskal–Wallis significance and the average Gini index across all feature genes.

## Code development and server implementation

MIA was developed by using standard C++ and Python. The pybind11 library was used to create Python bindings for C++ codes (https://pybind11.readthedocs.io/en/stable/index.html). In addition, a user-friendly web interface was developed, allowing users to perform online analysis. The MIA web interface and its command-line source codes are publicly available at https://ngdc.cncb.ac.cn/mia/.

## References


1. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biology* **18**, 83 (2017).
2. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics data integration, interpretation, and its application. *Bioinformatics and Biology Insights* **14**, 1177932219899051 (2020).
3. Karczewski, K.J. & Snyder, M.P. Integrative omics for health and disease. *Nature Reviews Genetics* **19**, 299-310 (2018).
4. Sandhu, C., Qureshi, A. & Emili, A. Panomics for precision medicine. *Trends in Molecular Medicine* **24**, 85-101 (2018).
5. Xu, Y. et al. Multi-omics analysis at epigenomics and transcriptomics levels reveals prognostic subtypes of lung squamous cell carcinoma. *Biomedicine & Pharmacotherapy* **125**, 109859 (2020).
6. Hoadley, K.A. et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173**, 291-304 (2018).
7. Yoon, S.J. et al. Comprehensive multi-omics analysis reveals aberrant metabolism of Epstein–Barr-virus-associated gastric carcinoma. *Cells* **8**, 1220 (2019).
8. Miao, R. et al. Identification of prognostic biomarkers in hepatitis B virus-related hepatocellular carcinoma and stratification by integrative multi-omics analysis. *Journal of Hepatology* **61**, 840-849 (2014).
9. Gisby, J.S. et al. Multi-omics identify falling LRRC15 as a COVID-19 severity marker and persistent pro-thrombotic signals in convalescence. *Nature Communications* **13**, 7775 (2022).
10. Chen, C. et al. Applications of multi-omics analysis in human diseases. *MedComm* **4**, e315 (2023).
11. Eddy, S., Mariani, L.H. & Kretzler, M. Integrated multi-omics approaches to improve classification of chronic kidney disease. *Nature Reviews Nephrology* **16**, 657-668 (2020).
12. Rappoport, N. & Shamir, R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Research* **46**, 10546-10562 (2018).
13. Wang, B. et al. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods* **11**, 333-337 (2014).
14. Nguyen, T., Tagett, R., Diaz, D. & Draghici, S. A novel approach for data integration and disease subtyping. *Genome Research* **27**, 2025-2039 (2017).
15. Rappoport, N. & Shamir, R. NEMO: cancer subtyping by integration of partial multi-omic data. *Bioinformatics* **35**, 3348-3356 (2019).
16. Tyler, S.R. et al. Merged affinity network association clustering: Joint multi-omic/clinical clustering to identify disease endotypes. *Cell Reports* **35**, 108975 (2021).
17. Argelaguet, R. et al. Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology* **14**, e8124 (2018).
18. Shi, Q. et al. Pattern fusion analysis by adaptive alignment of multiple heterogeneous omics data. *Bioinformatics* **33**, 2706-2714 (2017).

19. Chalise, P. & Fridley, B.L. Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. *PLoS One* **12**, e0176278 (2017).
20. Fang, J. Tightly integrated genomic and epigenomic data mining using tensor decomposition. *Bioinformatics* **35**, 112-118 (2019).
21. Harshman, R.A. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA working papers in phonetics* **16**, 84 (1970).
22. Kroonenberg, P.M. Three-mode principal component analysis: Theory and applications, Vol. 2. (DSWO press, 1983).
23. Carroll, J.D. & Chang, J.J. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika* **35**, 283-319 (1970).
24. Cichocki, A. et al. Tensor networks for dimensionality reduction and large-scale optimization: part 1 low-rank tensor decompositions. *Foundations and Trends in Machine Learning* **9**, 249-429 (2016).
25. Jung, I., Kim, M., Rhee, S., Lim, S. & Kim, S. MONTI: A multi-omics non-negative tensor decomposition framework for gene-level integrative analysis. *Frontiers in Genetics* **12**, 682841 (2021).
26. Zhang, J.Z., Xu, W. & Hu, P. Tightly integrated multiomics-based deep tensor survival model for time-to-event prediction. *Bioinformatics* **38**, 3259-3266 (2022).
27. Chang, S.M. et al. Gene-set integrative analysis of multi-omics data using tensor-based association test. *Bioinformatics* **37**, 2259-2265 (2021).
28. Schwammle, V. & Jensen, O.N. A simple and fast method to determine the parameters for fuzzy c-means cluster analysis. *Bioinformatics* **26**, 2841-2848 (2010).
29. Adair, J.E. et al. Gene therapy enhances chemotherapy tolerance and efficacy in glioblastoma patients. *Journal of Clinical Investigation* **124**, 4082-4092 (2014).
30. Grochans, S. et al. Epidemiology of Glioblastoma Multiforme–Literature Review. *Cancers* **14**, 2412 (2022).
31. Verhaak, R. et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**, 98-110 (2010).
32. Mehta, K.R. et al. Fractional genomic alteration detected by array-based comparative genomic hybridization independently predicts survival after hepatic resection for metastatic colorectal cancer. *Clinical Cancer Research* **11**, 1791-1797 (2005).
33. Malta, T.M. et al. Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell* **173**, 338-354 (2018).
34. Joseph, J.V. et al. Serum-induced differentiation of glioblastoma neurospheres leads to enhanced migration/invasion capacity that is associated with increased MMP9. *PLoS One* **10**, e0145393 (2015).
35. Wang, X. et al. Reciprocal signaling between glioblastoma stem cells and differentiated tumor dells promotes malignant progression. *Cell Stem Cell* **22**, 514-528 (2018).
36. Sun, X. et al. BRD8 maintains glioblastoma by epigenetic reprogramming of the p53 network. *Nature* **613**, 195-202 (2023).
37. Tang, Q. et al. Comprehensive analysis reveals that LTBR is a immune-related biomarker for glioma. *Computers in Biology and Medicine* **174**, 108457 (2024).
38. Rusu, P. et al. GPD1 specifically marks dormant glioma stem cells with a distinct metabolic profile. *Cell Stem Cell* **25**, 241-257 (2019).
39. Lozinski, M. et al. ATR inhibition using gartisertib enhances cell death and synergises with temozolomide and radiation in patient-derived glioblastoma cell lines. *Oncotarget* **15**, 1-18 (2024).

40. Baroni, M. et al. Musashi1 Contribution to Glioblastoma Development via Regulation of a Network of DNA Replication, Cell Cycle and Division Genes. *Cancers* **13**, 1494 (2021).
41. Zhang, Y. et al. METTL3-mediated N6-methyladenosine modification and HDAC5/YY1 promote IFFO1 downregulation in tumor development and chemo-resistance. *Cancer Lett* **553**, 215971 (2023).
42. He, Y. et al. Examination of the functions and mechanism of KCP in mediating paclitaxel resistance in cervical squamous carcinoma cells. *Front Oncol* **15**, 1550032 (2025).
43. Majmudar, P.R. & Keri, R.A. The neural stem cell gene PAFAH1B1 controls cell cycle progression, DNA integrity, and paclitaxel sensitivity of triple-negative breast cancer cells. *J Biol Chem* **301**, 110235 (2025).
44. Zhang, L. et al. NRBP2 Overexpression Increases the Chemosensitivity of Hepatocellular Carcinoma Cells via Akt Signaling. *Cancer Research* **76**, 7059-7071 (2016).
45. Lu, M. et al. Long non-coding RNA LINC00665 promotes gemcitabine resistance of Cholangiocarcinoma cells via regulating EMT and stemness properties through miR-424-5p/BCL9L axis. *Cell Death Dis* **12**, 72 (2021).
46. Liu, X. et al. FOXP4 Is a Direct YAP1 Target That Promotes Gastric Cancer Stemness and Drives Metastasis. *Cancer Research* **84**, 3574-3588 (2024).
47. Wang, Y. et al. Expression of small glutamine-rich TPR-containing protein A (SGTA) in Non-Hodgkin's Lymphomas promotes tumor proliferation and reverses cell adhesion-mediated drug resistance (CAM-DR). *Leuk Res* **38**, 955-963 (2014).
48. Kim, J. et al. MAPK13 stabilization via m(6)A mRNA modification limits anticancer efficacy of rapamycin. *J Biol Chem* **299**, 105175 (2023).
49. You, J.S. & Jones, P.A. Cancer genetics and epigenetics: two sides of the same coin? *Cancer Cell* **22**, 9-20 (2012).
50. Claussen, H., Ghosh, S. & Chen, J. Exploring combined effects of DNA methylation and copy number on gene expression with a two-stage approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **21**, 1667-1675 (2024).
51. Brat, D.J. et al. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *New England Journal of Medicine* **372**, 2481-2498 (2015).
52. Hurst, C.D. et al. Stage-stratified molecular profiling of non-muscle-invasive bladder cancer enhances biological, clinical, and therapeutic insight. *Cell Reports Medicine* **2**, 100472 (2021).
53. Chakraborty, G. et al. Fraction genome altered (FGA) to regulate both cell autonomous and non-cell autonomous functions in prostate cancer and its effect on prostate cancer aggressiveness. *Journal of Clinical Oncology* **38**, 347 (2020).
54. Zhou, W. et al. Counting alleles to predict recurrence of early-stage colorectal cancers. *Lancet* **359**, 219-225 (2002).
55. Choi, S.W. et al. Genetic classification of colorectal cancer based on chromosomal loss and microsatellite instability predicts survival. *Clinical Cancer Research* **8**, 2311-2322 (2002).
56. Zhao, R. et al. Clonal dynamics and Stereo-seq resolve origin and phenotypic plasticity of adenosquamous carcinoma. *Npj Precision Oncology* **7**, 80 (2023).
57. Cahill, D.P., Kinzler, K.W., Vogelstein, B. & Lengauer, C. Genetic instability and darwinian selection in tumours. *Trends in Cell Biology* **9**, M57-M60 (1999).
58. Liu, L. et al. BrainBase: a curated knowledgebase for brain diseases. *Nucleic Acids Research* **50**, D1131-D1138 (2022).
59. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Systems* **1**, 417-425 (2015).

60. Li, M. et al. EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. *Nucleic Acids Res* **47**, D983-d988 (2019).
61. Wu, P. et al. Lnc-TALC promotes O(6)-methylguanine-DNA methyltransferase expression via regulating the c-Met pathway by competitively binding with miR-20b-3p. *Nature Communications* **10**, 2045 (2019).
62. Tso, J.L. et al. Bone morphogenetic protein 7 sensitizes O6-methylguanine methyltransferase expressing-glioblastoma stem cells to clinically relevant dose of temozolomide. *Molecular Cancer* **14**, 189 (2015).
63. Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E. & Storey, J.D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882-883 (2012).
64. Kossaifi, J., Panagakis, Y., Anandkumar, A. & Pantic, M. TensorLy: Tensor Learning in Python. *Journal of Machine Learning Research* **20**, 1-6 (2019).
65. Zhu, L., Ma, B. & Zhao, X. Clustering validity analysis based on silhouette coefficient. *Journal of Computer Applications* **30**, 139-141 (2010).
66. Xiao, J., Lu, J. & Li, X. Davies Bouldin Index based hierarchical initialization K-means. *Intelligent Data Analysis* **21**, 1327-1338 (2017).
67. Archdeacon, T.J. Correlation and regression analysis: a historian's guide. (University of Wisconsin Press, USA; 1994).

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (T2425005 and 32030021).

## Author contributions

Yang Zhang: Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Validation, Writing - original draft, Writing - review & editing; Lin Liu: Data curation; Lina Ma: Conceptualization, Project administration, Writing - review & editing; Zhang Zhang: Conceptualization, Funding acquisition, Project administration, Writing - review & editing. All authors have read and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

# Figures

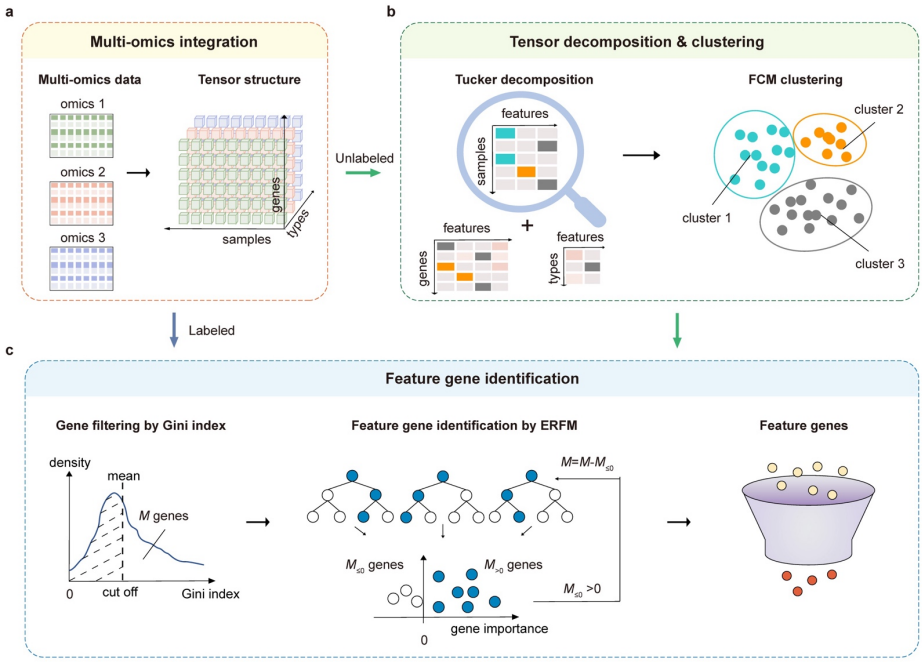



**Fig. 1. Schematic of the MIA framework.**

(a) Various types of omics data (e.g., CNV, expression, and methylation data) are integrated into a tensor, forming a three-dimensional matrix comprising samples, genes, and omics types. (b) For unlabeled samples, Tucker decomposition is used to reduce the tensor dimensionality, transforming it into three two-dimensional factor matrices: sample-feature matrix, gene-feature matrix, and type-feature matrix. Subsequently, the FCM clustering algorithm is used to cluster the sample-feature matrix with labeled information. (c) Feature gene identification is performed based on the tensor and clustering/labeled information. After gene filtering by the Gini index, an ERFM designed for three-dimentional matrix is applied to evaluate gene importance. The importance scores of all genes are iteratively calculated until all remaining genes show positive importance values (see details in **Methods**).

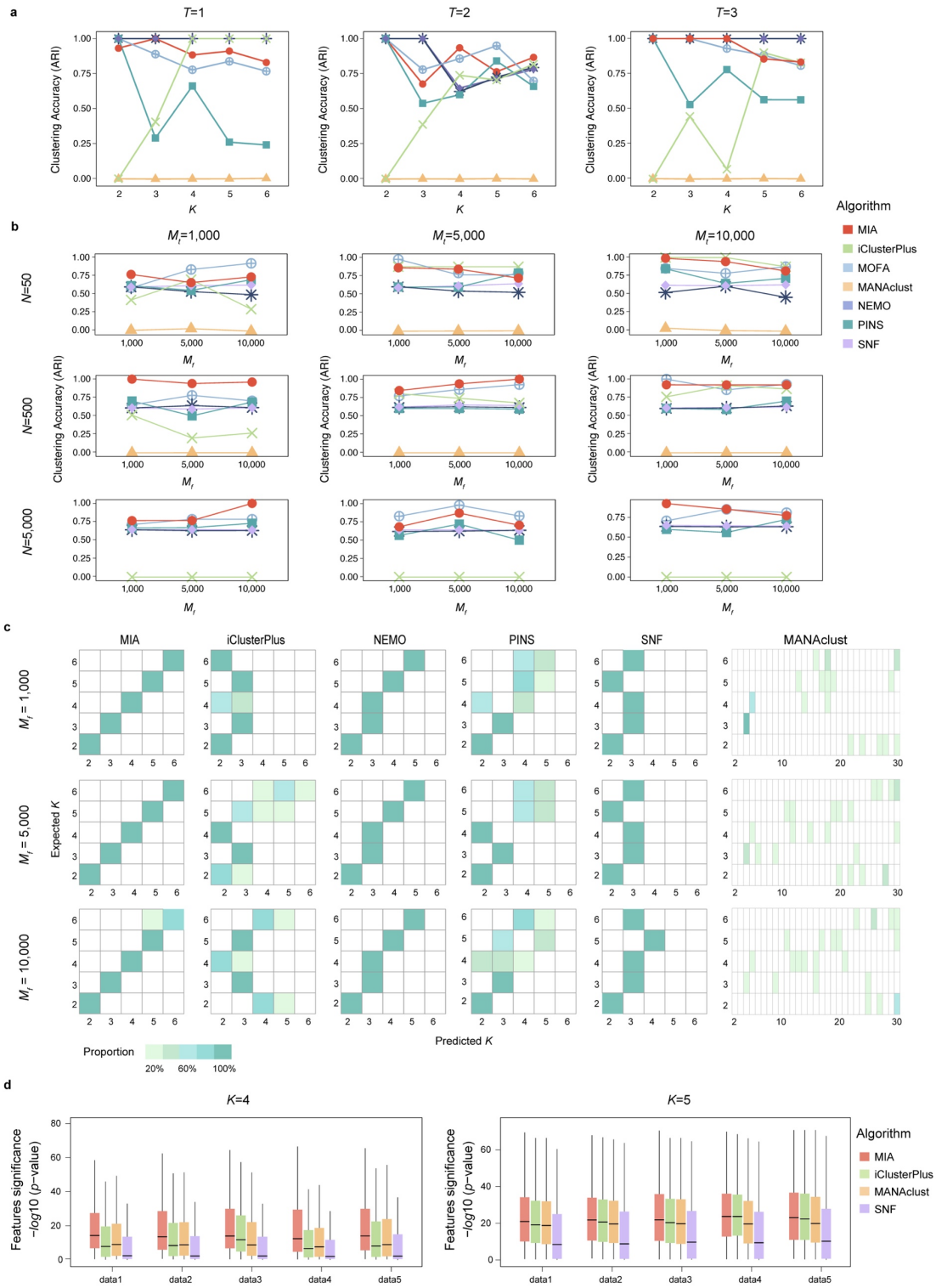



**Fig. 2. Performance of MIA in clustering and feature selection on simulated datasets by comparison against several existing algorithms.**

(a) Clustering performance across different numbers of clusters ($K$=2~6) and omics types ($T$=1~3), with fixed sample size ($N$=500), real features ($M_r$=5,000), and noise features ($M_f$=5,000). Clustering accuracy is assessed by the Adjusted Rand Index (ARI). (b) Clustering performance across varied sample sizes ($N$=50, 500, and 5,000), real features ($M_r$=1,000, 5000, and 10,000), and noise features ($M_f$=1,000, 5,000,

and 10,000), with fixed $K$=4 and $T$=2. (c) Evaluation of optimal $K$ estimation by different algorithms based on simulated datasets consisting of $K$=2~6, $M_f$ =1,000, 5,000, and 10,000, with fixed $N$=500 and $M_t$=1,000. The proportion of predicted $K$ values identified by each algorithm is calculated across simulated replicates. (d) Performance of feature genes identified by different algorithms is evaluated by the Kruskal-Wallis test on five simulated replicates (data1~data5) with $K$=4~5, $T$=2, $N$=500, $M_t$ =1,000, and $M_f$=10,000.

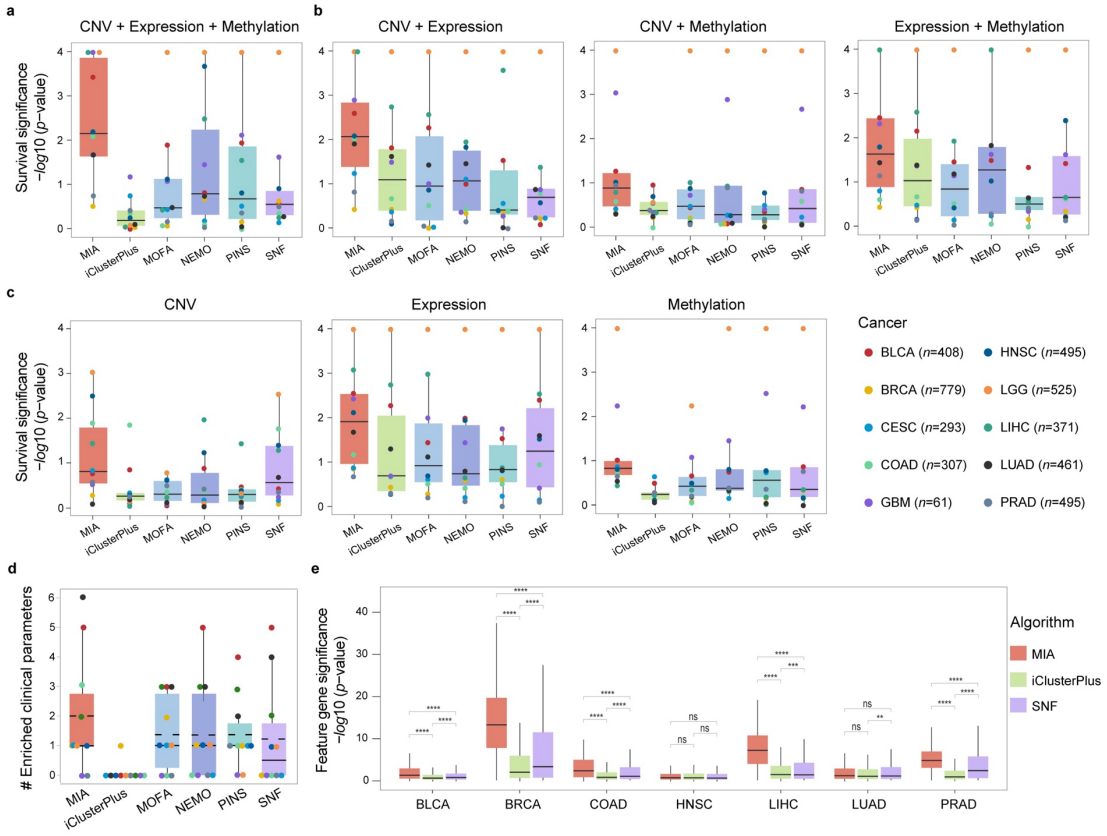



**Fig. 3. Evaluation of survival outcomes and feature genes determined by MIA on TCGA datasets in comparison with existing algorithms.**

Comparison of clustering results from different algorithms based on survival significance in (a) three-omics, (b) two-omics, and (c) single-omics datasets. (d) Number of significantly enriched clinical parameters ($p$<0.05) across ten cancer types. Horizontal lines indicate the median and dashed lines represent the mean. (e) Feature gene signifiance between cancer and normal samples across seven cancer types. The Kruskal-Wallis test was used to evaluate the average significance of feature genes identified by each algorithm ($p$-value, with ****<0.0001, ***<0.001, **<0.01, *<0.05, and ns not signifiacnt).

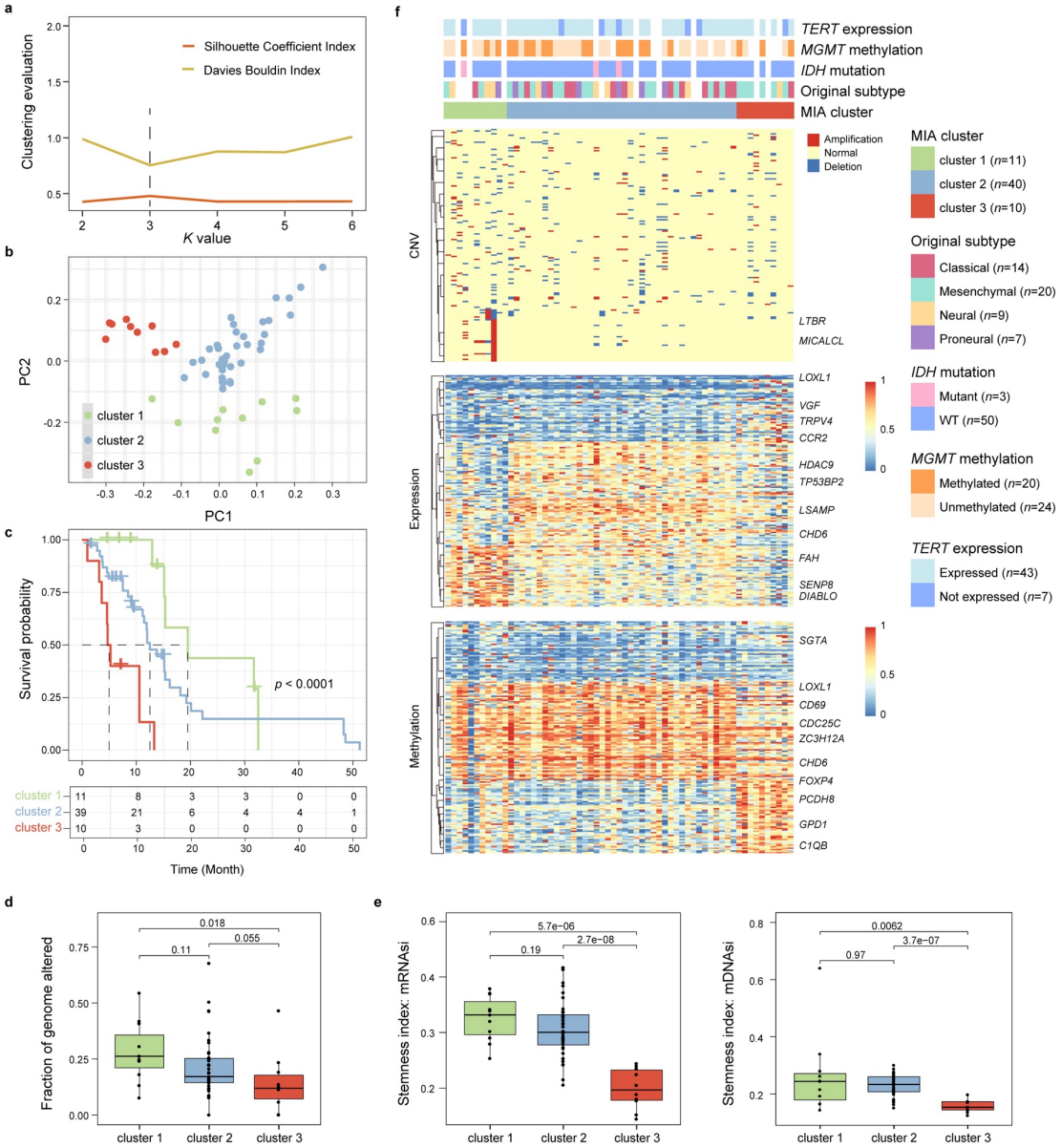



**Fig. 4. TCGA GBM subtyping analysis by MIA.**

(a) Clustering evaluation at different *K* values using the silhouette coefficient and Davies-Bouldin index. A higher silhouette coefficient and lower Davies-Bouldin index indicate better cluster separation. (b) Principal component analysis (PCA) of the sample-feature matrix obtained from Tucker decomposition. (c) Kaplan-Meier survival curves for the three clusters identified by MIA. The log-rank test was used to assess the statistical significance and the accompanying table shows the number of samples and their survivals at specific time points. (d) Fractions of genome altered across the three clusters. (e) Stemness indices (mRNAsi and mDNAsi) across the three clusters, with statistical significance assessed by the Wilcoxon rank-sum test. (f) Heatmaps of 167 GBM feature genes identified by MIA.

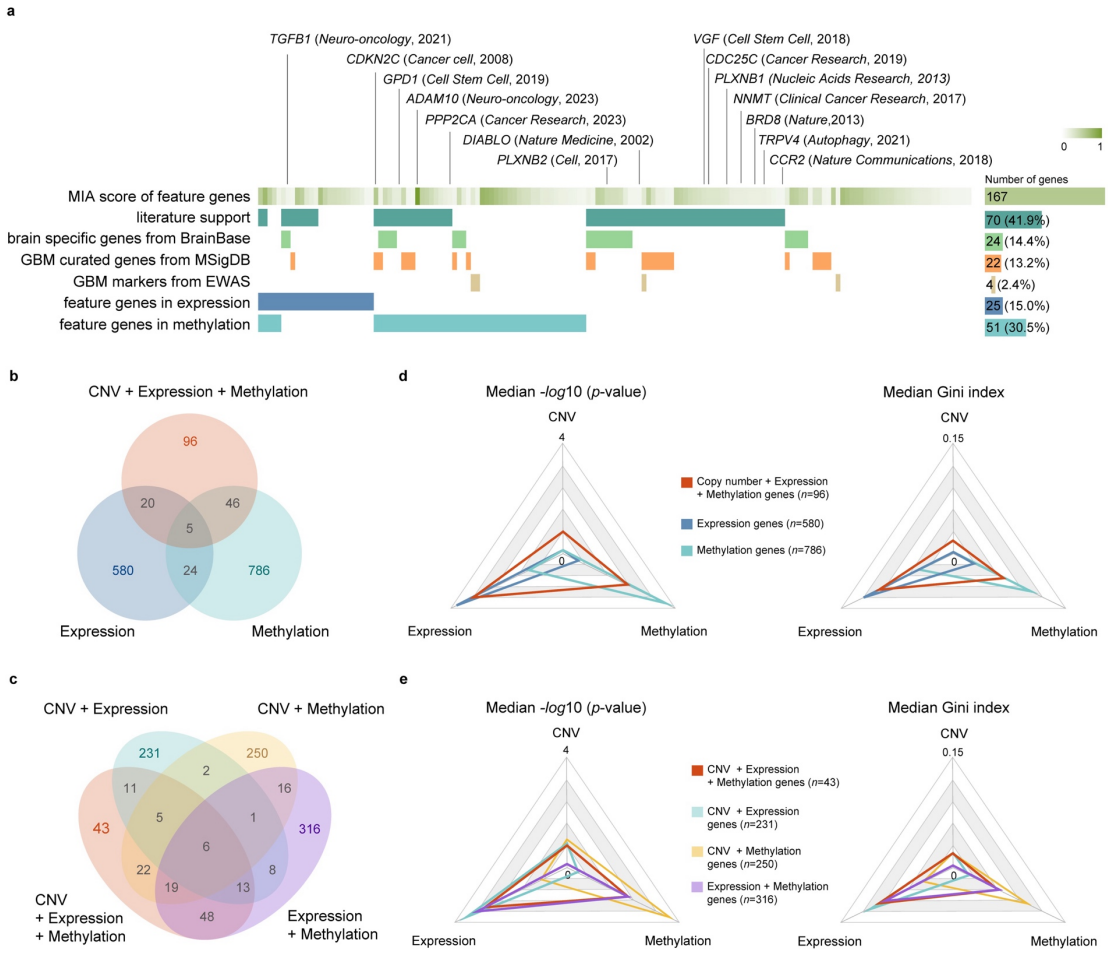



**Fig. 5. Characterization of 167 GBM feature genes identified by MIA.**

(a) Feature genes ranked by MIA scores (0~1), with literaure evidence, database support, and genes identified by single-omics data analysis (genes identified by CNV did not reach the threshold) (**Supplementary Table 9**). (b-c) Feature genes identified from three-omics, by comparsion with single-omics and two-omics, respectively. (d-e) Clustering power of feature genes specifically identified from three-omics, by comparsion with single-omics and two-omics, respectively. The Kruskal-Wallis test and the Gini index were employed to evaluate the clustering power.

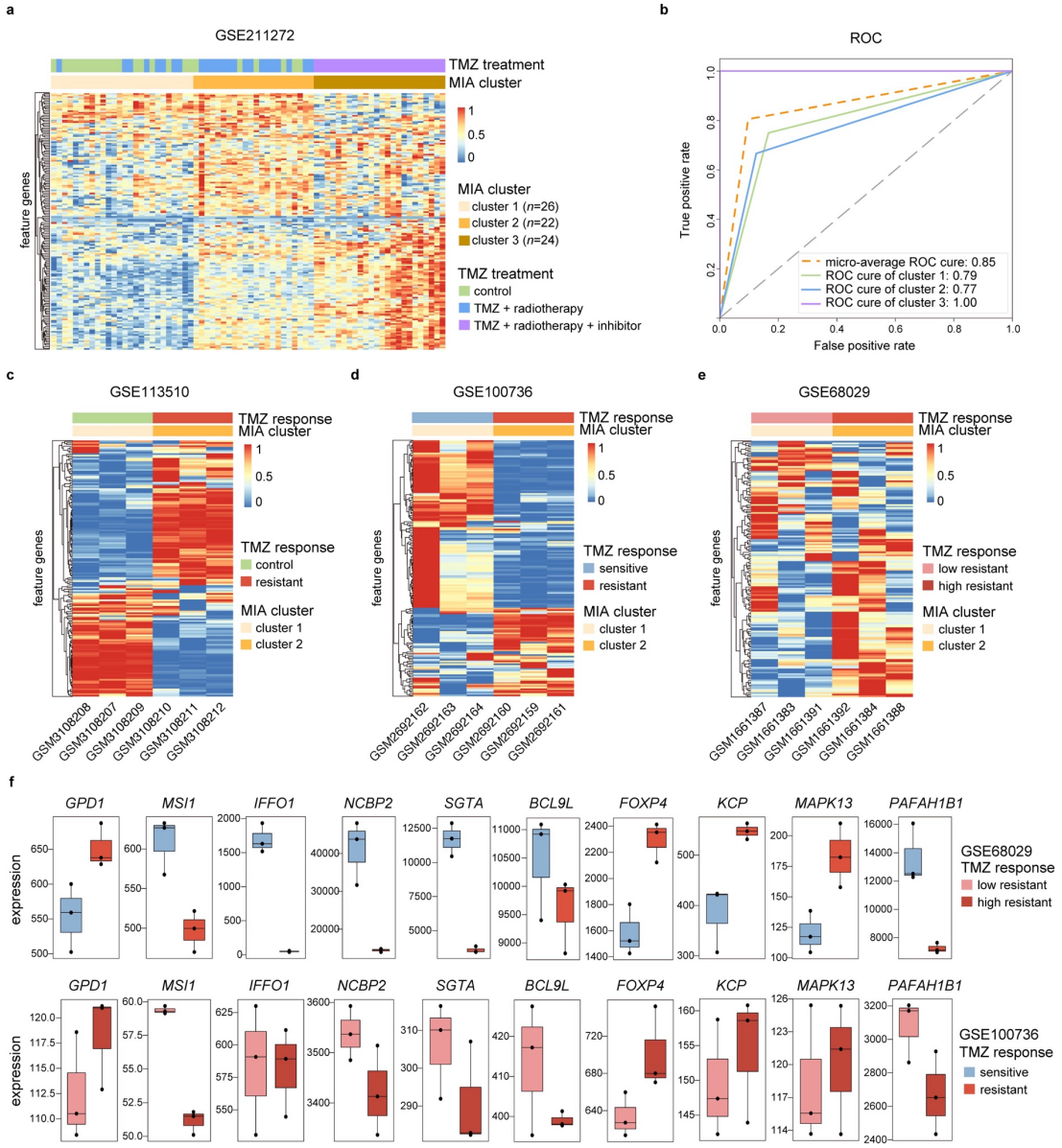
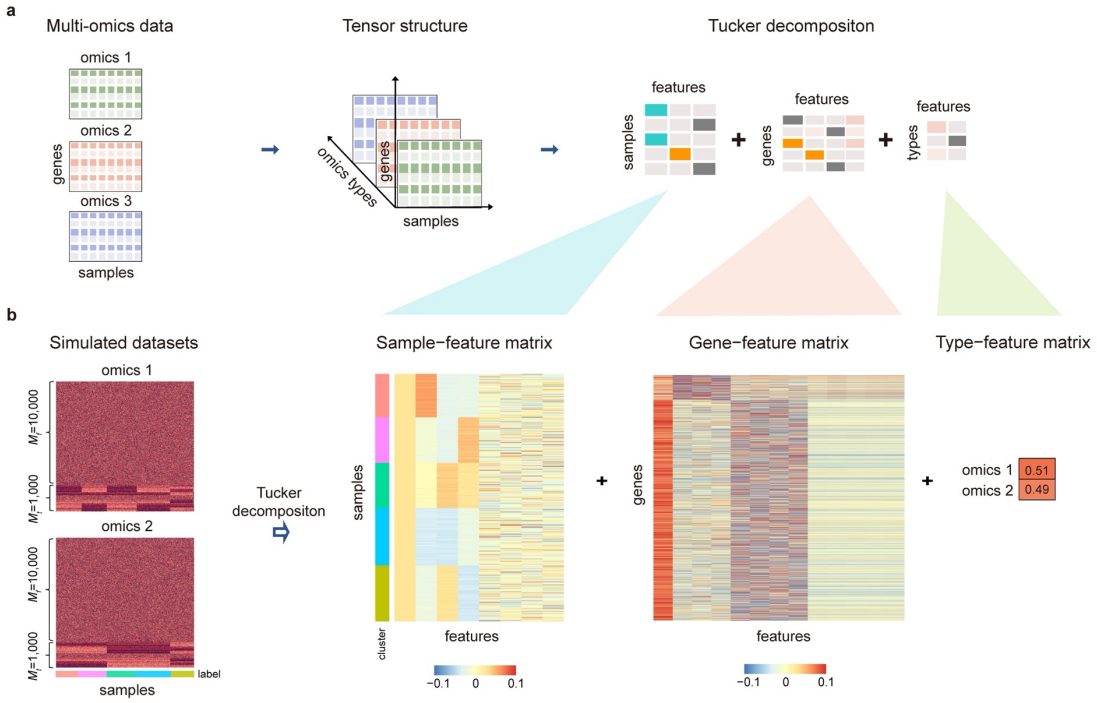



**Fig. 6. MIA Clustering of 167 GBM feature genes in TMZ treatment datasets.**

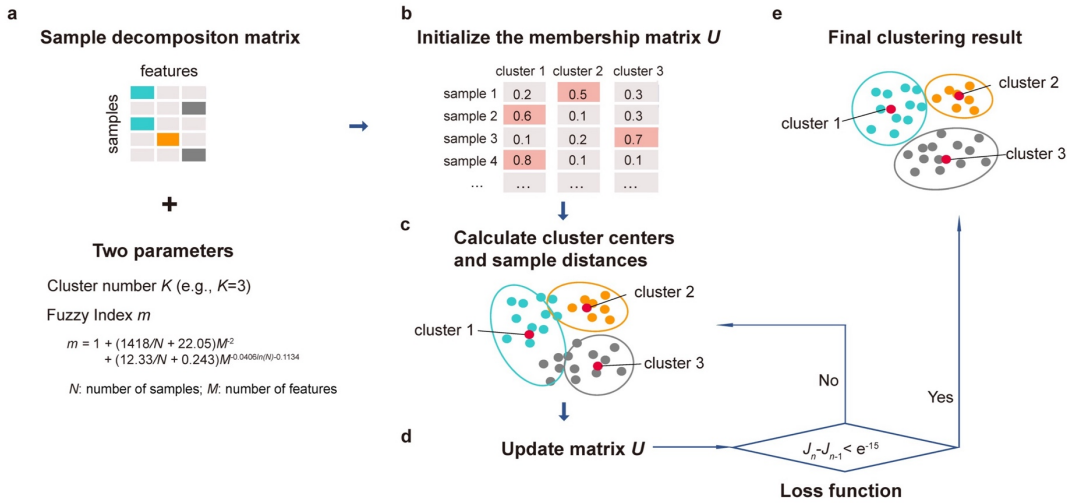
(a) MIA Clustering of 167 GBM feature genes in GSE211272, including control, TMZ + radiotherapy, and TMZ + radiotherapy + inhibitor samples. (b) Multi-class ROC curve illustrating clustering performance on GSE211272. (c-e) MIA Clustering of 167 GBM feature genes in three TMZ treatment datasets, viz., GSE113510 (control vs resistant), GSE100736 (resistant vs sensitive), and GSE68029 (low resistant vs high resistant). (f) The literature-reported feature genes, exhibiting consistent expression patterns in both GSE100736 and GSE68029.

## Supplementary Figures

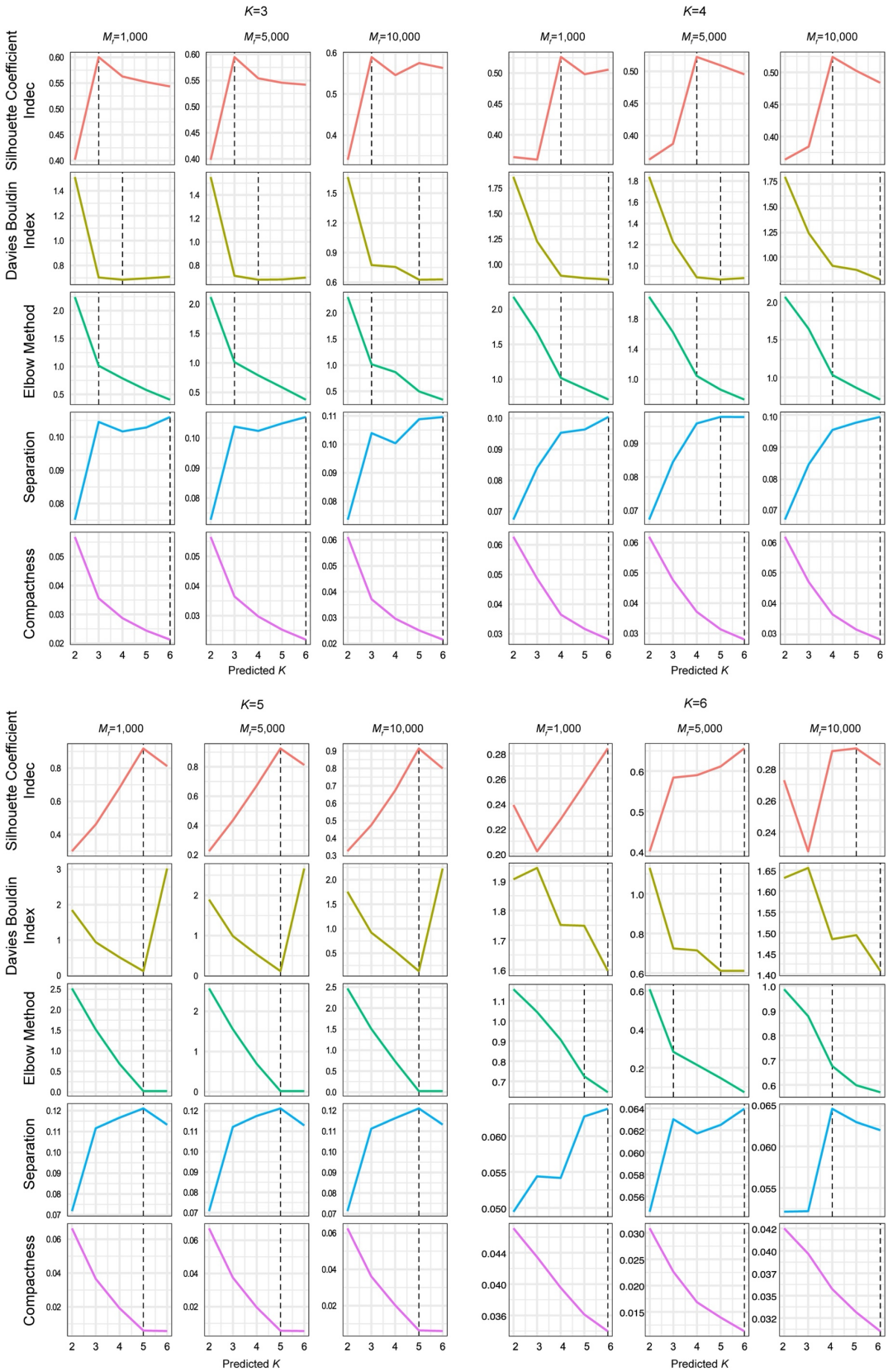



**Supplementary Fig. 1. Schematic of Tucker decomposition for multi-omics data.**

(a) Tensor integration of multi-omics data into a three-dimensional matrix, and Tucker decomposition reducing the tensor into three factor matrices: sample-feature matrix, gene-feature matrix, and type-feature matrix. (b) Tucker decomposition results on simulated datasets with $T$=2, $K$=5, $N$=500, $M_t$=1,000, and $M_f$=10,000.

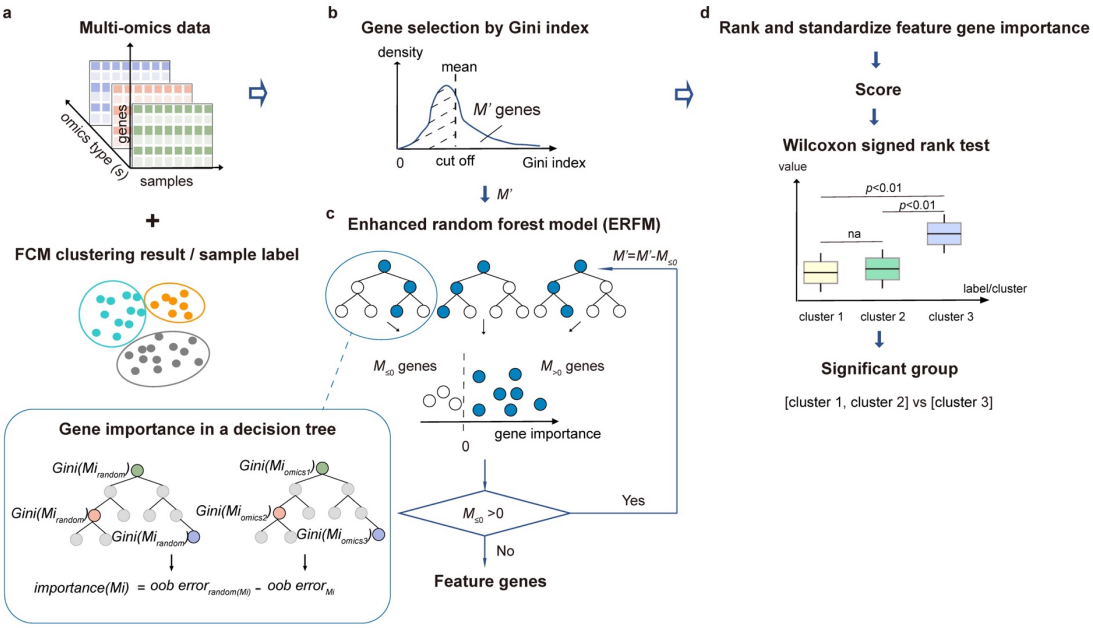



**Supplementary Fig. 2. Schematic of multi-omics data integration and FCM clustering.** (a) FCM clustering using the sample-feature matrix with specified cluster count (*K*) and fuzzy index (*m*). (b) Membership matrix *U* initialized to represent samples-cluster probability. (c) Cluster center vectors calculated from *U*. (d) *U* iteratively updated based on cluster centers and *m*, stopping when the loss function converges below a threshold. (e) Each sample assigned to the cluster with the highest membership probability in *U*.



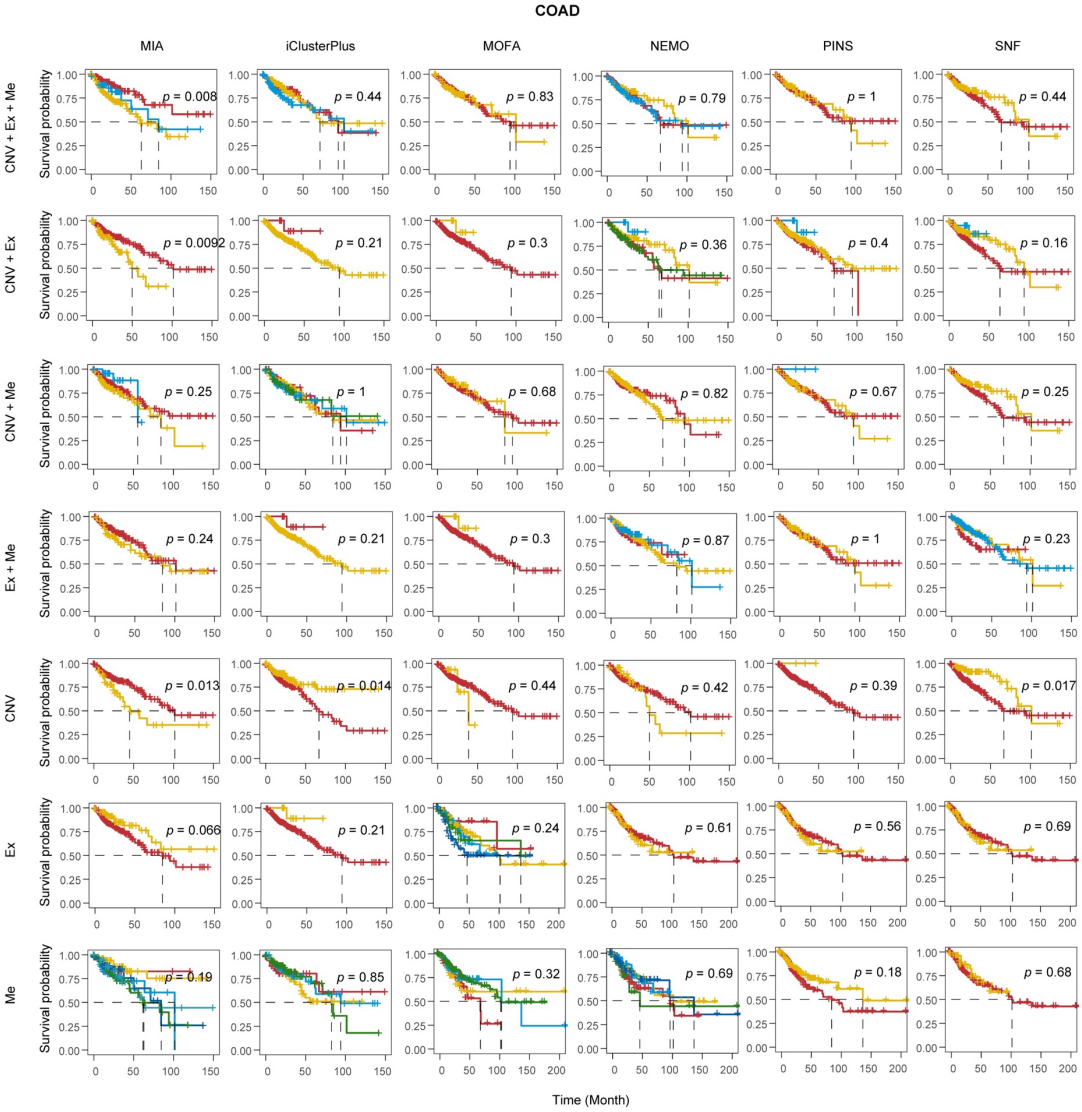

**Supplementary Fig. 3. Evaluation of five clustering evaluation indicators for optimal *K* selection on simulated datasets when *K* varies from 3 to 6.**

Simulated datasets comprise $K$=3~6 and $M_f$=1,000, 5,000, and 10,000, with fixed $N$=500 and $M_t$=1,000. The optimal $K$ values are evaluated by five indicators: Silhouette Coefficient, Davis Bouldin Index, Elbow method, Separation index and Compactness index.



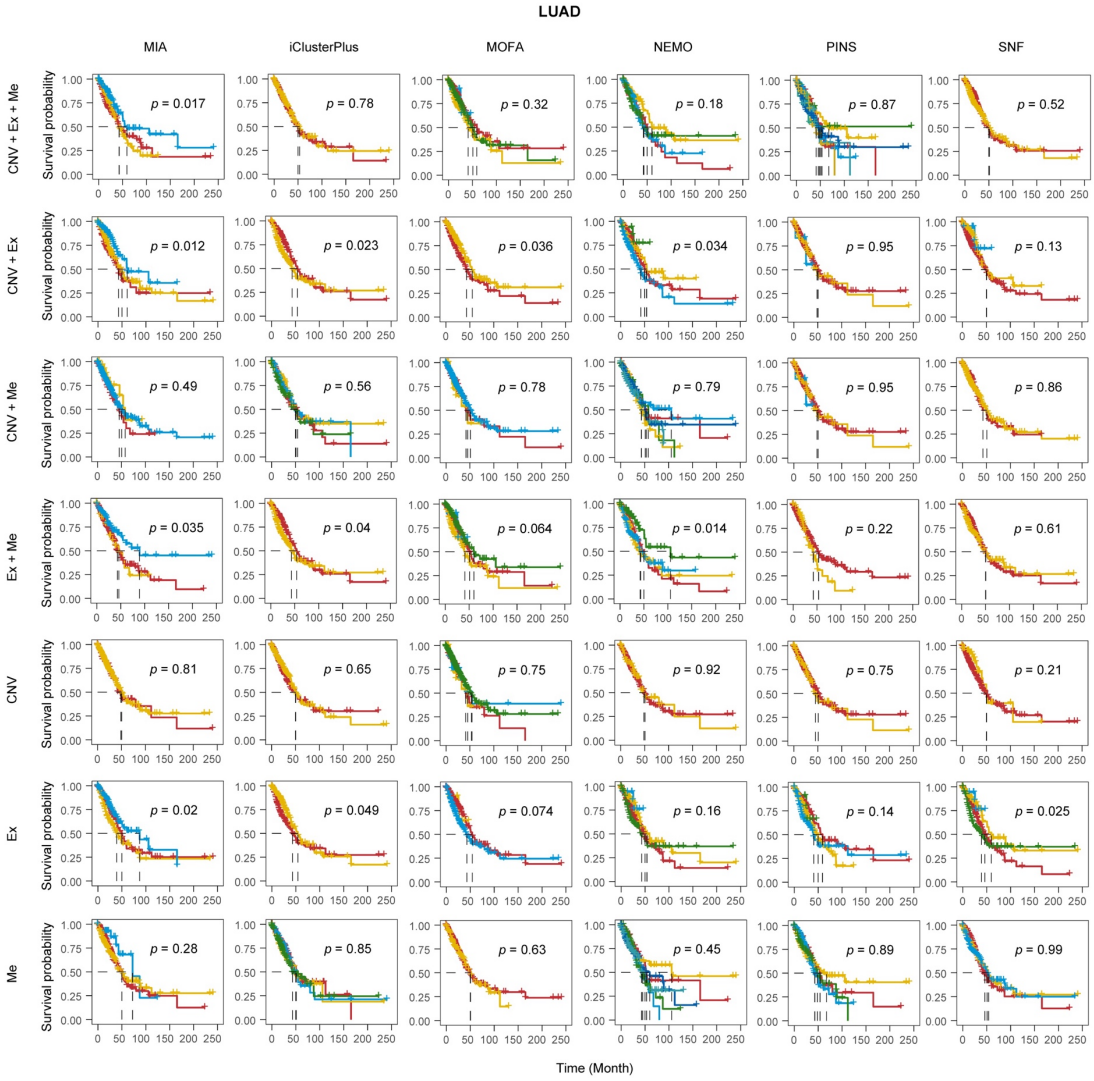

**Supplementary Fig. 4. Feature gene identification by ERFM.**

(a) Input multi-omics data with corresponding sample labels or clustering results. (b) Preliminary feature selection using the Gini index. Genes with values above the mean across all omics are retained. (c) Construction of the ERFM adapted for three-dimensional matrix. Gene importance is determined by the reduction in classification accuracy after randomizing multi-omics data for each gene. The computational procedure iterates until all genes have positive importance values. (d) Final feature-gene scoring: importance rankings are standardized, and the Wilcoxon rank-sum test evaluates significant differences across clusters (see **Methods** for details).



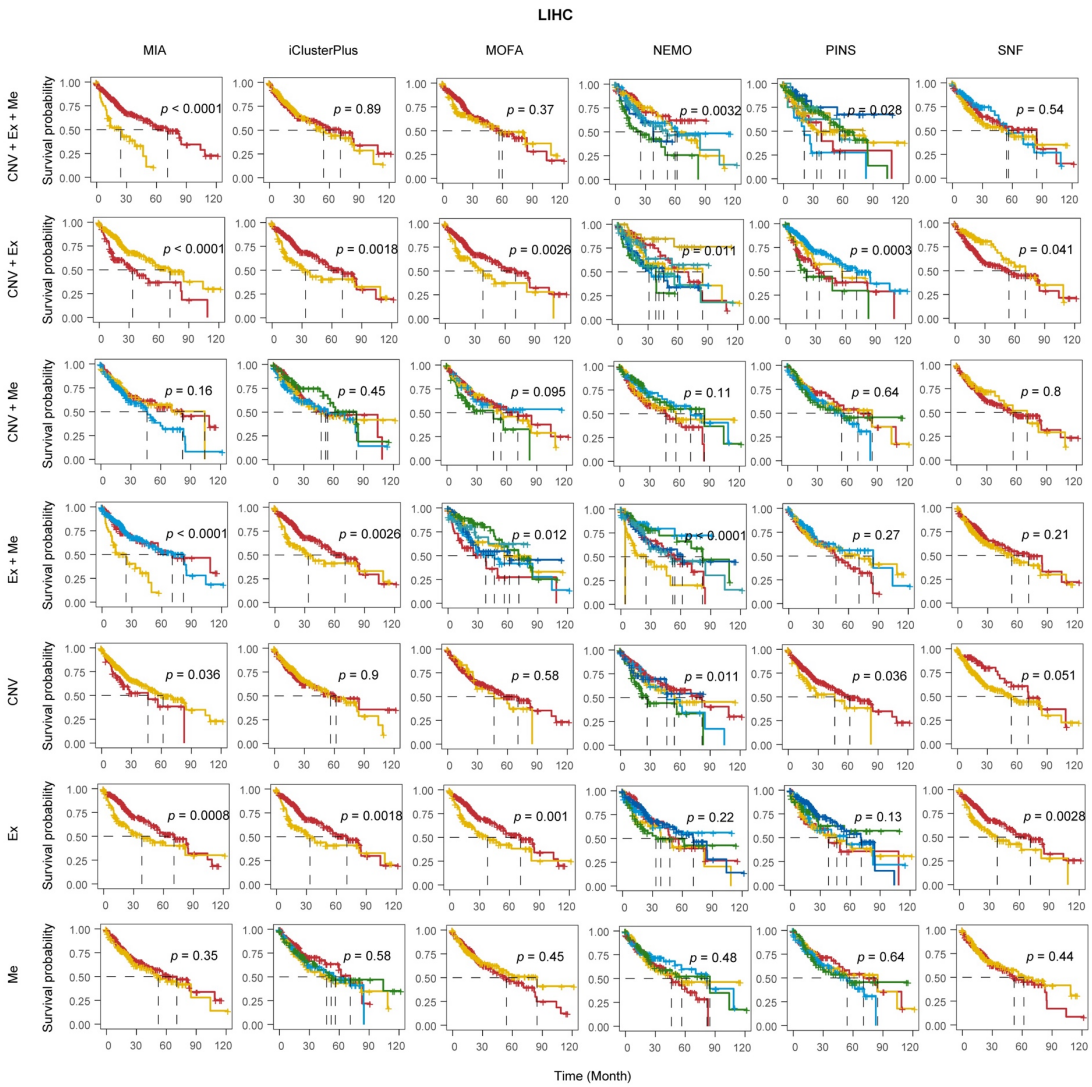

**Supplementary Fig. 5. Kaplan-Meier survival analysis on TCGA COAD dataset clustered by different algorithms across various omics combinations.**

The log-rank test was used to assess the statistical significance between survival curves. Abbreviations: CNV, copy number variation; Ex, gene expression; Me, DNA methylation.



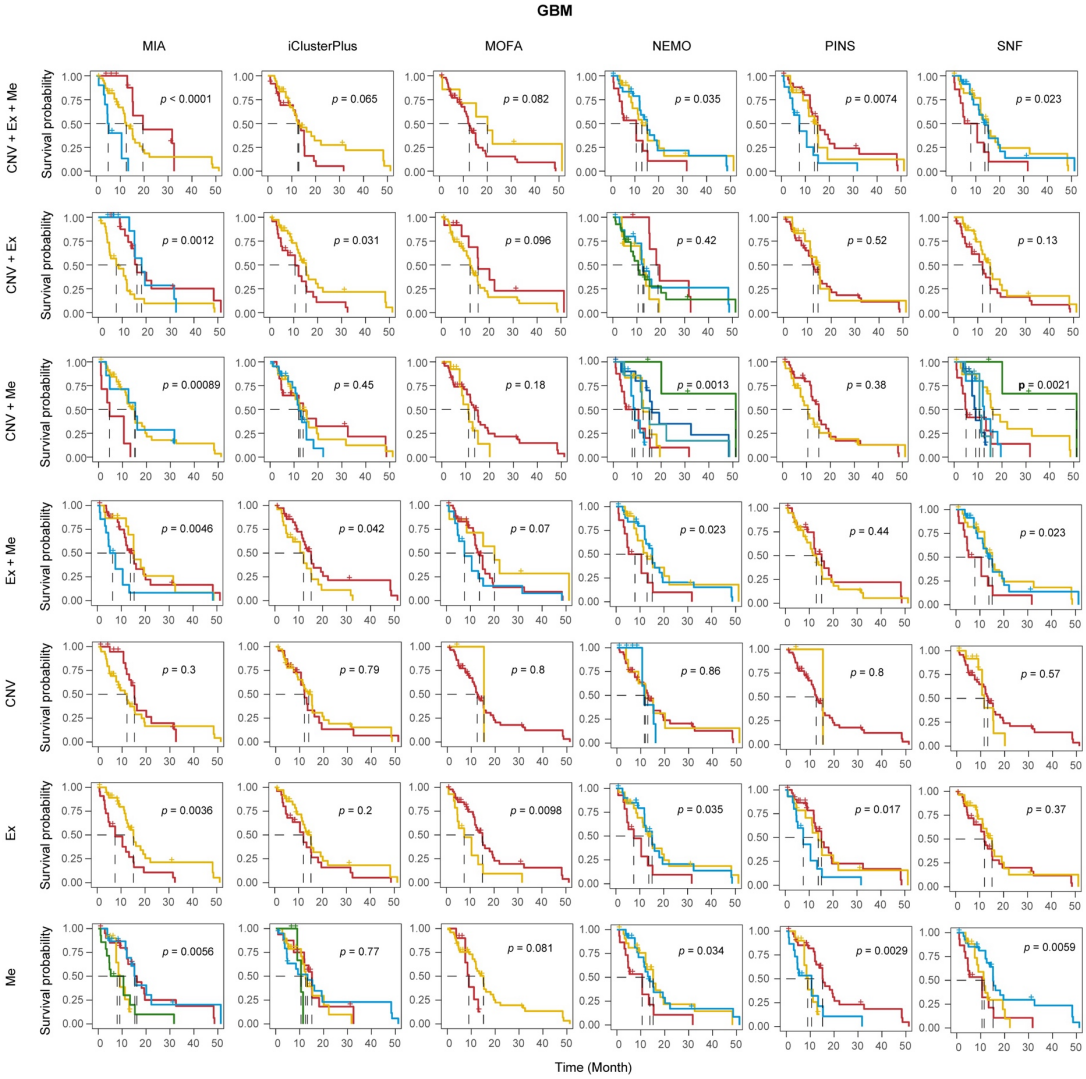

**Supplementary Fig. 6. Kaplan-Meier survival analysis on TCGA LUAD dataset clustered by different algorithms across various omics combinations.**

The log-rank test was used to assess statistical significance between survival curves. Abbreviations: CNV, copy number variation; Ex, gene expression; Me, DNA methylation.



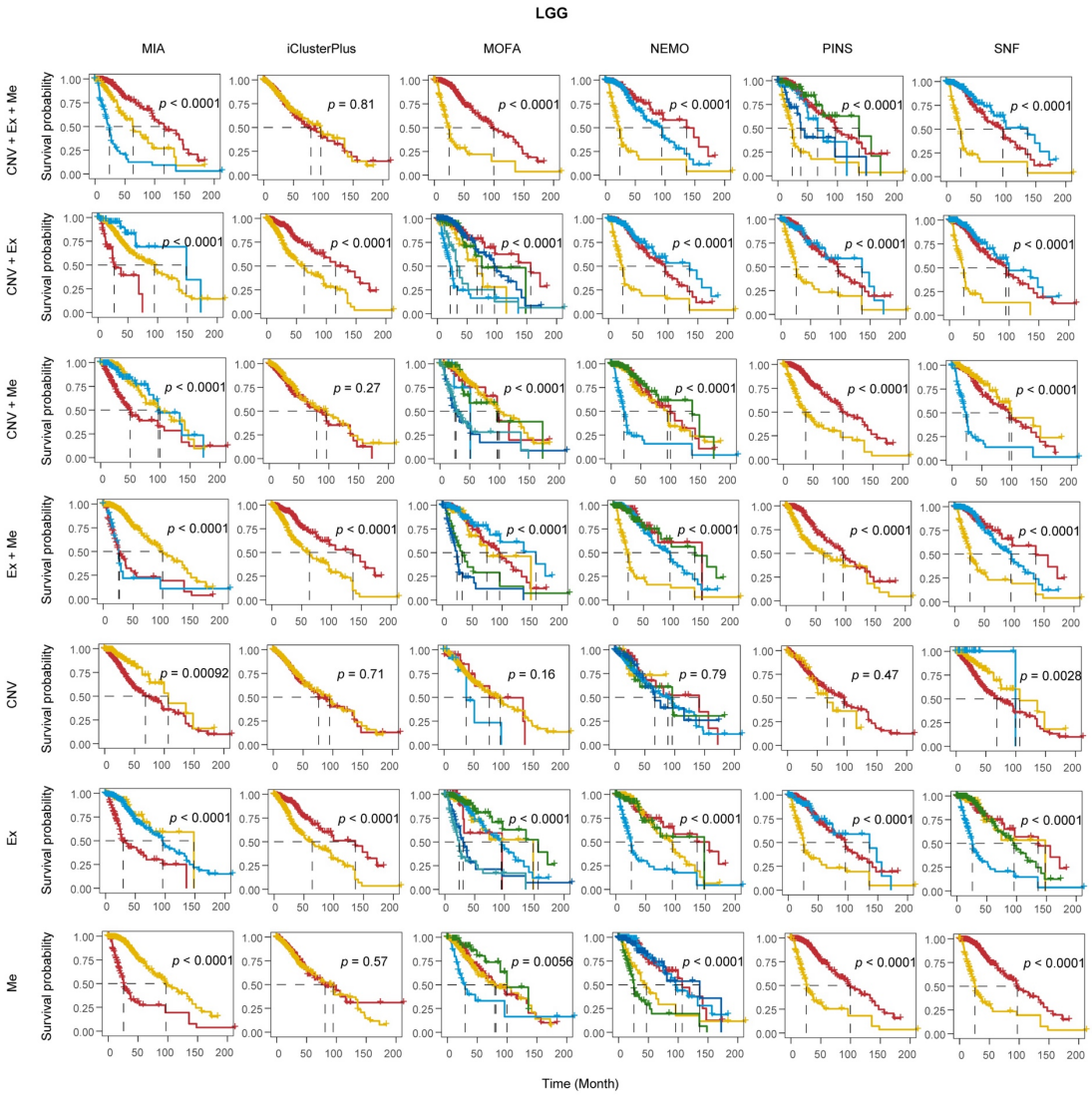

**Supplementary Fig. 7. Kaplan-Meier survival analysis on TCGA LIHC dataset clustered by different algorithms across various omics combinations.**

The log-rank test was used to assess statistical significance between survival curves. Abbreviations: CNV, copy number variation; Ex, gene expression; Me, DNA methylation.



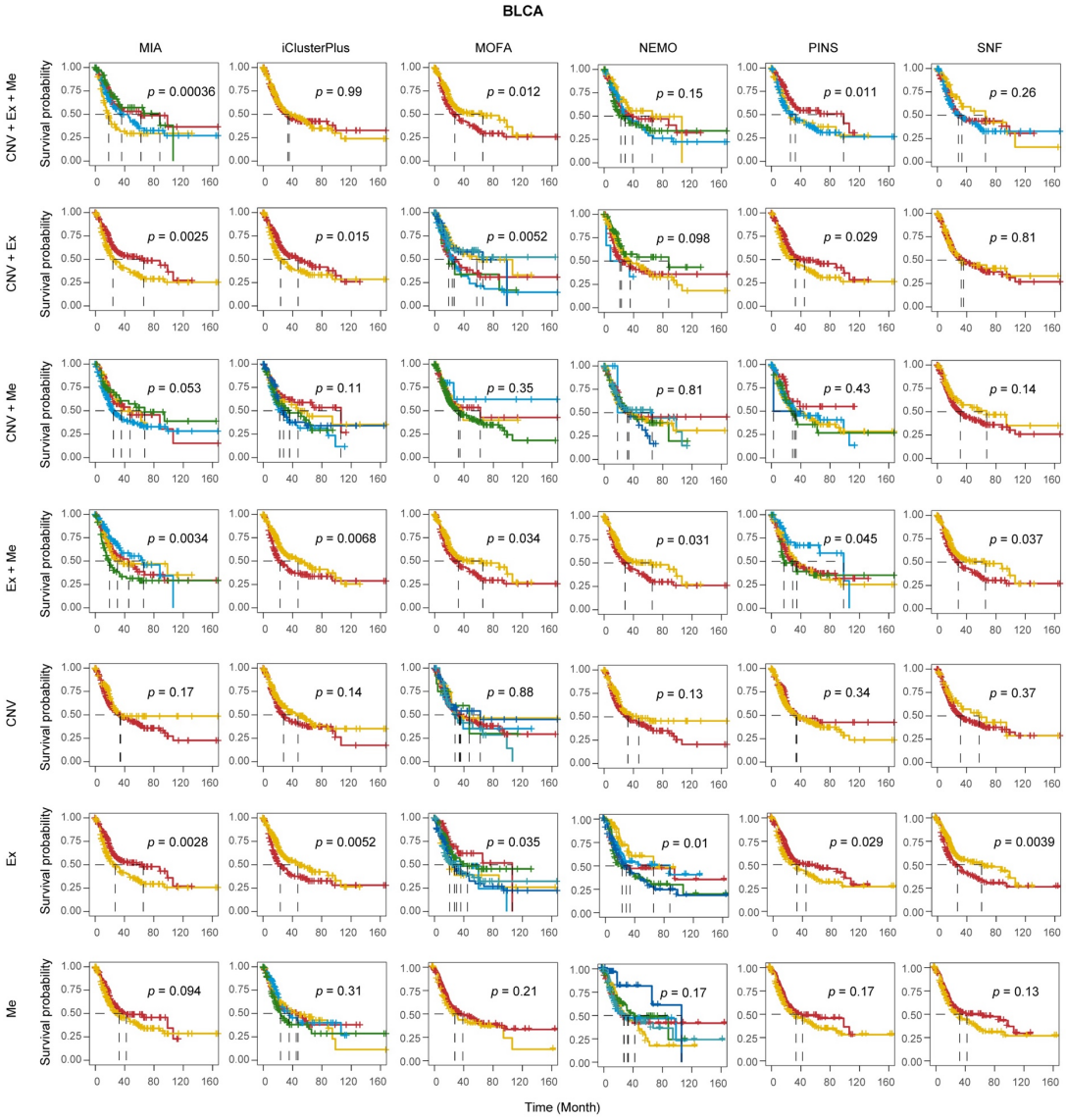

**Supplementary Fig. 8. Kaplan-Meier survival analysis on TCGA GBM dataset clustered by different algorithms across various omics combinations.**

The log-rank test was used to assess statistical significance between survival curves. Abbreviations: CNV, copy number variation; Ex, gene expression; Me, DNA methylation.



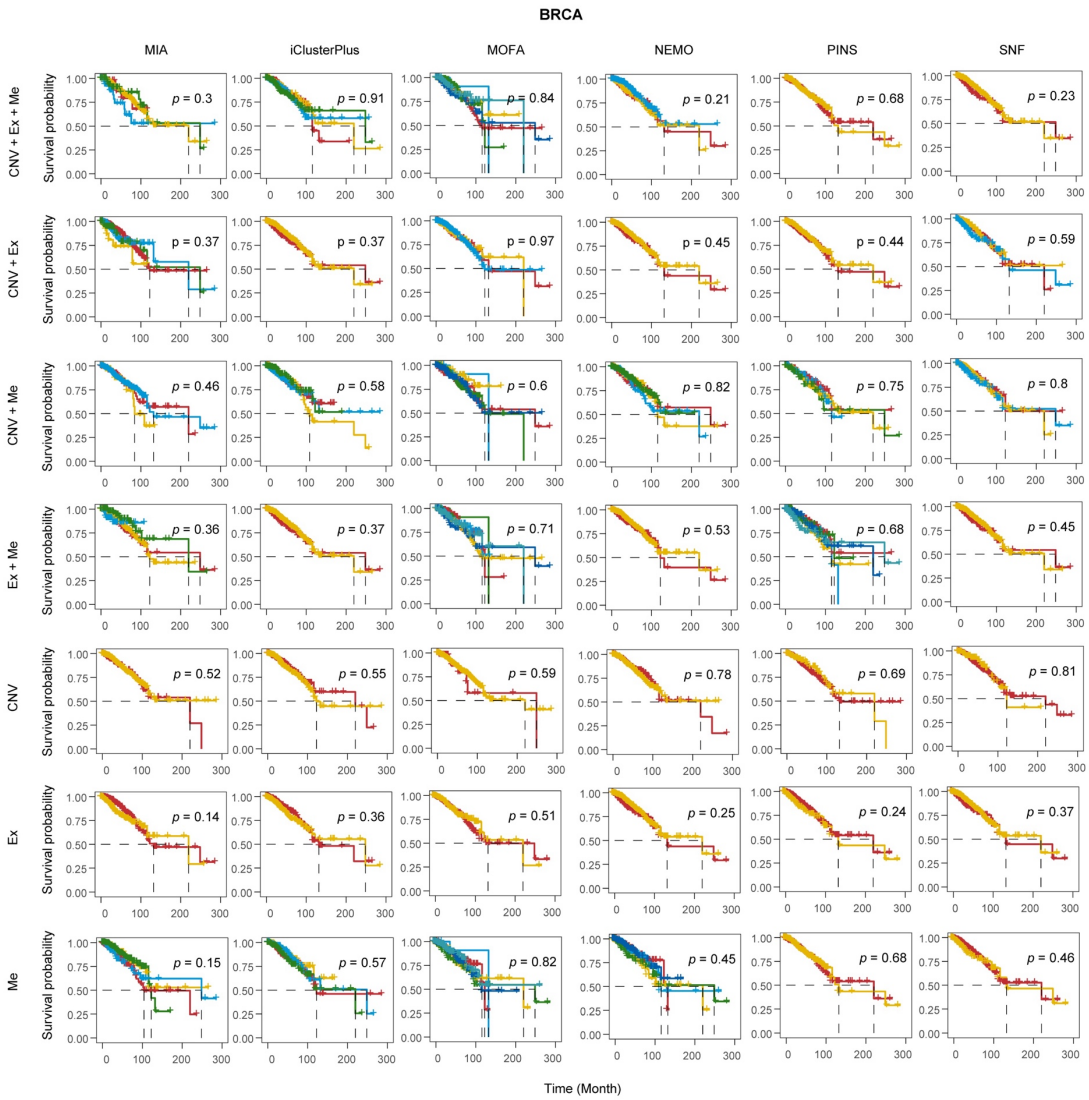

**Supplementary Fig. 9. Kaplan-Meier survival analysis on TCGA LGG dataset clustered by different algorithms across various omics combinations.**

The log-rank test was used to assess statistical significance between survival curves. Abbreviations: CNV, copy number variation; Ex, gene expression; Me, DNA methylation.



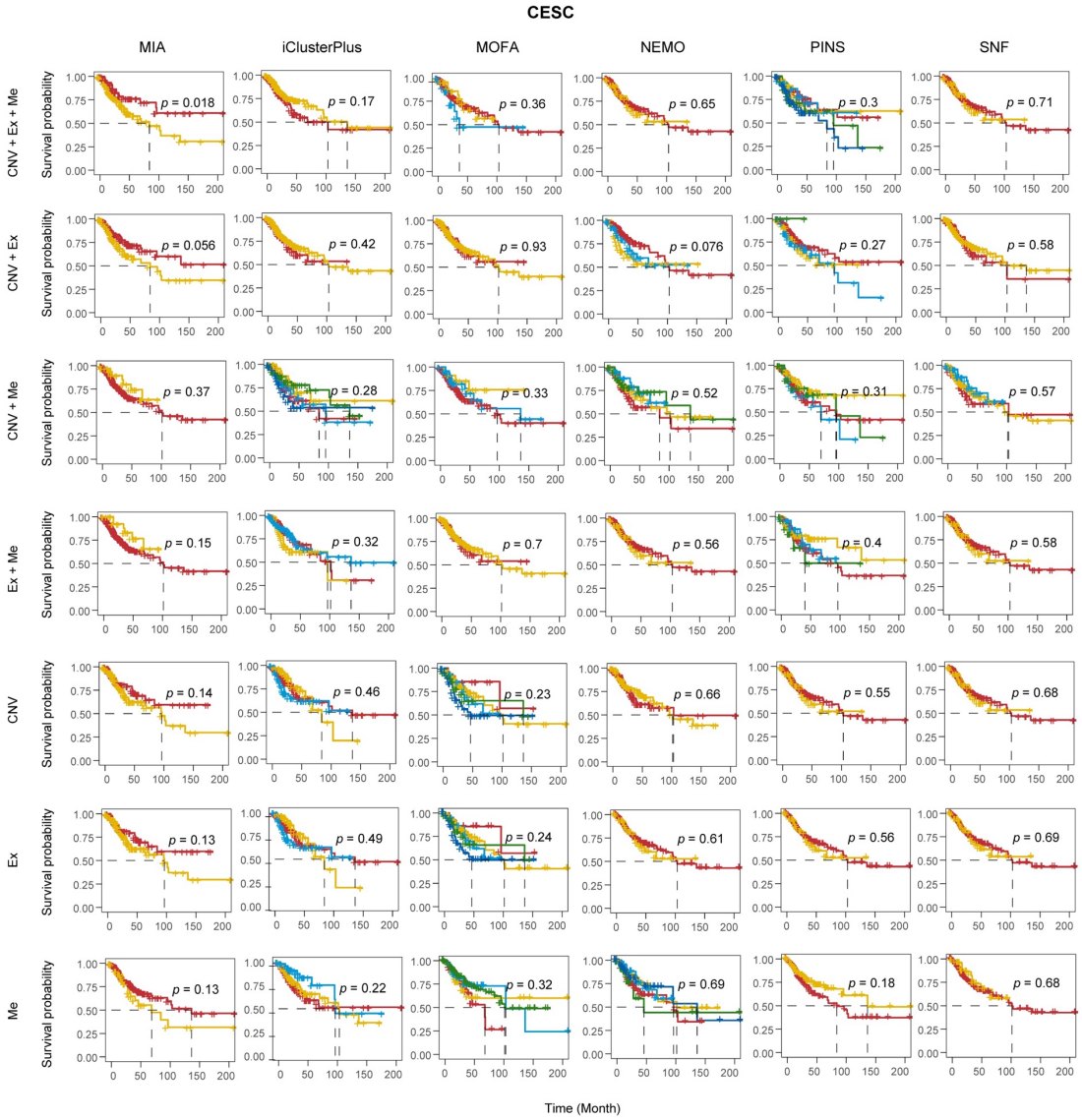

**Supplementary Fig. 10. Kaplan-Meier survival analysis on TCGA BLCA dataset clustered by different algorithms across various omics combinations.**

The log-rank test was used to assess statistical significance between survival curves. Abbreviations: CNV, copy number variation; Ex, gene expression; Me, DNA methylation.



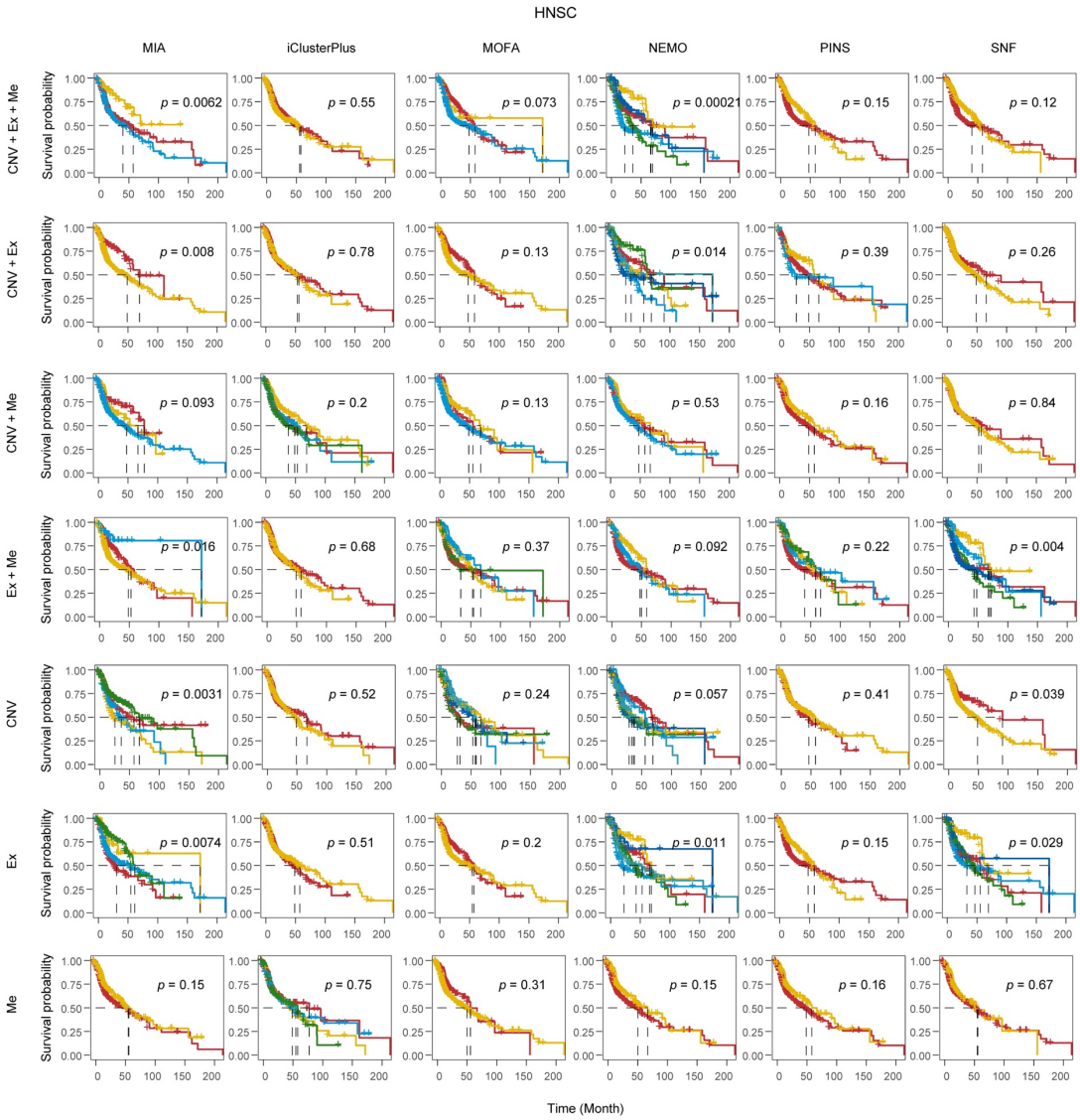

**Supplementary Fig. 11. Kaplan-Meier survival analysis on TCGA BRCA dataset clustered by different algorithms across various omics combinations.**

The log-rank test was used to assess statistical significance between survival curves. Abbreviations: CNV, copy number variation; Ex, gene expression; Me, DNA methylation.



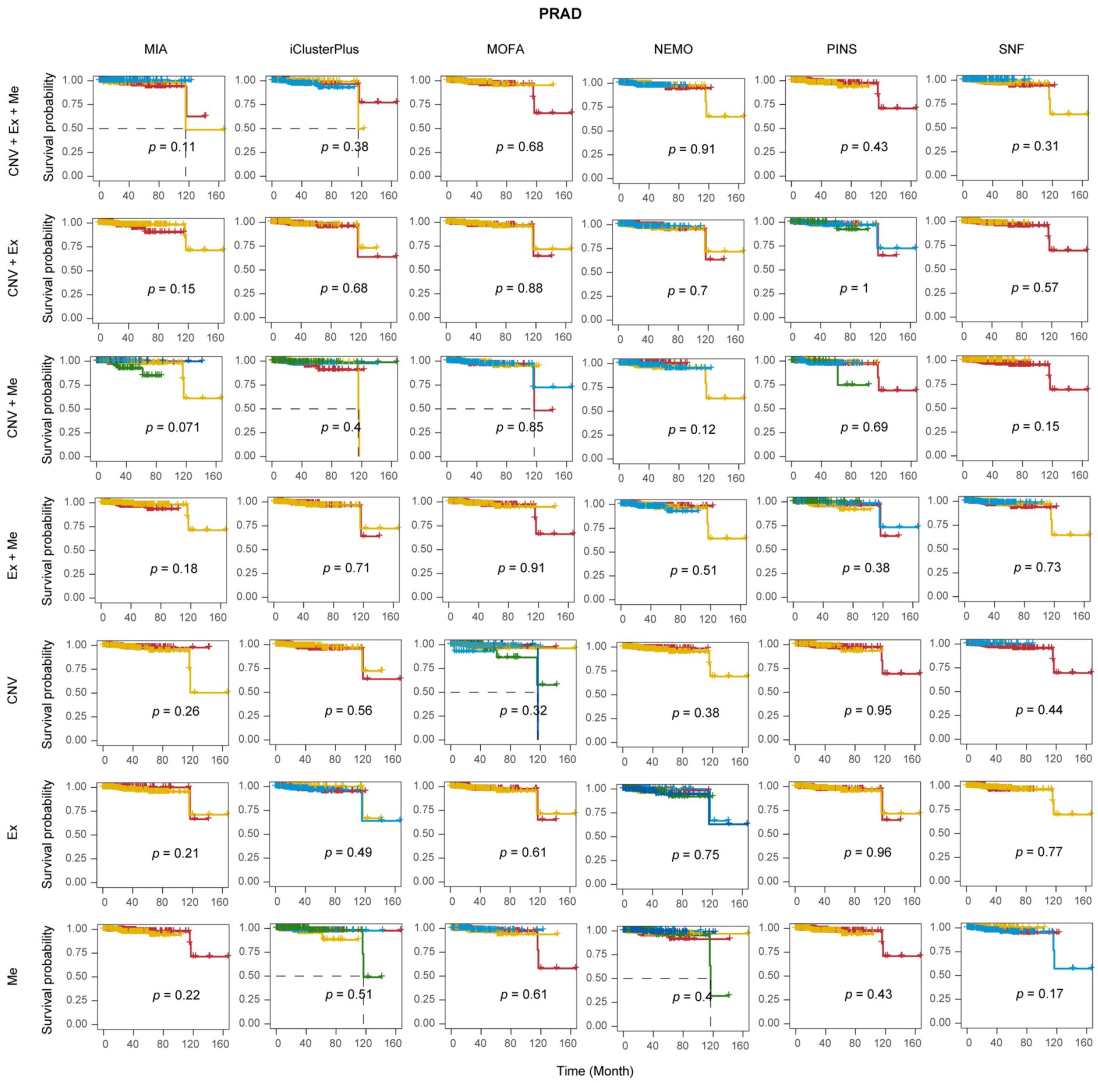

**Supplementary Fig. 12. Kaplan-Meier survival analysis on TCGA CESC dataset clustered by different algorithms across various omics combinations.**

The log-rank test was used to assess statistical significance between survival curves. Abbreviations: CNV, copy number variation; Ex, gene expression; Me, DNA methylation.



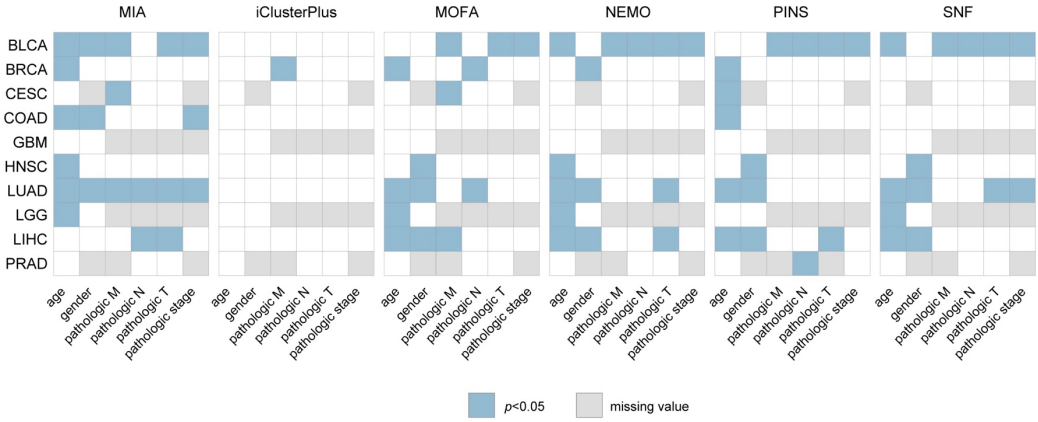

**Supplementary Fig. 13. Kaplan-Meier survival analysis on TCGA HNSC dataset clustered by different algorithms across various omics combinations.**

The log-rank test was used to assess statistical significance between survival curves. Abbreviations: CNV, copy number variation; Ex, gene expression; Me, DNA methylation.



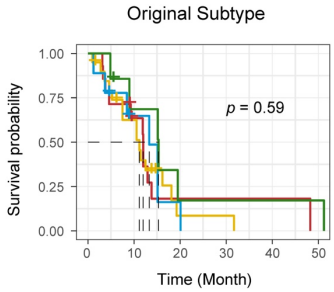

**Supplementary Fig. 14. Kaplan-Meier survival analysis on TCGA PRAD dataset clustered by different algorithms across various omics combinations.**

The log-rank test was used to assess statistical significance between survival curves. Abbreviations: CNV, copy number variation; Ex, gene expression; Me, DNA methylation.



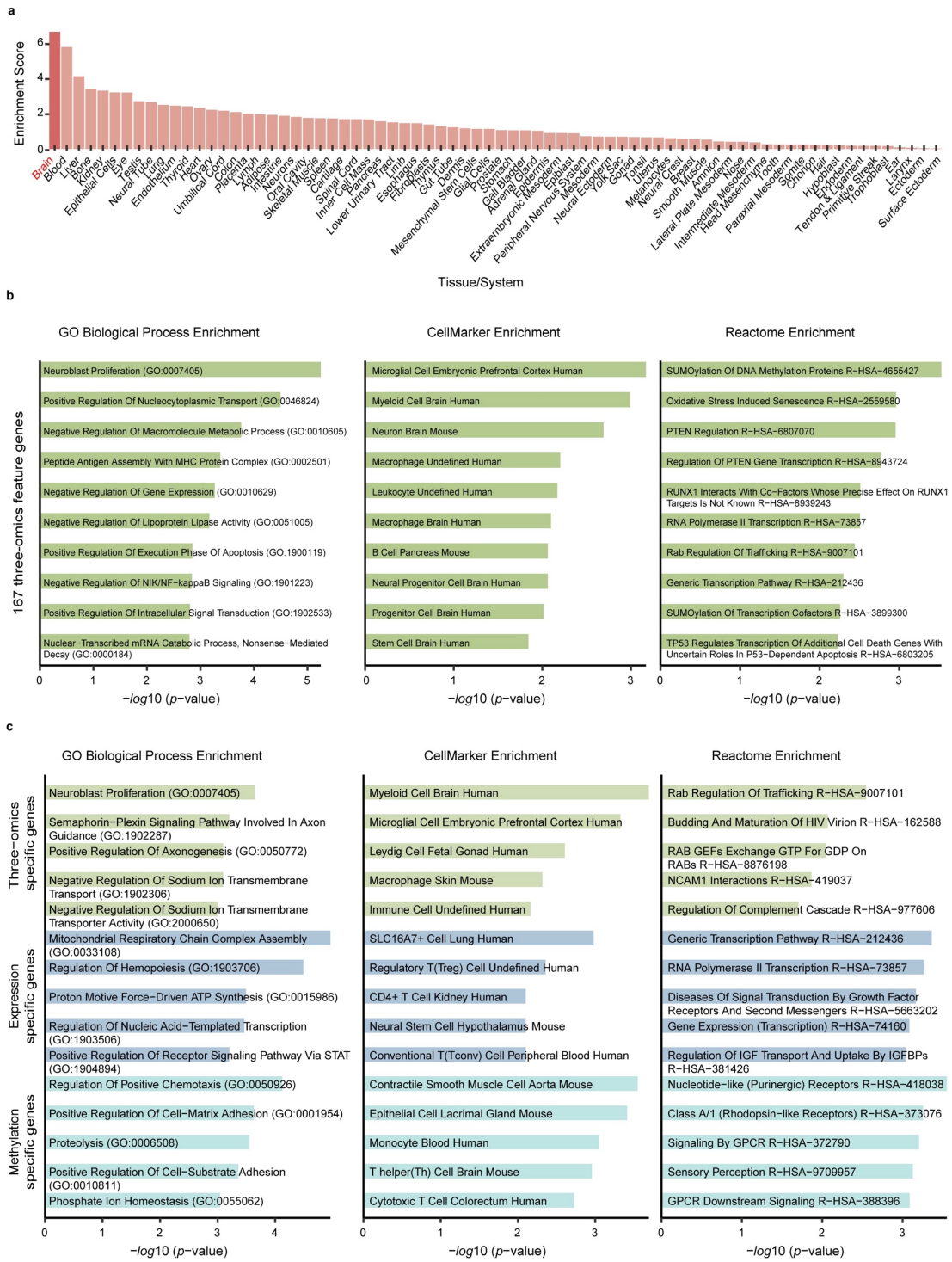

**Supplementary Fig. 15. Enrichment of clinical parameters by different algorithms across ten cancer types.**

Clinical parameters include Pathologic M, metastasis progression; Pathologic N, lymph node involvement; Pathologic T, tumor progression; Pathologic stage, overall disease stage.



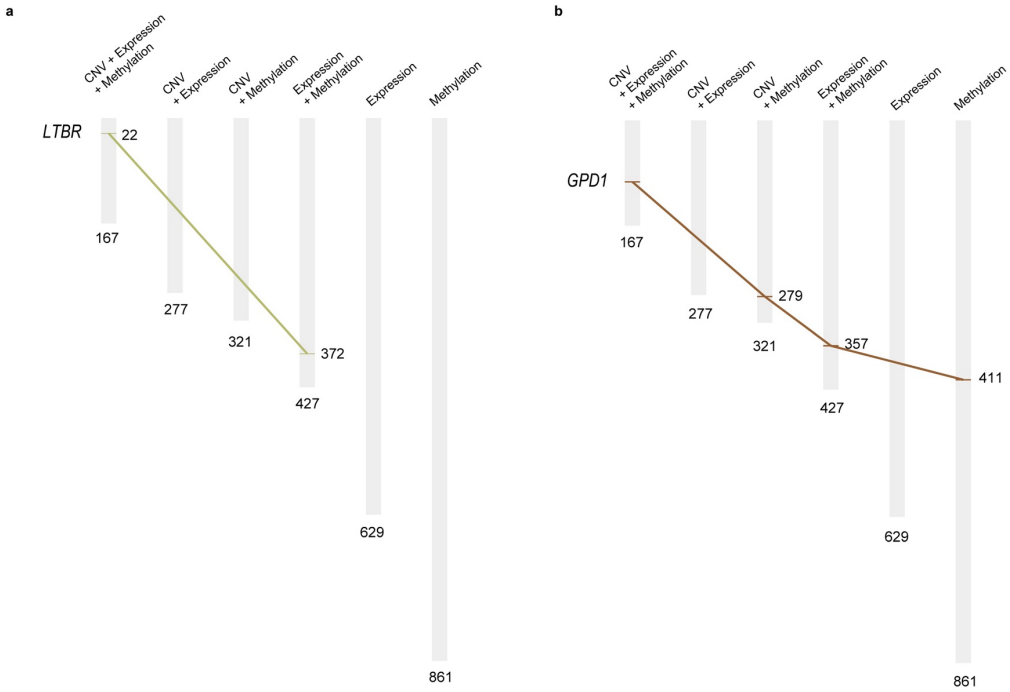

**Supplementary Fig. 16. Kaplan-Meier survival analysis of GBM classical subtypes.**

The log-rank test was used to assess statistical significance between survival curves.




**Supplementary Fig. 17. Functional enrichment of 167 GBM feature genes.**

(a) Enrichment scores of 167 feature genes across tissues/systems. (b) Enrichment analysis of 167 feature genes identified from multi-omics data. (c) Enrichment analysis of feature genes identified specifically from multi-omics, expression, and methylation data, respectively.




**Supplementary Fig. 18. Importance ranking of *LTBR* (a) and *GPD1* (b) identified from different omics combinations.**

Feature genes identified by CNV (not shown) did not reach the threshold.

774 # Supplementary Tables

775 **Supplementary Table 1. Clustering performance evaluation using the Adjusted Rand Index (ARI)**
776 **for different algorithms on simulated datasets.**
777 **Supplementary Table 2. Important features identified by different algorithms and their**
778 **significance on simulated datasets.**
779 **Supplementary Table 3. Optimal *K* values and corresponding *p*-values for survival outcomes**
780 **across different omics combinations on real datasets.**
781 **Supplementary Table 4. Sample clustering results generated by different algorithms on real**
782 **datasets.**
783 **Supplementary Table 5. *P*-values for enrichment of clinical parameters across different algorithm**
784 **clusters in ten cancer types.**
785 **Supplementary Table 6. GBM subtyping results with associated clinical information.**
786 **Supplementary Table 7. Summary of 167 GBM feature genes.**
787 **Supplementary Table 8. Importance ranking of feature genes across different omics combinations.**
788 **Supplementary Table 9. Sample clustering results of 167 feature genes in GSE211272, GSE113510,**
789 **GSE100736 and GSE68029 datasets.**
790 **Supplementary Table 10. Associations between genes and chemotherapy drugs across different**
791 **cancers.**