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

Yang Zhang[1,2], Lin Liu[1,2,§], Zhaoqi Liu[1,2,3], Qi Liu[4], 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*

[4] *Department of Hematology, Tongji Hospital, Frontier Science Center for Stem Cell Research, Bioinformatics Department, School of Life Sciences and Technology, Tongji University, Shanghai 200092, 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 due to the high dimensionality and heterogeneity of multi-omics data. Here we describe MIA, a machine learning framework for multi-omics integrative analysis. Unlike existing algorithms that rely on two-dimensional representations, MIA employs a three-dimensional tensor representation coupled with tensor decomposition, Fuzzy C-Means, and an enhanced random forest model to jointly enable accurate sample stratification and feature discovery. Benchmarking on simulated datasets demonstrates that MIA achieves higher accuracy in both clustering and feature identification by comparison with extant algorithms. Application to TCGA datasets further shows the ability of MIA to stratify samples and identify features associated with significantly clinical outcomes. Notably, as applied to glioblastoma, MIA is capable to detect three previously uncharacterized subtypes with distinct prognostic profiles and uncover critical feature genes linked to glioblastoma subtyping and therapeutic response. Collectively, these results establish MIA as a generalizable computational 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; machine learning; tensor; clustering

Introduction

Powered by high-throughput sequencing technologies, data from multiple omics layers, such as genomics, transcriptomics, proteomics, and epigenomics, have 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 networks across different omics types, but are sensitive to data perturbations, outliers and noise, which may reduce robustness and increase 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 factorize 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 analyses using both simulated and empirical datasets, and apply it to 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.

To analyze unlabeled biomedical samples, MIA applies tensor decomposition for dimensionality reduction, followed by FCM clustering for sample stratification. Among available tensor decomposition 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 factorize the tensor into three matrices: sample-feature matrix, gene-feature matrix, and omics type-feature matrix, and then employs FCM for clustering based on the sample-feature matrix (**Fig. 1b** and **Supplementary Fig. 1**). Unlike *K*-means, FCM allows samples to be classified into 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 details in Methods; **Supplementary Figs. 2 and 3**).

Using the clustering results as sample labels, MIA further utilizes the Gini index to perform preliminary 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 parameter 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.

When varying K and T with fixed N, M_t and M_f, MIA overall achieves the stable and higher accuracy across all parameter 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 identified by the other algorithms (**Fig. 2d and Supplementary Table 2**).

Benchmarking of MIA on TCGA datasets

We evaluate MIA in ten empirical 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 differences across different omics data combinations (**Fig. 3a-c** and **Supplementary Table**

144 **3**), particularly for COAD and LUAD (**Supplementary Figs. 5 and 6**). Compared to all possible
145 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 striking than that of other algorithms. Particularly, MIA achieves significantly different
148 survival outcomes (p<0.0001, Log-rank test) in LIHC and GBM based on three-omics data, whereas
149 single-omics or two-omics combinations fail to achieve similar differences (**Supplementary Figs.**
150 **7 and 8**). Additionally, we observe that although nearly all algorithms exhibit comparable survival
151 differences in LGG on multi-omics combinations, clusters determined by MIA show more
152 divergences (**Supplementary Fig. 9**). Detailed survival analyses for all these ten cancer types
153 (**Supplementary Figs. 5-14**) show that multi-omics data integrative analysis by MIA is superior to
154 other algorithms in precise clustering.

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

168 We further assess the MIA performance of feature gene identification in seven cancer types with
169 sufficient samples. Compared with existing algorithms, feature genes identified by MIA present
170 greater significant differences between cancer and normal samples across nearly all examined
171 cancer types (**Fig. 3e**). Clearly, this result demonstrates the superior capability of MIA in feature
172 gene 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 index signifies the stem cell-like characteristics of cancer cells[33], together
186 indicating tumor aggressiveness and prognosis. Thus, we further evaluate the MIA clustering in
187 GBM based on these metrics. We find that the fraction of altered genome decreases significantly
188 from cluster 1 to cluster 3 ($p<0.05$; **Fig. 4d**), corresponding well with the declining survival rates.
189 Consistently, the stemness index gradually reduce, namely, higher in cluster 1, medium in cluster 2,
190 and lower in cluster 3 (**Fig. 4e** and **Supplementary Table 6**). These results indicate that GBM with
191 fewer genomic alterations may exhibit more aggressive behavior and lower survival rates, consistent
192 with observations in advanced tumors[32]. In addition, GBM with poor prognosis has lower stemness
193 and higher cancer cell differentiation[34, 35]. Overall, the clinical analysis demonstrates that MIA is
194 effective to identify GBM subtypes with distinct aggressiveness and prognosis.

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

207 We also use MIA to identify feature genes based on single-omics and two-omics data. Among these
208 167 feature genes, 96 (57.5%) and 43 (25.7%) genes are specifically identified based on three-omics
209 data, by comparison against single-omics and two-omics data, respectively (**Fig. 5b,c**). Functional
210 enrichment analysis shows that these feature genes specifically identified by three-omics data
211 exhibit stronger associations with brain tissue and tumor development (**Supplementary Fig. 17c**).
212 We further analyze the clustering power of feature genes specifically identified by different omics
213 combinations. We observe that feature genes identified from single-omics data exhibit strong
214 significance within their respective omics layers, but feature genes from three-omics always show
215 higher clustering scores, clearly indicating that multi-omics integrative analysis achieves more
216 enhanced clustering power (**Fig. 5d,e**). Notably, several literature-reported GBM marker genes,
217 such as *VGF* (promoting GBM stem cell survival and tumor growth[36]) and *BRD8* (hindering *p53*
218 transcriptional activation and promoting GBM cell proliferation[37]), are successfully identified by
219 three-omics data integrative analysis, yet absent in single-omics or two-omics data analysis
220 (**Supplementary Table 8**). It is worth mentioning that based on three-omics integrative analysis,
221 *LTBR* (activating anti-tumor immune response in glioma[38]) and *GPD1* (whose inhibition prolongs
222 GBM mouse survival[39]) are computationally identified as feature genes with markedly enhanced

223 priority (**Supplementary Fig. 18**), thus highlighting the great potential of MIA to uncover
224 biologically significant features from multi-omics data and thereby guiding in-depth experimental
225 validation and mechanism investigation.

226 Characterization of Temozolomide treatment sensitivity through GBM feature 227 genes

228 Temozolomide (TMZ) is one of the most effective chemotherapy agents for GBM treatment[29].
229 Based on these 167 feature genes identified by MIA, we further explore whether these genes can be
230 used to predict TMZ chemotherapy response. Toward this end, firstly, we collect a publicly
231 available dataset (expression data only) with GBM TMZ treatment, containing samples of control,
232 TMZ + radiotherapy, and TMZ + radiotherapy + inhibitor[40]. MIA clustering based on the 167 gene
233 expression profiles separates the samples into three clusters that correspond closely to the three
234 treatment groups. We find that the control samples are most enriched in cluster 1, the TMZ +
235 radiotherapy samples are most in cluster 2, and the TMZ + radiotherapy + inhibitor samples are
236 exclusively included in cluster 3 (**Fig. 6a** and **Supplementary Table 9**). A multi-class ROC analysis
237 yields an average AUC of 0.85, indicating that these genes effectively distinguish different TMZ
238 treatment regimens. (**Fig. 6b**). To further assess the relevance of these genes to TMZ treatment
239 sensitivity, we collect three independent datasets containing TMZ-resistant and TMZ-sensitive
240 GBM cell-line samples. For these three datasets, clustering results align perfectly with the outcomes
241 of TMZ treatment (**Fig. 6c-e** and **Supplementary Table 9**), implying that these 167 genes are
242 closely associated with TMZ chemotherapy response.

243 To systematically identify genes associated with TMZ chemotherapy sensitivity, we discover that
244 64 genes display consistent expression patterns in the above datasets. We collect these 64 genes'
245 associations with chemotherapy drugs across different cancers (**Supplementary Table 10)**. Notably,
246 ten genes (*GPD1*, *MSI1*, *IFFO1*, *NCBP2*, *SGTA*, *BCL9L*, *FOXP4*, *KCP*, *MAPK13*, and *PAFAH1B1*)
247 have already been reported to be related to chemotherapy sensitivity (**Fig. 6f**). Specifically, *GPD1*
248 is specifically expressed in brain tumor stem cells and involved in tumor initiation, metastasis, and
249 recurrence; increased expression of *GPD1* correlates with TMZ resistance in GBM[39]. *MSI1* is a
250 stem cell gene that controls the balance between self-renewal and differentiation; GBM lines with
251 *MSI1* knockout display increased sensitivity to cell cycle and DNA replication inhibitors[41]. The rest
252 eight genes, albeit unexplored with GBM chemotherapy, associate closely with gynecological
253 cancers[42-44], LIHC[45], gastrointestinal cancers[46, 47], hematological malignancies[48], and other
254 cancers[49], presumably indicating their potential relevance in GBM treatment.

255 Discussion

256 In this study, we presented MIA, a powerful framework for multi-omics integrative analysis capable
257 of accurate sample clustering and feature gene identification. MIA utilizes various machine learning
258 techniques, including Tucker decomposition, FCM, and ERFM, to effectively integrate and analyze
259 multi-omics data and identify biologically important features contributing to sample clustering.

Comparative analyses on simulated and real datasets revealed that MIA outperforms existing algorithms in both integrative clustering and feature identification. 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[50], together exerting an influence on protein production, cell function, and thereby, phenotypic and clinical outcomes[51]. 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[52].

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[53, 54], and is also a sign of poor prognosis[55, 56]. On the contrary, in advanced tumors, several studies have also found that tumors with lower genomic alterations can exhibit more aggressive behavior and decreased survival rates[32]. For example, in metastases compared to primary lung adenocarcinoma, the fraction of genome altered was significantly lower[57]. Additionally, excessive genomic instability may cumulate damage, exceeding the threshold of tumor cell activity, thereby inhibiting advanced tumor growth and

301 metastasis[58]. Our observations in GBM are consistent with these findings of advanced tumors, but
302 more research is needed to explore the association between this indicator and the prognosis of
303 advanced tumors.

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

308 Methods

309 Simulated datasets

310 Simulated multi-omics datasets were generated by a Python script. Several parameters—number of
311 clusters (K), number of omics types (T), sample size (N), and number of real features (M_t), and
312 number of noise features (M_f)—were considered and five replicates were produced for each
313 parameter combination. Samples were assigned to K clusters using a Gaussian probability
314 distribution. Each omics layer was represented as an $N \times (M_t + M_f)$ matrix, where M_t features were
315 differentially expressed across clusters with group-specific means and Gaussian noise and M_f
316 features were randomly generated. For clustering evaluation, we considered the effects of K and T
317 (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
318 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,
319 with fixed K=4 and T=2). For feature gene identification, datasets were simulated with K=4 and 5,
320 N=500, M_t=1,000, and M_f=10,000. The code for dataset simulation is provided in
321 *simulate_datasets.py* within the MIA package (available at https://ngdc.cncb.ac.cn/mia/download).

322 The TCGA datasets

323 Multi-omics datasets for ten TCGA cancer types, including CNV (gene-level), gene expression
324 (FPKM), and DNA methylation data, were downloaded from UCSC Xena
325 (https://xenabrowser.net/). For DNA methylation data, the methylation level of each gene was
326 determined by averaging methylation levels of all sites within the first 1,500bp upstream of the gene.
327 For integrative analysis, only samples and genes shared across all three omics types were retained
328 for each cancer type. Survival data and clinical information were obtained from cBioPortal
329 (https://www.cbioportal.org/). Tumor stemness index (mRNA expression-based stemness index,
330 mRNAsi) were adopted by referencing a previous study[33]. Feature genes identified for GBM were
331 annotated by PubMed (https://pubmed.ncbi.nlm.nih.gov/), BrainBase[59], MSigDB[60], and EWAS[61].

332 GBM temozolomide treatment datasets

333 Gene expression datasets related to GBM temozolomide (TMZ) treatment were retrieved from the
334 Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/gds/). Specifically, GSE211272[40]

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[62] 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[63] 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[64] 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.

Tensor integration and decomposition

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 factorize 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"[65] (http://tensorly.org/stable/index.html).

FCM clustering and parameter optimization

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^(-0.0406ln(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[66], Davies-Bouldin Index[67], Elbow method[68], 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 defined by clustering results or sample labels. 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.

Feature 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 $M_{\leqslant 0}$.

(e) Iterative computational procedure: To enhance consistency of the identified feature gene set, an iterative procedure is applied with a termination condition: if $M_{\leqslant 0}>0$, these $M_{\leqslant 0}$ genes are removed, and $M'=M'-M_{\leqslant 0}$ is updated to repeat **steps (a-d)**; if $M_{\leqslant 0}=0$, 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**).

Clustering evaluation

To benchmark the performance of MIA, we compared it with existing algorithms using their default or recommended parameter settings. All algorithms were applied to the same preprocessed multi-omics datasets, and clustering results were evaluated using a unified evaluation framework to ensure consistency and comparability. For simulated datasets, clustering accuracy was assessed using the ARI, which measures concordance between predicted and true labels on a scale from -1 to 1, with higher values indicating higher accuracy. For TCGA empirical datasets, we evaluated both survival and clinical relevance of the derived clusters. Survival differences were assessed using Kaplan-Meier estimation and log-rank testing. Significances between clusters and clinical records were examined using chi-square tests for categorical variables and Kruskal-Wallis tests for continuous variables, with empirical significance estimated via permutation to obtain confidence intervals, following established procedures[12]. For TMZ response datasets, clustering accuracy was quantified

using multi-class ROC analysis. Each cluster was evaluated in a one-vs-rest framework, ROC curves were generated for all classes, and overall performance was summarized using the macro-averaged area under the curve (AUC).

Feature gene evaluation

For each gene, the Kruskal-Wallis test was employed to quantify its discriminative ability among groups by using true labels in simulated datasets and identified clusters in empirical datasets. Based on this test, significance level was obtained at each omics layer (e.g., CNV, expression, methylation). For any multi-omics dataset, significance level was averaged across its corresponding omics layers. The clustering power of a feature-gene set was quantified by the Kruskal-Wallis test and the 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. 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, Z. et al. Machine learning revealed stemness features and a novel stemness-based classification with appealing implications in discriminating the prognosis, immunotherapy and temozolomide responses of 906 glioblastoma patients. *Brief Bioinform* **22**, bbab032 (2021).
36. Wang, X. et al. Reciprocal signaling between glioblastoma stem cells and differentiated tumor dells promotes malignant progression. *Cell Stem Cell* **22**, 514-528 (2018).
37. Sun, X. et al. BRD8 maintains glioblastoma by epigenetic reprogramming of the p53 network. *Nature* **613**, 195-202 (2023).
38. Tang, Q. et al. Comprehensive analysis reveals that LTBR is a immune-related biomarker for glioma. *Computers in Biology and Medicine* **174**, 108457 (2024).
39. Rusu, P. et al. GPD1 specifically marks dormant glioma stem cells with a distinct metabolic profile. *Cell Stem Cell* **25**, 241-257 (2019).
40. 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).
41. 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).
42. 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).
43. 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).
44. 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).
45. Zhang, L. et al. NRBP2 Overexpression Increases the Chemosensitivity of Hepatocellular Carcinoma Cells via Akt Signaling. *Cancer Research* **76**, 7059-7071 (2016).
46. 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).

557 47. Liu, X. et al. FOXP4 Is a Direct YAP1 Target That Promotes Gastric Cancer Stemness and Drives
558 Metastasis. *Cancer Research* **84**, 3574-3588 (2024).
559 48. Wang, Y. et al. Expression of small glutamine-rich TPR-containing protein A (SGTA) in Non-
560 Hodgkin's Lymphomas promotes tumor proliferation and reverses cell adhesion-mediated drug
561 resistance (CAM-DR). *Leuk Res* **38**, 955-963 (2014).
562 49. Kim, J. et al. MAPK13 stabilization via m(6)A mRNA modification limits anticancer efficacy of
563 rapamycin. *J Biol Chem* **299**, 105175 (2023).
564 50. You, J.S. & Jones, P.A. Cancer genetics and epigenetics: two sides of the same coin? *Cancer Cell*
565 **22**, 9-20 (2012).
566 51. Claussen, H., Ghosh, S. & Chen, J. Exploring combined effects of DNA methylation and copy
567 number on gene expression with a two-stage approach. *IEEE/ACM Transactions on Computational*
568 *Biology and Bioinformatics* **21**, 1667-1675 (2024).
569 52. Brat, D.J. et al. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *New*
570 *England Journal of Medicine* **372**, 2481-2498 (2015).
571 53. Hurst, C.D. et al. Stage-stratified molecular profiling of non-muscle-invasive bladder cancer
572 enhances biological, clinical, and therapeutic insight. *Cell Reports Medicine* **2**, 100472 (2021).
573 54. Chakraborty, G. et al. Fraction genome altered (FGA) to regulate both cell autonomous and non-
574 cell autonomous functions in prostate cancer and its effect on prostate cancer aggressiveness.
575 *Journal of Clinical Oncology* **38**, 347 (2020).
576 55. Zhou, W. et al. Counting alleles to predict recurrence of early-stage colorectal cancers. *Lancet* **359**,
577 219-225 (2002).
578 56. Choi, S.W. et al. Genetic classification of colorectal cancer based on chromosomal loss and
579 microsatellite instability predicts survival. *Clinical Cancer Research* **8**, 2311-2322 (2002).
580 57. Zhao, R. et al. Clonal dynamics and Stereo-seq resolve origin and phenotypic plasticity of
581 adenosquamous carcinoma. *Npj Precision Oncology* **7**, 80 (2023).
582 58. Cahill, D.P., Kinzler, K.W., Vogelstein, B. & Lengauer, C. Genetic instability and darwinian
583 selection in tumours. *Trends in Cell Biology* **9**, M57-M60 (1999).
584 59. Liu, L. et al. BrainBase: a curated knowledgebase for brain diseases. *Nucleic Acids Research* **50**,
585 D1131-D1138 (2022).
586 60. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell*
587 *Systems* **1**, 417-425 (2015).
588 61. Li, M. et al. EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. *Nucleic*
589 *Acids Res* **47**, D983-d988 (2019).
590 62. Wu, P. et al. Lnc-TALC promotes O(6)-methylguanine-DNA methyltransferase expression via
591 regulating the c-Met pathway by competitively binding with miR-20b-3p. *Nature Communications*
592 **10**, 2045 (2019).
593 63. Tso, J.L. et al. Bone morphogenetic protein 7 sensitizes O6-methylguanine methyltransferase
594 expressing-glioblastoma stem cells to clinically relevant dose of temozolomide. *Molecular Cancer*
595 **14**, 189 (2015).
596 64. Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E. & Storey, J.D. The sva package for removing
597 batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882-
598 883 (2012).

65. Kossaifi, J., Panagakis, Y., Anandkumar, A. & Pantic, M. TensorLy: Tensor Learning in Python. *Journal of Machine Learning Research* **20**, 1-6 (2019).
66. Zhu, L., Ma, B. & Zhao, X. Clustering validity analysis based on silhouette coefficient. *Journal of Computer Applications* **30**, 139-141 (2010).
67. Xiao, J., Lu, J. & Li, X. Davies Bouldin Index based hierarchical initialization K-means. *Intelligent Data Analysis* **21**, 1327-1338 (2017).
68. 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).

Author contributions

Yang Zhang: Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Validation, Writing - original draft, Writing - review & editing; Lin Liu: Data curation; Zhaoqi Liu: Writing - review & editing; Qi Liu: Writing - review & editing; 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

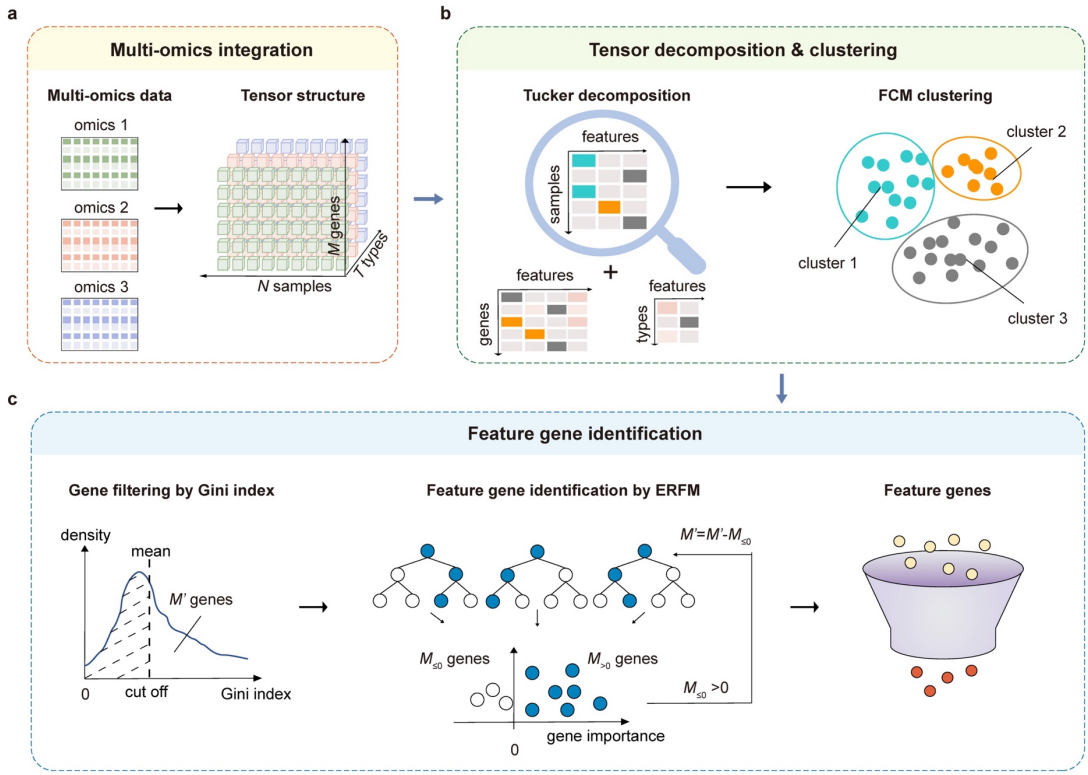


Fig. 1. Schematic of the MIA framework.

(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) Tucker decomposition is used to factorize the tensor into three two-dimensional factor matrices: sample-feature matrix, gene-feature matrix, and type-feature matrix. Subsequently, the FCM clustering is used to cluster the sample-feature matrix. (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 quantify the contribution of each gene in clustering and thus to yield a set of feature genes (see details in Methods).

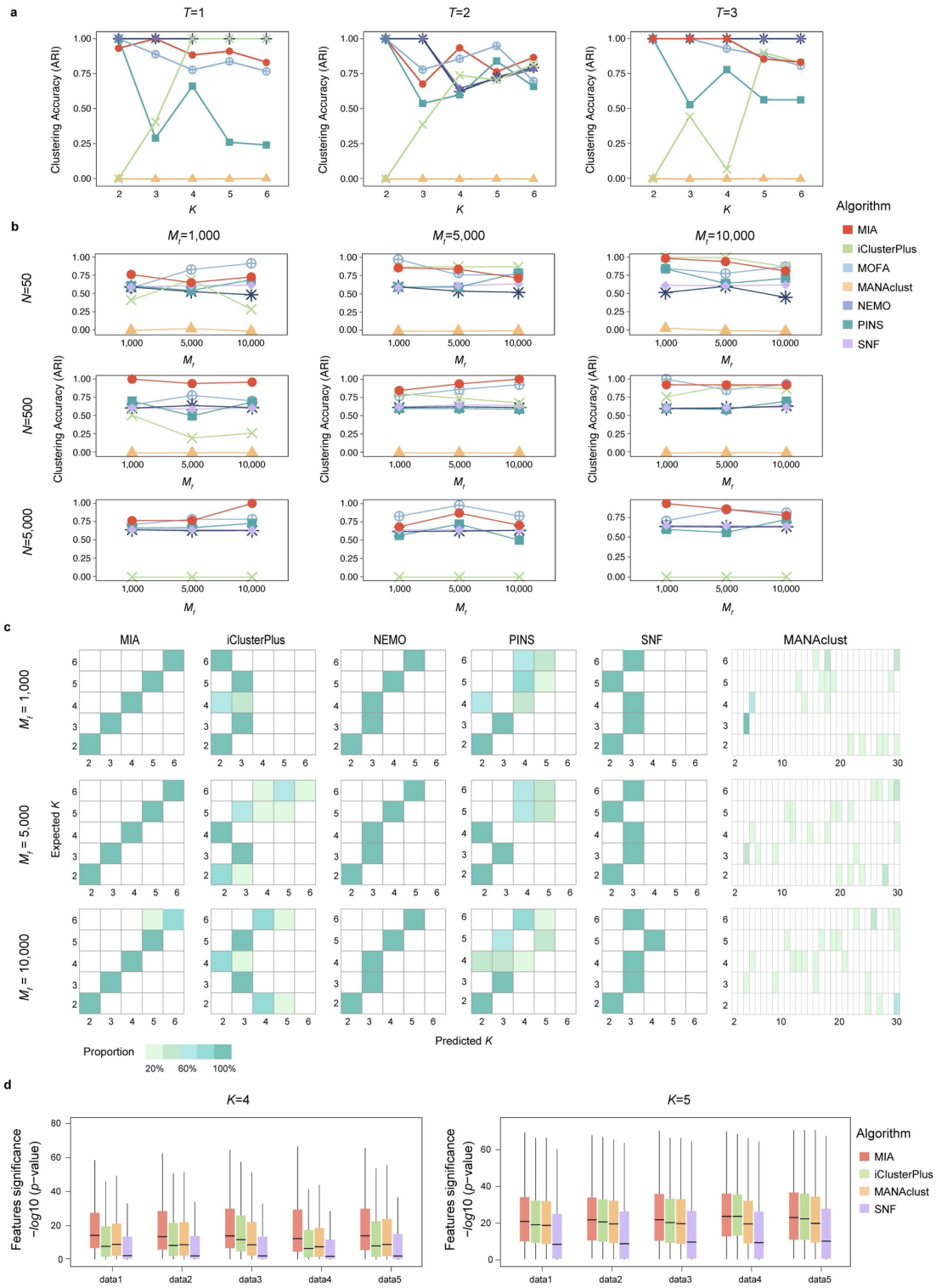


Fig. 2. Performance of MIA in clustering and feature gene identification on simulated datasets by comparison against several existing algorithms.

(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_t=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_t=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.

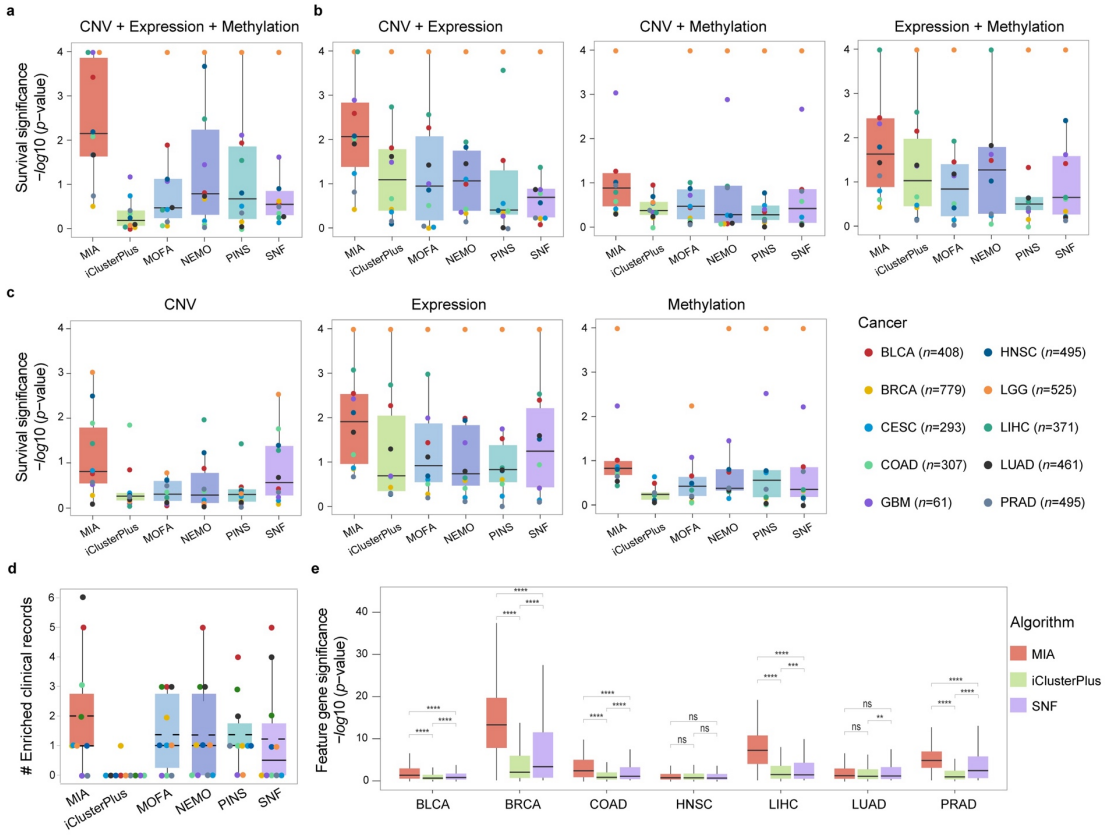


Fig. 3. Evaluation of survival outcomes and feature genes determined by MIA on TCGA datasets in comparison with existing algorithms.

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 records (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).

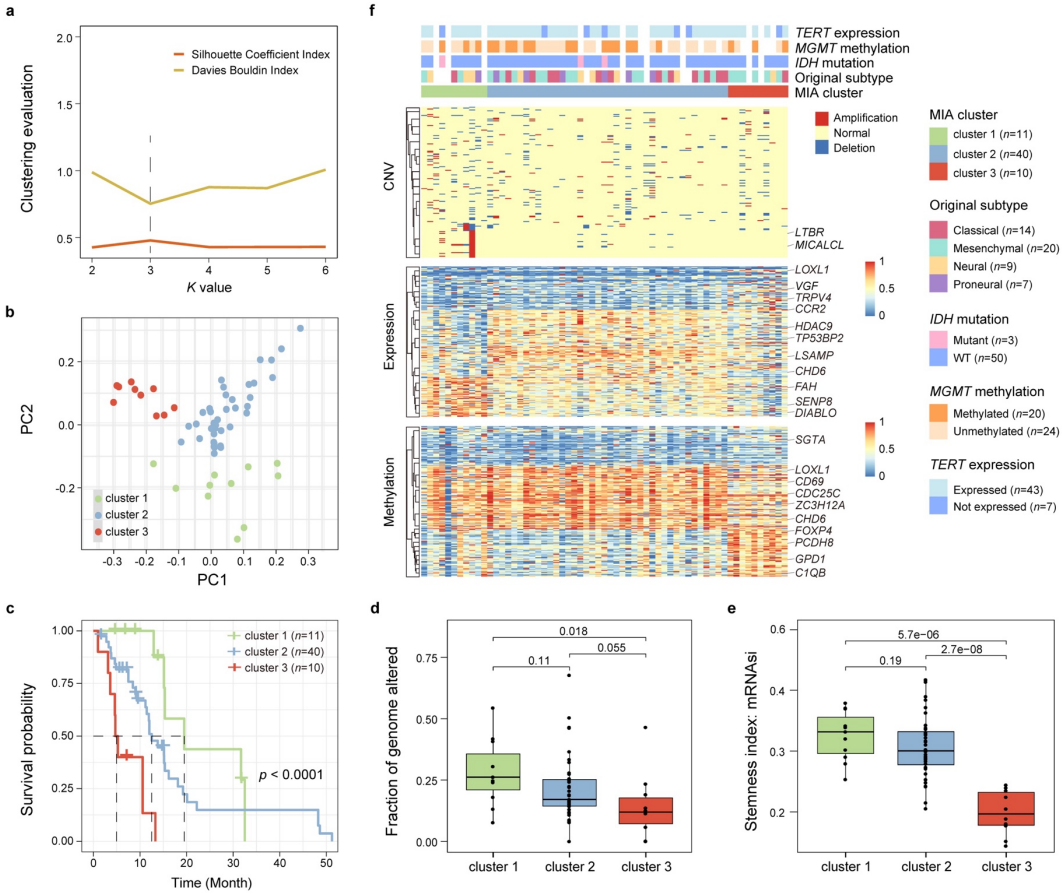


Fig. 4. TCGA GBM subtyping analysis by MIA.

(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 index (mRNA expression-based stemness index, mRNAsi) across the three clusters, with statistical significance assessed by the Wilcoxon rank-sum test. (f) Heatmaps of 167 GBM feature genes identified by MIA.

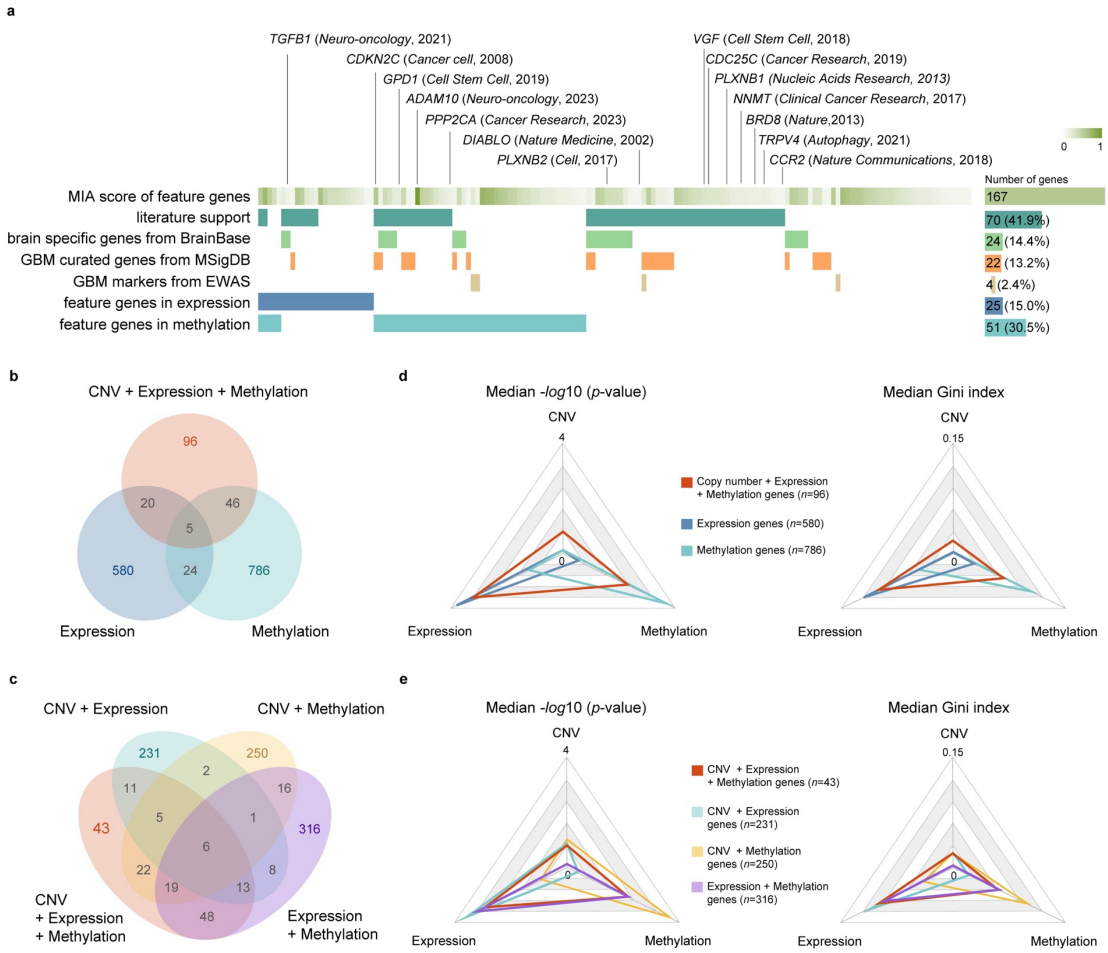


Fig. 5. Characterization of 167 GBM feature genes identified by MIA.

(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.

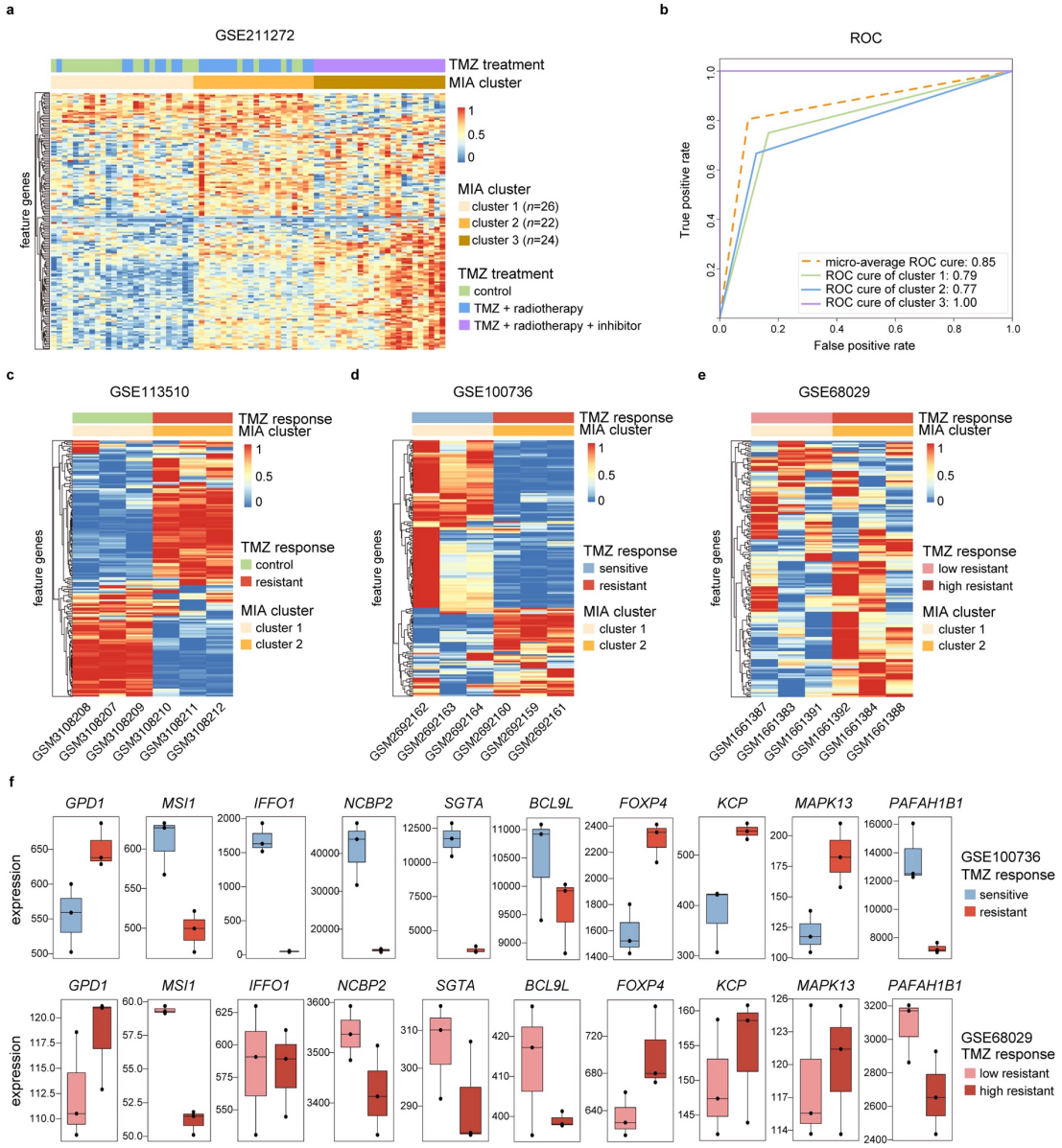
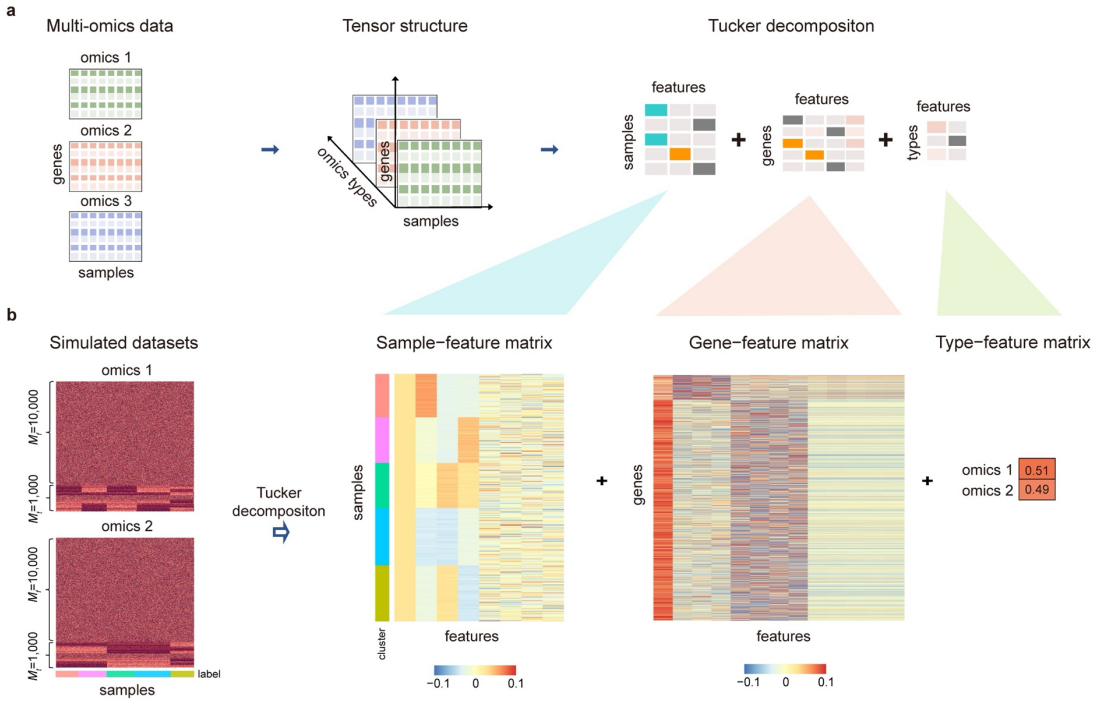


Fig. 6. MIA Clustering of 167 GBM feature genes in TMZ treatment datasets.

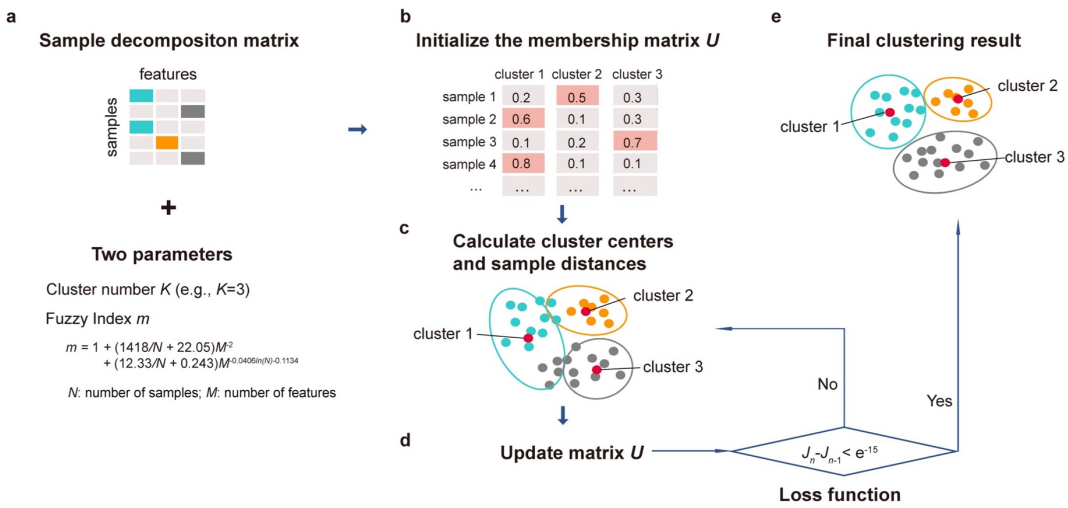
(a) MIA Clustering of 167 GBM feature genes in GSE211272, including control, TMZ + radiotherapy, and TMZ + radiotherapy + inhibitor samples. (b) Multi-class ROC curve illustrating clustering performance on GSE211272. (c-e) MIA Clustering of 167 GBM feature genes in three TMZ treatment datasets, viz., 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

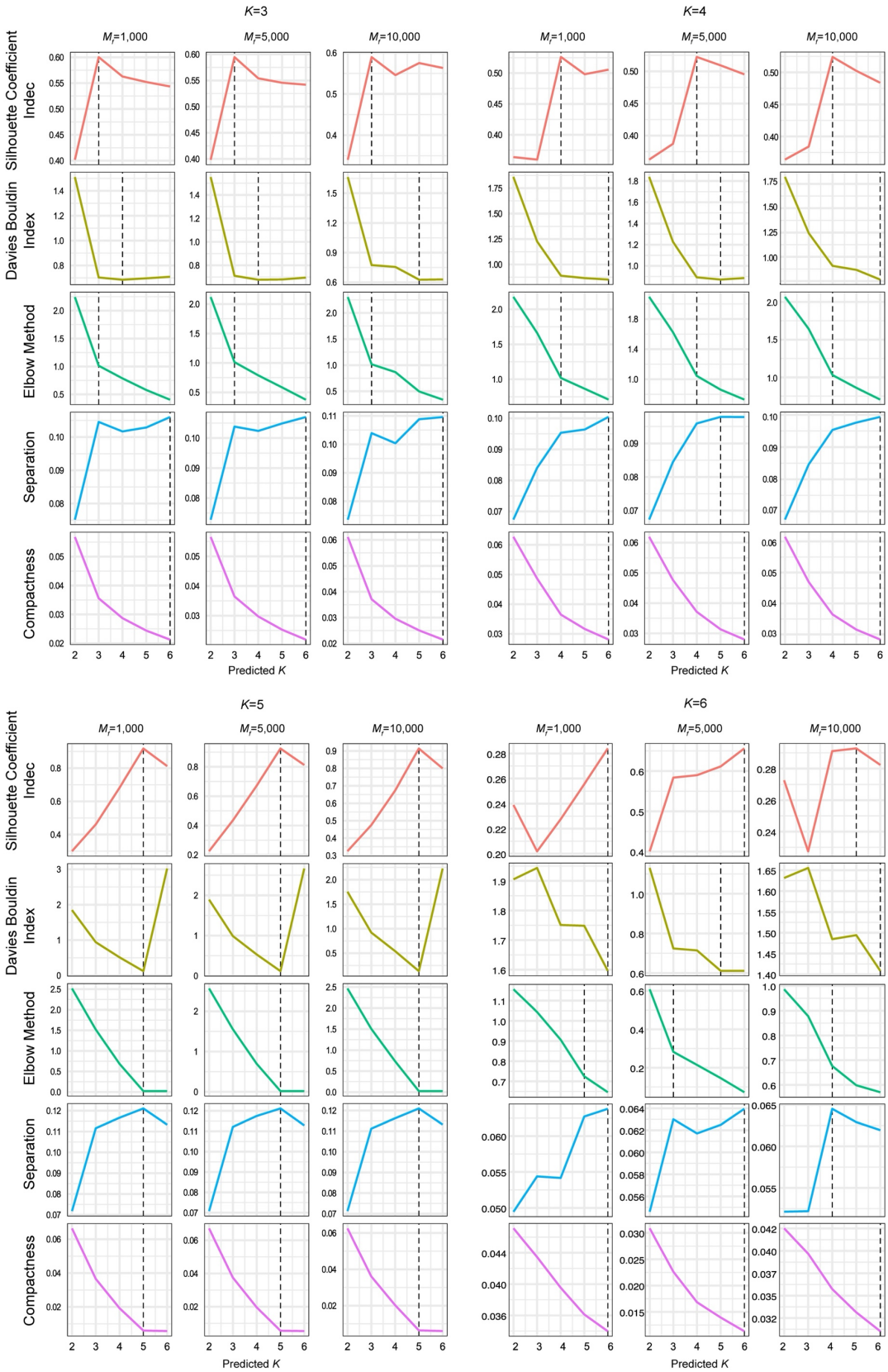


Supplementary Fig. 1. Schematic of Tucker decomposition for multi-omics data.

(a) Tensor integration of multi-omics data into a three-dimensional matrix, and Tucker decomposition factorizing 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.

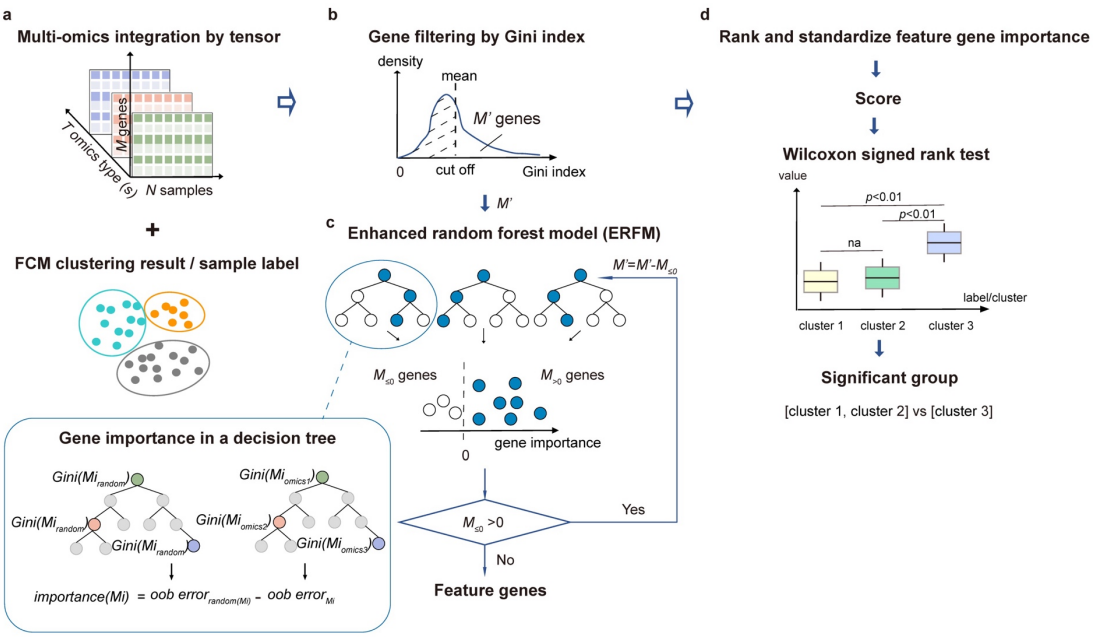


Supplementary Fig. 2. Schematic of multi-omics data integration and FCM clustering.
(a) FCM clustering using the sample-feature matrix with specified cluster count (K) and fuzzy index (m). (b) Membership matrix U initialized to represent samples-cluster probability. (c) Cluster center vectors calculated 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.



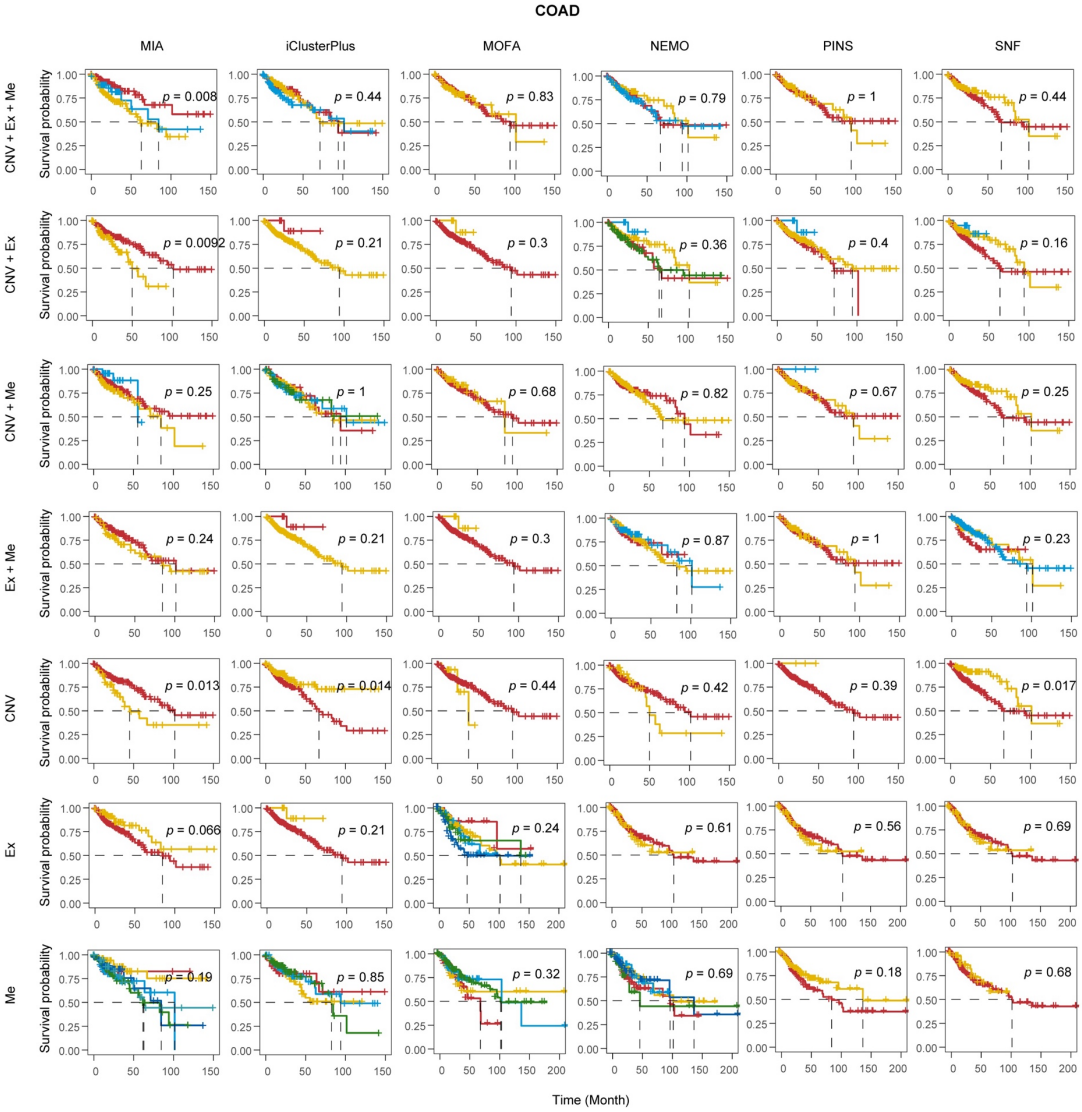
Supplementary Fig. 3. Evaluation of five clustering evaluation indicators for optimal *K* selection on simulated datasets when *K* varies from 3 to 6.

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.



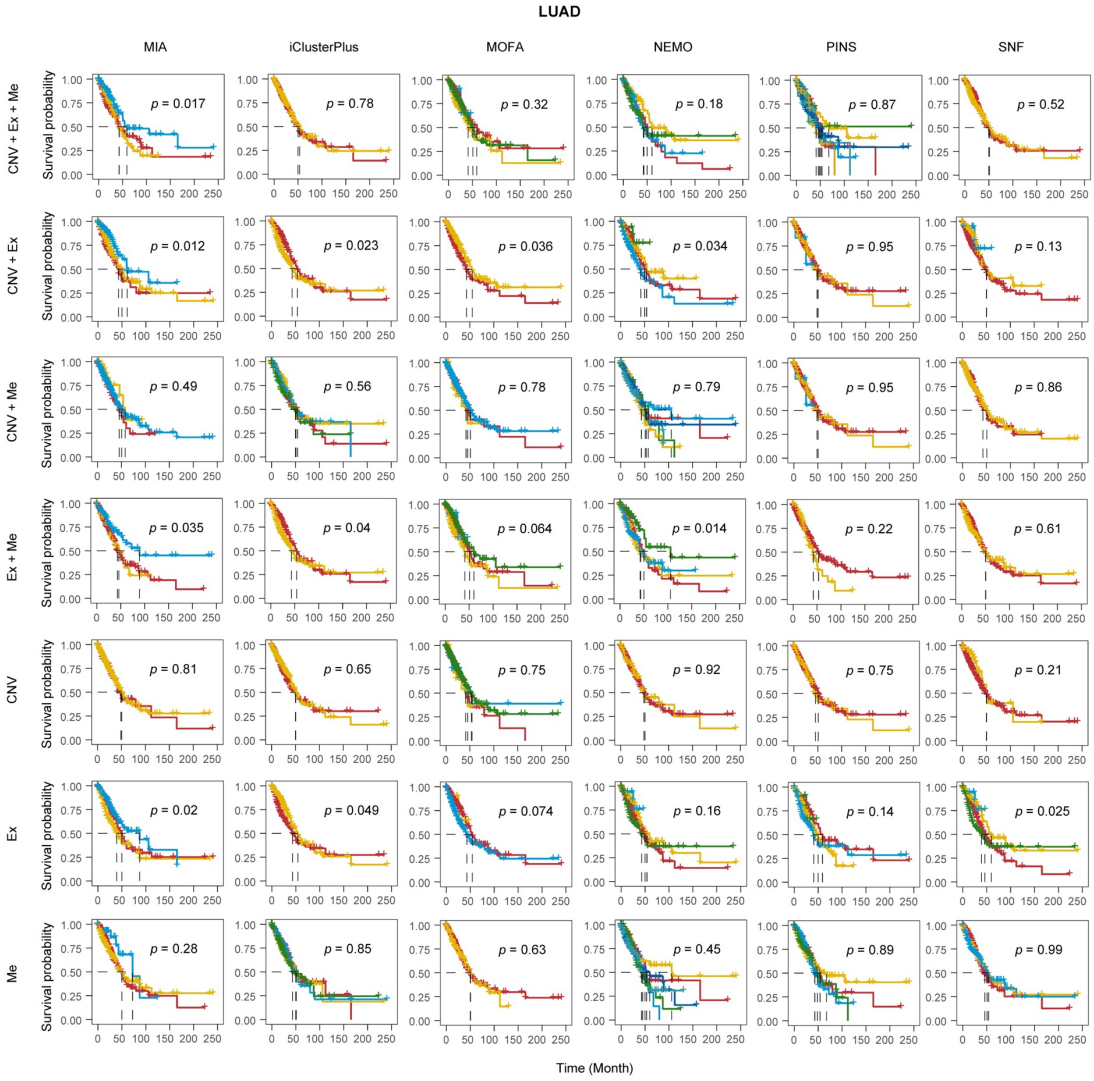
Supplementary Fig. 4. Feature gene identification by ERFM.

(a) Input multi-omics data with corresponding clustering results or sample labels. (b) Preliminary gene filtering based on the Gini index. Genes with values exceeding the mean across all omics are retained. (c) Construction of the ERFM adapted for three-dimensional matrix. Gene importance is quantified by the reduction in classification accuracy following simultaneous randomization of all omics types for each gene. The computational procedure iterates until all genes have positive importance values. (d) Final feature-gene scoring, in which importance values are standardized and significant differences across clusters are evaluated using the Wilcoxon rank-sum test (see details in Methods).



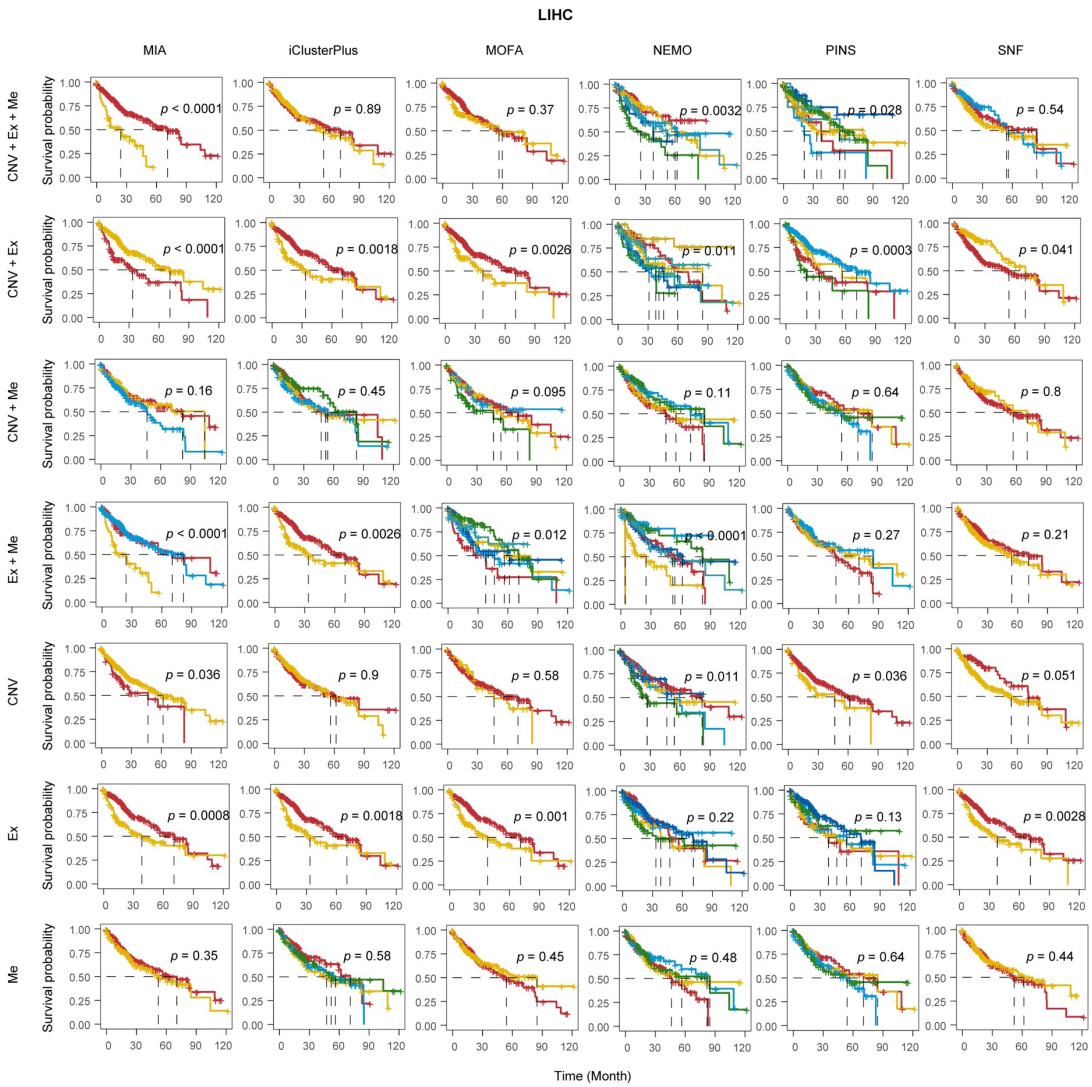
Supplementary Fig. 5. Kaplan-Meier survival analysis on TCGA COAD dataset clustered by different algorithms across various omics combinations.

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.



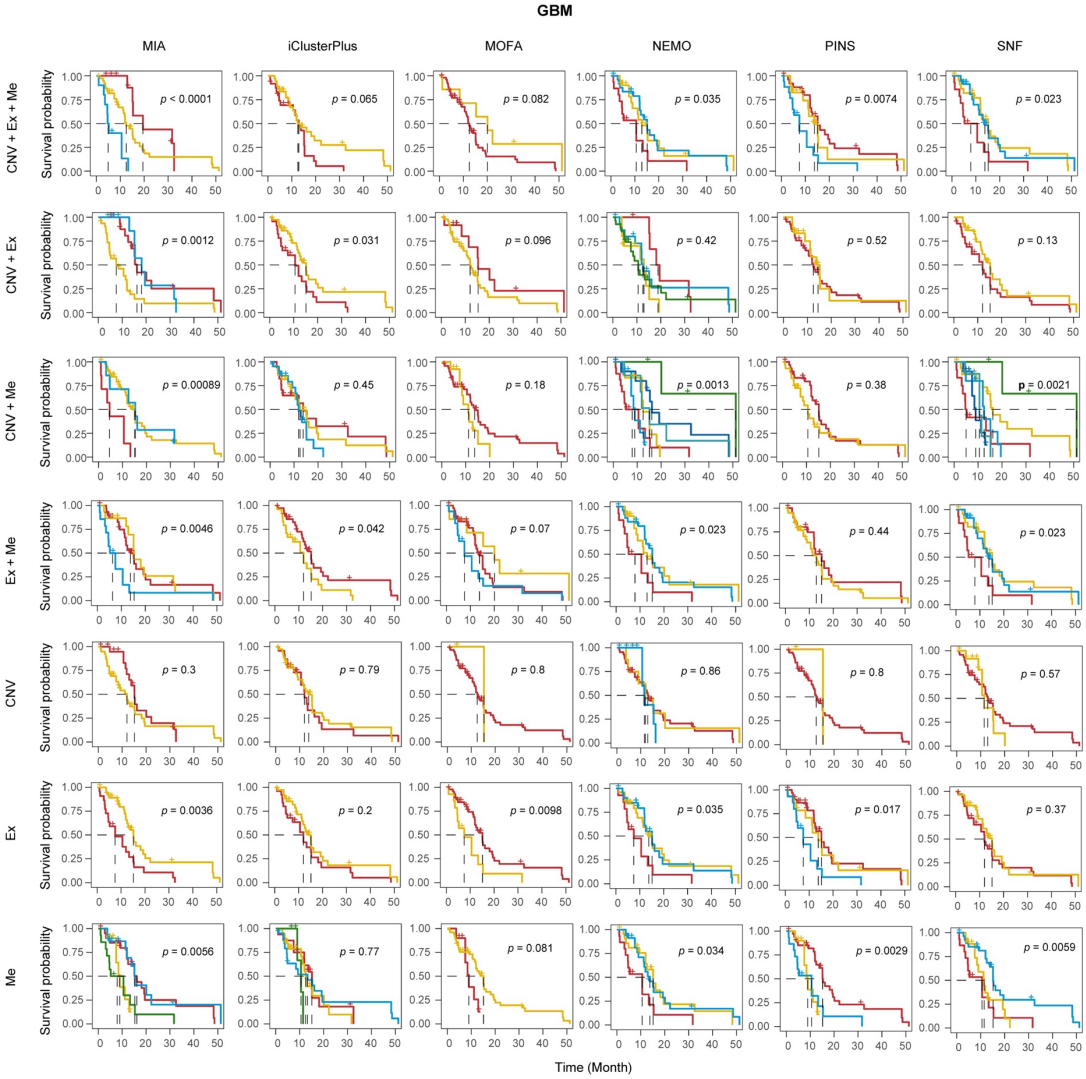
Supplementary Fig. 6. Kaplan-Meier survival analysis on TCGA LUAD dataset clustered by different algorithms across various omics combinations.

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



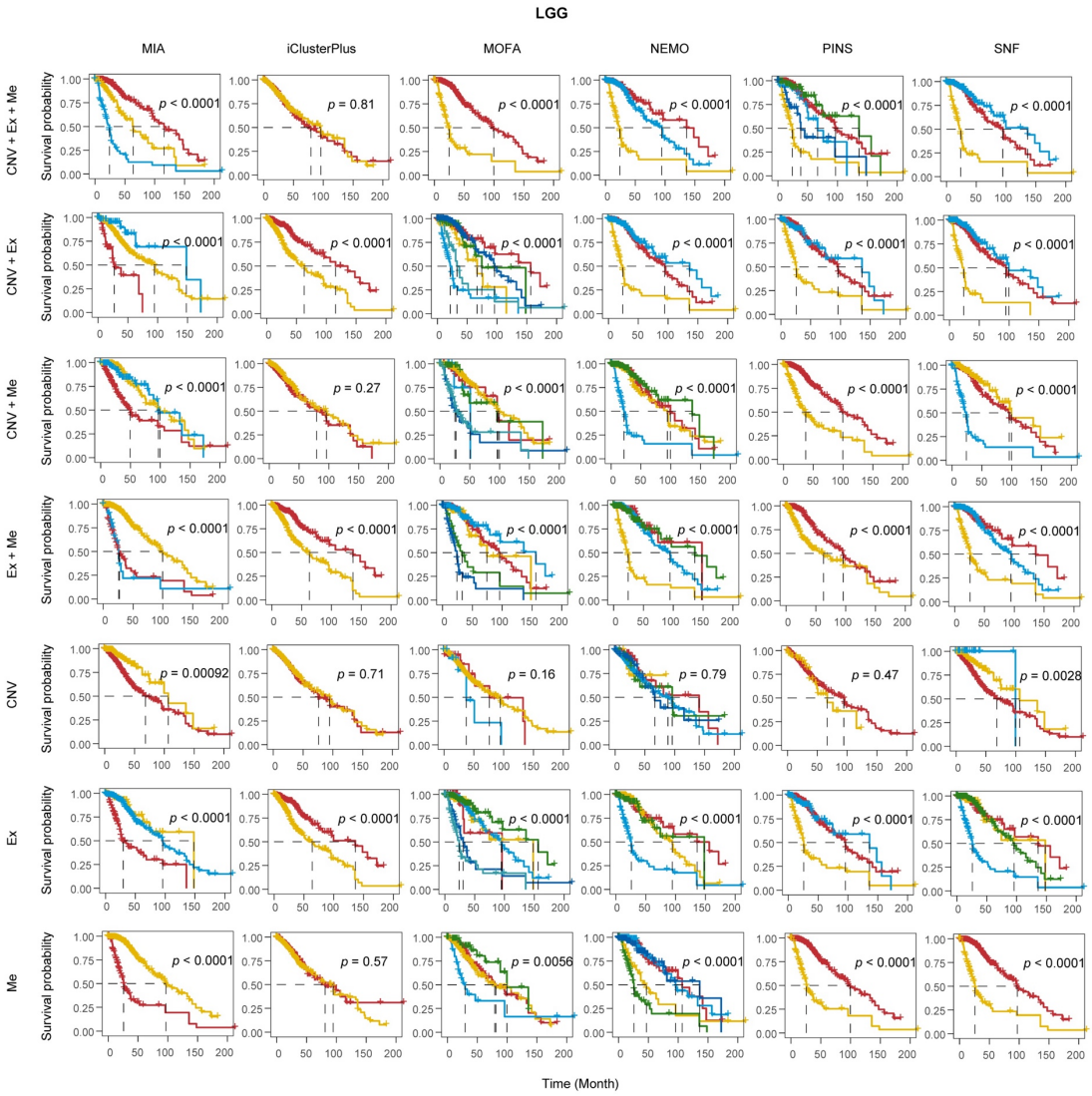
Supplementary Fig. 7. Kaplan-Meier survival analysis on TCGA LIHC dataset clustered by different algorithms across various omics combinations.

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



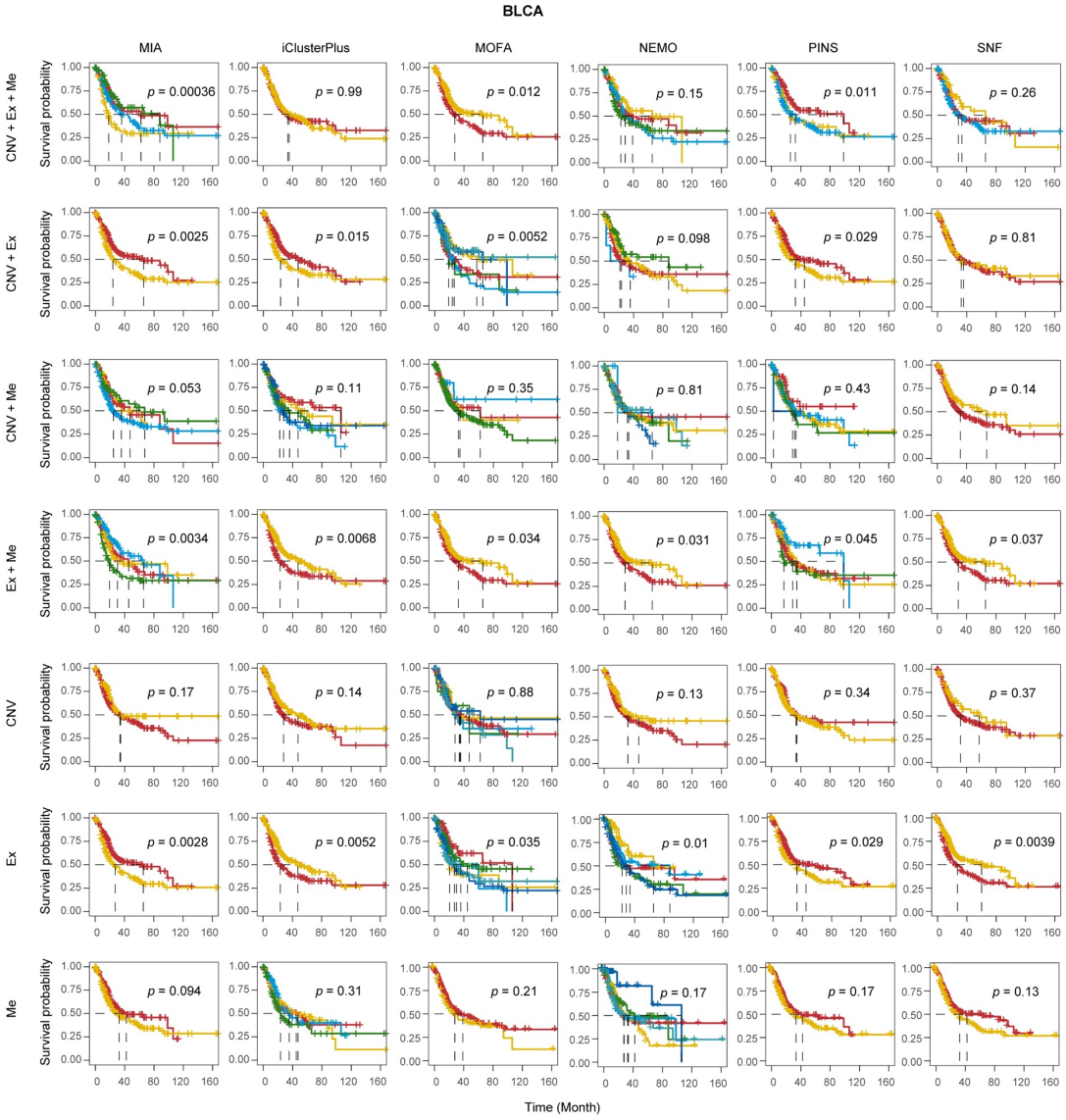
Supplementary Fig. 8. Kaplan-Meier survival analysis on TCGA GBM dataset clustered by different algorithms across various omics combinations.

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



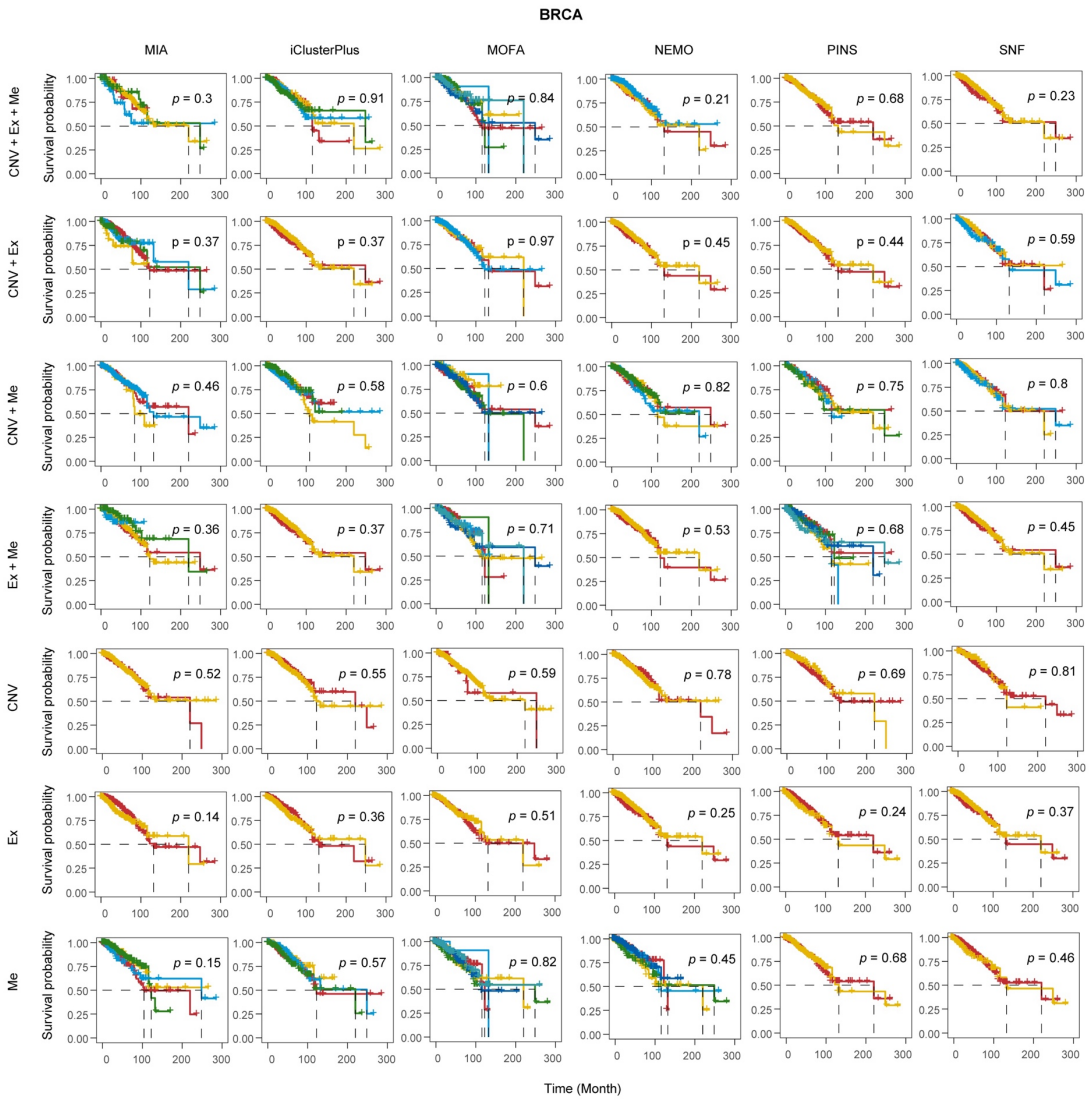
Supplementary Fig. 9. Kaplan-Meier survival analysis on TCGA LGG dataset clustered by different algorithms across various omics combinations.

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



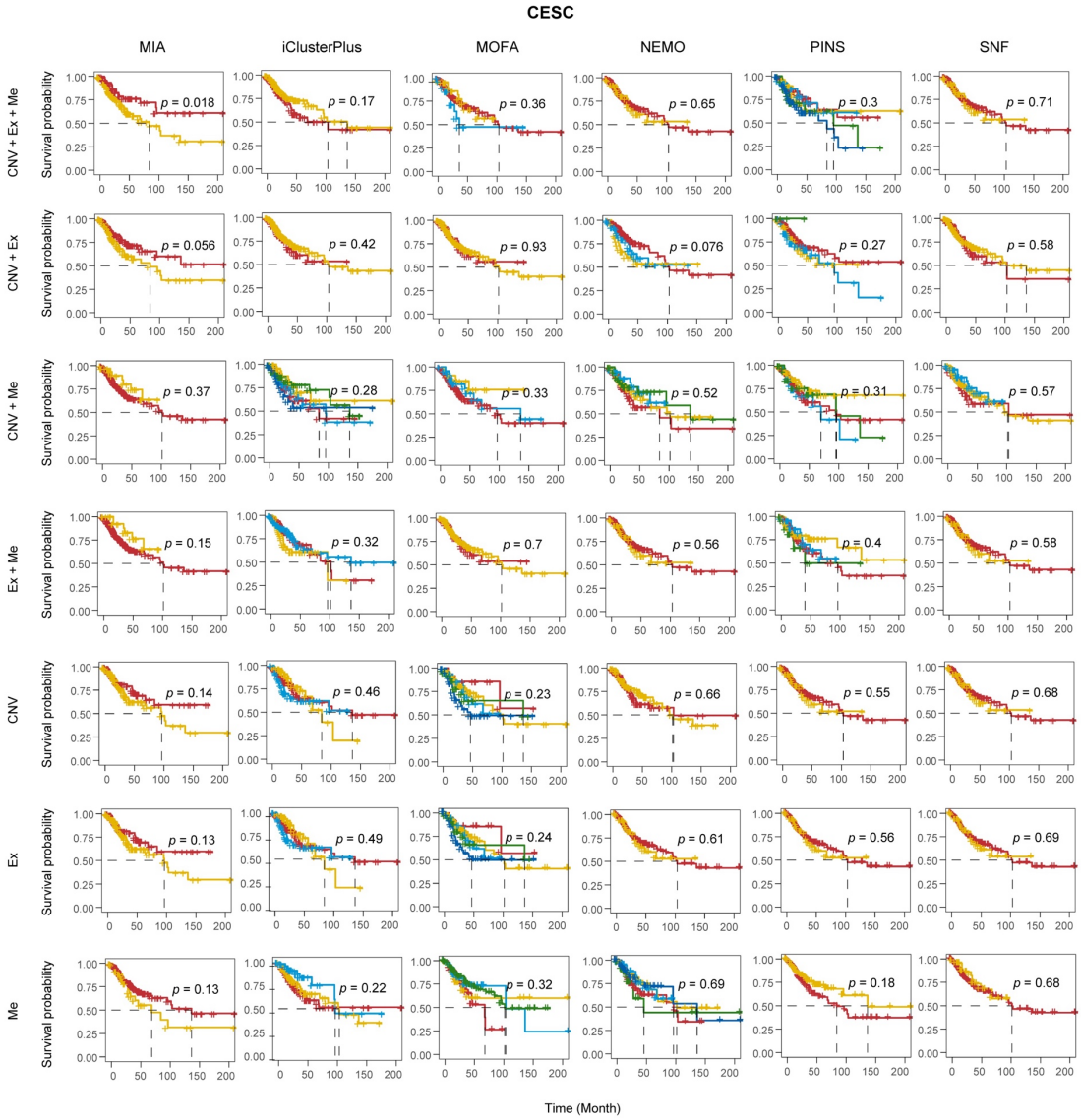
Supplementary Fig. 10. Kaplan-Meier survival analysis on TCGA BLCA dataset clustered by different algorithms across various omics combinations.

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



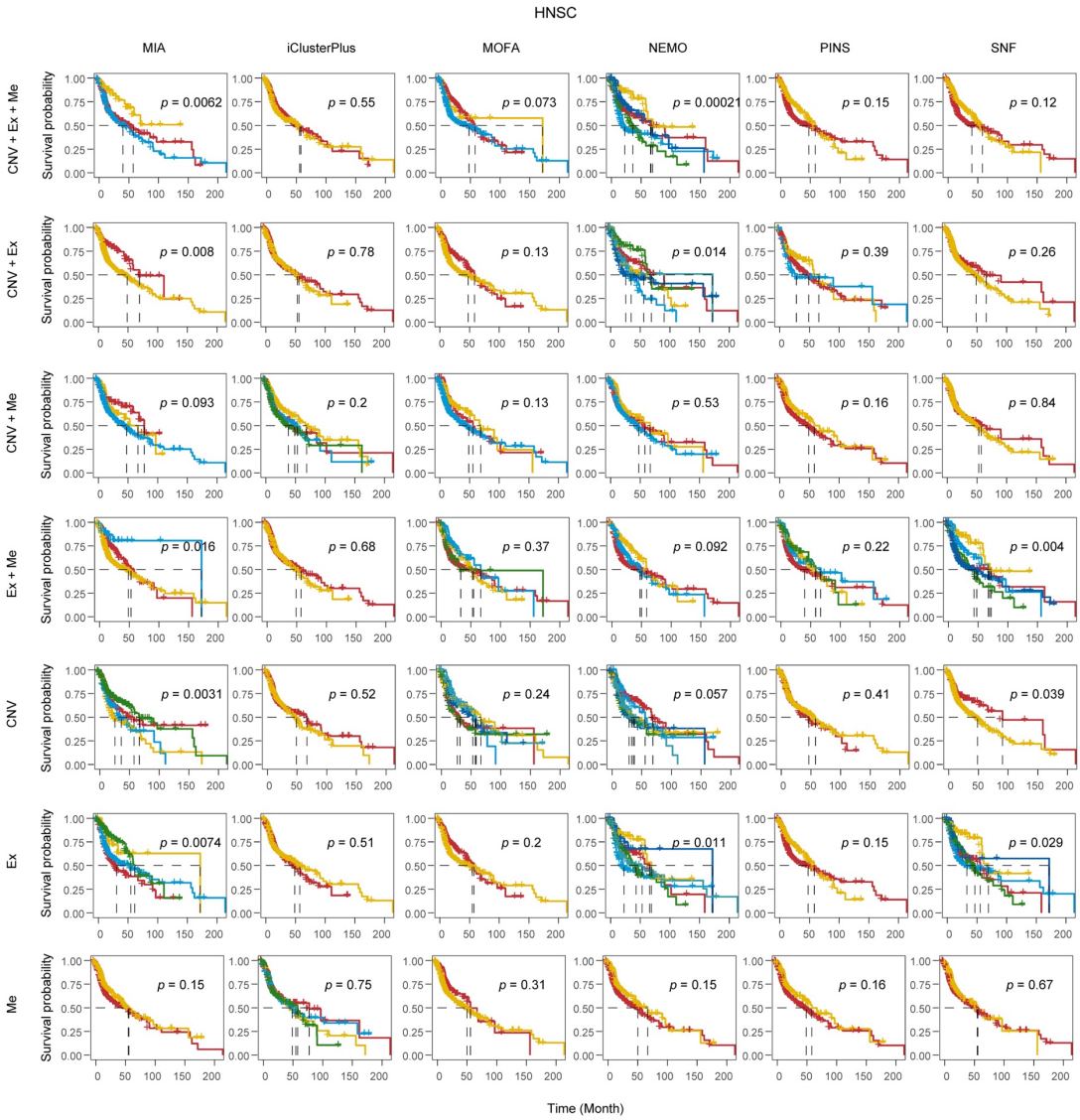
Supplementary Fig. 11. Kaplan-Meier survival analysis on TCGA BRCA dataset clustered by different algorithms across various omics combinations.

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



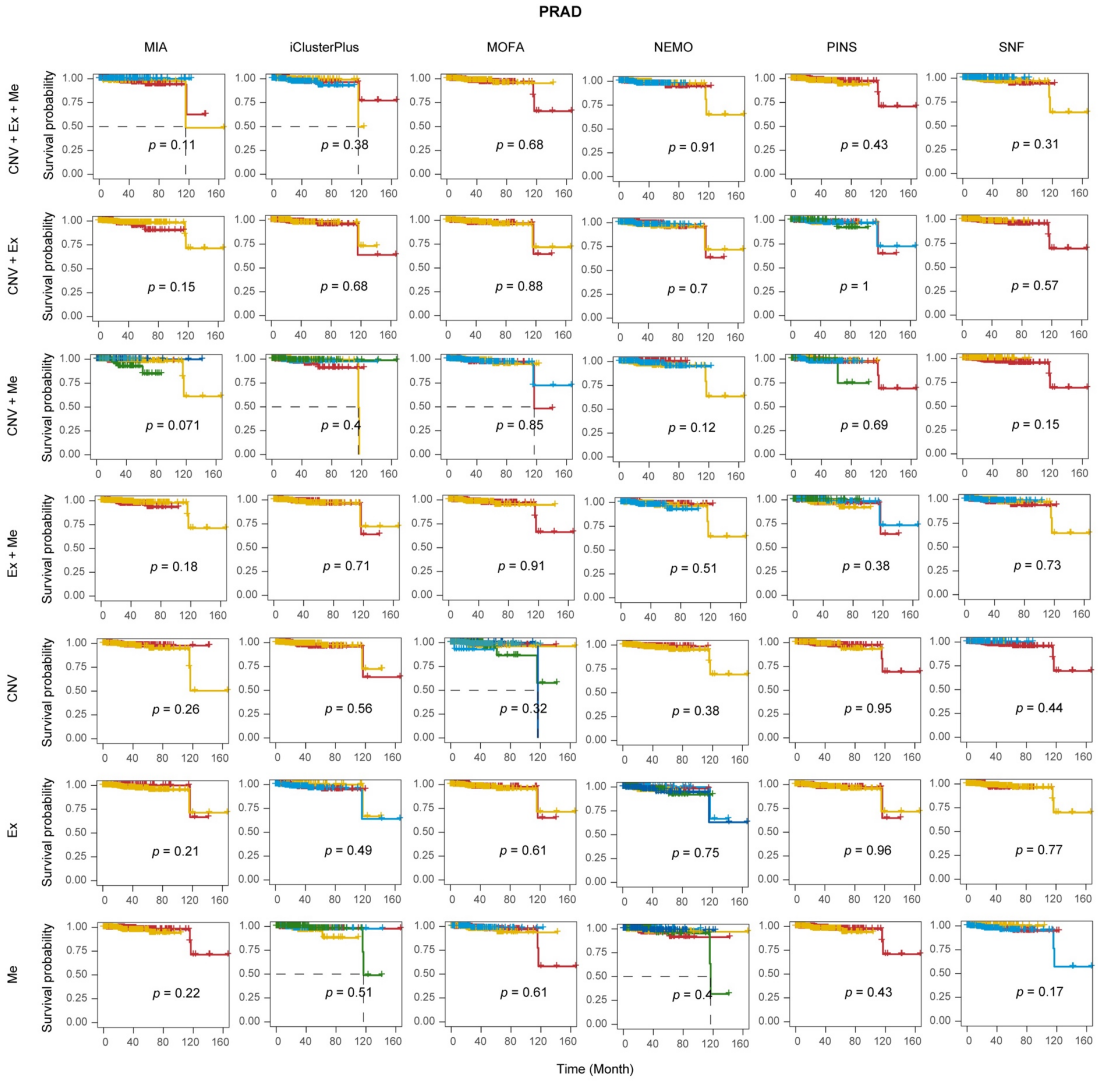
Supplementary Fig. 12. Kaplan-Meier survival analysis on TCGA CESC dataset clustered by different algorithms across various omics combinations.

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



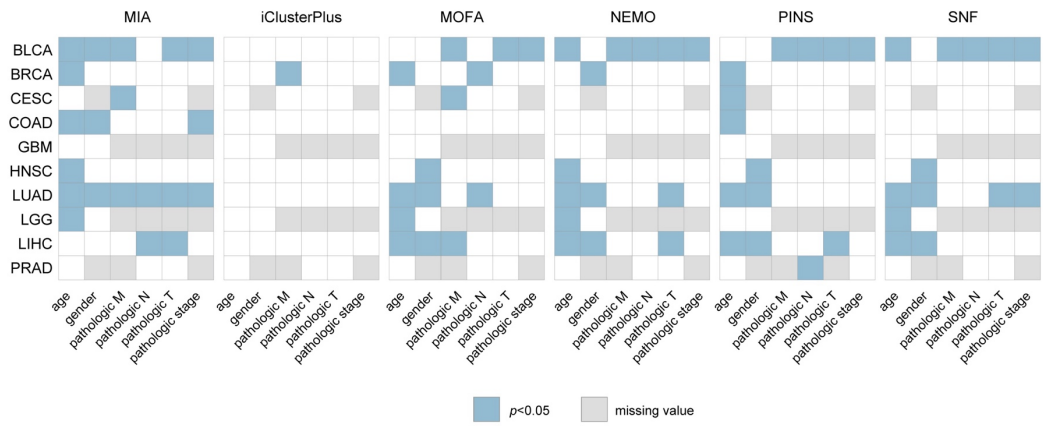
Supplementary Fig. 13. Kaplan-Meier survival analysis on TCGA HNSC dataset clustered by different algorithms across various omics combinations.

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



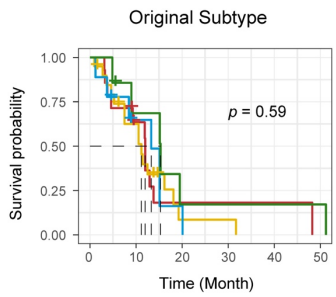
Supplementary Fig. 14. Kaplan-Meier survival analysis on TCGA PRAD dataset clustered by different algorithms across various omics combinations.

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



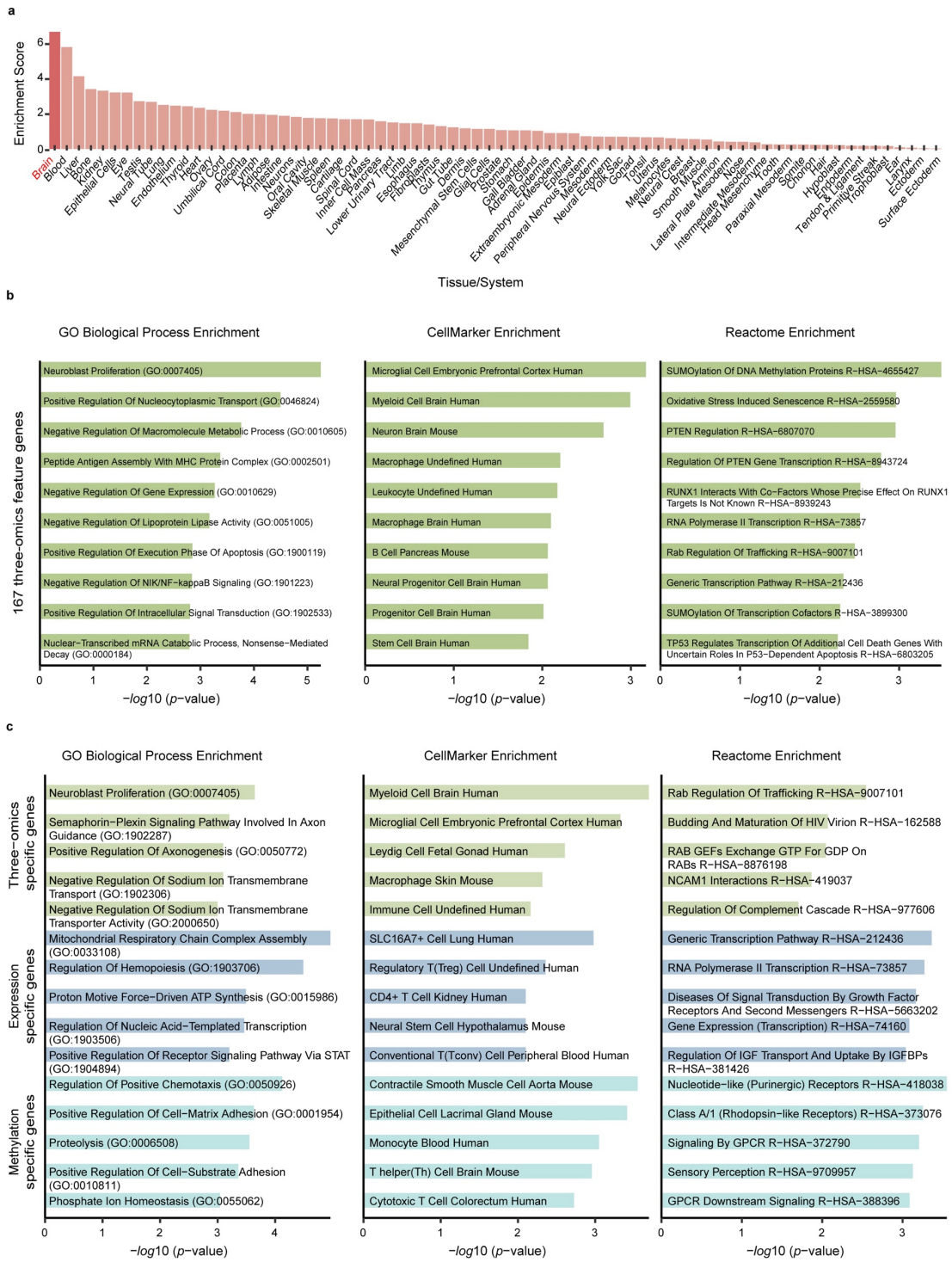
Supplementary Fig. 15. Enrichment of clinical records by different algorithms across ten cancer types.

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



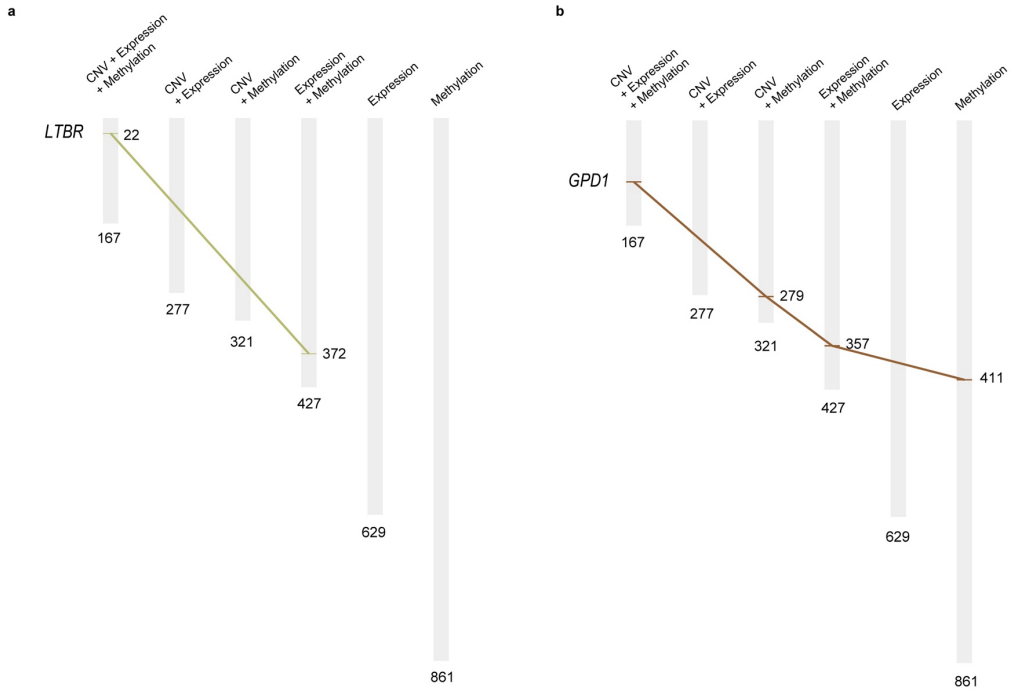
Supplementary Fig. 16. Kaplan-Meier survival analysis of GBM classical subtypes.

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.

776 # Supplementary Tables

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