frontiers | Frontiers in Genetics

TYPE Methods
PUBLISHED 24 January 2023
DOI 10.3389/fgene.2023.1082032

OPEN ACCESS

EDITED BY
Francesca Lantieri,
University of Genoa, Italy

REVIEWED BY
Jin Li,
Hainan Medical University, China
Feng Gao,
Tianjin University, China
Congmin Xu,
Georgia Institute of Technology, United States

*CORRESPONDENCE
Yunping Zhu,
zhuyunping@ncpsb.org.cn

†These authors share first authorship

SPECIALTY SECTION
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

RECEIVED 27 October 2022
ACCEPTED 11 January 2023
PUBLISHED 24 January 2023

CITATION
Chen X, Han M, Li Y, Li X, Zhang J and Zhu Y (2023), Identification of functional gene modules by integrating multi-omics data and known molecular interactions. *Front. Genet.* 14:1082032. doi: 10.3389/fgene.2023.1082032

Identification of functional gene modules by integrating multi-omics data and known molecular interactions

Xiaoqing Chen[1,2†], Mingfei Han[2†], Yingxing Li[3], Xiao Li[2], Jiaqi Zhang[2] and Yunping Zhu[1,2*]

[1]Basic Medical School, Anhui Medical University, Hefei, China, [2]National Center for Protein Sciences (Beijing), Beijing Proteome Research Center, Beijing Institute of Lifeomics, Beijing, China, [3]Central Research Laboratory, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

Multi-omics data integration has emerged as a promising approach to identify patient subgroups. However, in terms of grouping genes (or gene products) into co-expression modules, data integration methods suffer from two main drawbacks. First, most existing methods only consider genes or samples measured in all different datasets. Second, known molecular interactions (e.g., transcriptional regulatory interactions, protein–protein interactions and biological pathways) cannot be utilized to assist in module detection. Herein, we present a novel data integration framework, Correlation-based Local Approximation of Membership (CLAM), which provides two methodological innovations to address these limitations: 1) constructing a trans-omics neighborhood matrix by integrating multi-omics datasets and known molecular interactions, and 2) using a local approximation procedure to define gene modules from the matrix. Applying Correlation-based Local Approximation of Membership to human colorectal cancer (CRC) and mouse B-cell differentiation multi-omics data obtained from The Cancer Genome Atlas (TCGA), Clinical Proteomics Tumor Analysis Consortium (CPTAC), Gene Expression Omnibus (GEO) and ProteomeXchange database, we demonstrated its superior ability to recover biologically relevant modules and gene ontology (GO) terms. Further investigation of the colorectal cancer modules revealed numerous transcription factors and KEGG pathways that played crucial roles in colorectal cancer progression. Module-based survival analysis constructed four survival-related networks in which pairwise gene correlations were significantly correlated with colorectal cancer patient survival. Overall, the series of evaluations demonstrated the great potential of Correlation-based Local Approximation of Membership for identifying modular biomarkers for complex diseases. We implemented Correlation-based Local Approximation of Membership as a user-friendly application available at https://github.com/free1234hm/CLAM.

KEYWORDS

multi-omics integration, gene module detection, proteomic, transcriptomic, genomic

1 Introduction

Increasing attention has been devoted to the integration of multi-omics data to discover coherent biological signatures. In a comprehensive review of multi-omics data integration methods, Huang et al. (Huang et al., 2017) categorized the existing algorithms into four classes: matrix factorization methods (e.g., NMF (Zhang et al., 2011; Zhang et al., 2012) and iCluster

(Shen et al., 2012), Bayesian methods (e.g., MDI (Kirk et al., 2012), BCC (Lock and Dunson, 2013) and CONEXIC (Akavia et al., 2010), network-based methods (e.g., SNF (Wang et al., 2014), MoGCN (Li et al., 2022) and Lemon-tree (Bonnet et al., 2015), and multi-step analysis (e.g., CNAmet (Louhimo and Hautaniemi, 2011) and iPAC (Aure et al., 2013). These methods can discover patient subgroups when using samples as clustering objects and genes (or gene products) as clustering features, or identify co-expressed gene modules by exchanging the clustering objects and features.

However, most of the existing methods are particularly suitable for patient subtyping. Although some methods can be applied to gene module detection, such as jNMF (Zhang et al., 2012), iNMF (Yang and Michailidis, 2016), moCluster (Meng et al., 2016), iCluster, CONEXIC, Lemon-tree (Bonnet et al., 2015), etc., they suffer from two main drawbacks. First, most methods are limited in terms of input data, requiring the datasets from different sources to share the same clustering objects (genes) or features (samples). For example, jNMF and iNMF require the input data to share the same samples, iCluster and moCluster require the input data to share the same genes. Second, because co-expressed genes are often functionally related or co-regulated, known molecular interactions (e.g., transcriptional regulatory interactions, protein–protein interactions and biological pathways) are valuable for improving module detection. Although there are approaches that integrate multi-omics data and molecular interactions, most of these methods are aimed at biomarker discovery. For example, EMOG (Schulte-Sasse et al., 2021) integrates multi-omics data and protein–protein interaction networks to identify new cancer genes. ModulOmics (Silverbush et al., 2019) integrates multi-omics data and molecular networks to improve the identification of cancer driver modules. To our knowledge, molecular interactions are rarely used to improve the identification of co-expressed gene modules.

Herein, we present a novel analytical framework referred to as Correlation-based Local Approximation of Membership (CLAM), which employs three methodological innovations to address the above challenges. First, CLAM constructs a k-nearest neighbor (KNN) matrix for each dataset and combines them into a trans-omics neighborhood matrix. The combined matrix includes all genes measured in at least one dataset. Therefore, this step does not require different datasets to share the same genes or samples. Second, CLAM uses various known molecular interactions, such as transcriptional regulatory interactions, protein–protein interactions and biological pathways, to adjust the neighborhood matrix. Third, CLAM applies a local approximation procedure to define gene modules and performs module-based survival analysis to evaluate module–disease relationships. We have implemented CLAM as a user-friendly application with extensive interactive interfaces available at https://github.com/free1234hm/CLAM.

By applying CLAM and state-of-the-art module detection methods to human colorectal cancer (CRC) and mouse B-cell differentiation multi-omics data obtained from The Cancer Genome Atlas (TCGA), Clinical Proteomics Tumor Analysis Consortium (CPTAC), Gene Expression Omnibus (GEO) and ProteomeXchange database, we demonstrated that CLAM showed the highest precision, recall, relevance and recovery metrics in recovering biologically relevant modules and identified the highest number of gene ontology (GO) terms in enrichment analysis. Additionally, further investigation of the CRC modules revealed numerous transcription factors (TFs) and KEGG pathways that played crucial roles in CRC progression. Module-based survival analysis constructed four gene networks significantly correlated with CRC survival. In contrast to traditional survival genes, which affect patient survival based on their own expression levels, genes in the four survival-related networks affect patient survival based on the levels of their co-expression. We found that many genes in these networks played crucial roles in cancer progression and could serve as potential prognostic biomarkers. Overall, our results demonstrated the superior ability of CLAM in reconstructing modular structure from multi-omics data and identifying modular biomarkers for CRC.

2 Materials and methods

CLAM includes modules with the following three key functions (Figure 1): 1) constructing a trans-omics neighborhood matrix by integrating the k-nearest neighbor matrices obtained from different data sources; 2) using known molecular interactions to adjust the combined neighborhood matrix; and 3) using a local approximation procedure to define gene modules.

2.1 Construction of the trans-omics neighborhood matrix

In each dataset, we first calculate the similarity between each pair of objects (genes or proteins) and extract the $\boldsymbol{k}$ (10 as default) nearest neighbors for each object. The similarity measure can be Euclidean distance, mutual information and Pearson correlation coefficient, etc. Second, the similarity measures between each object and its nearest neighbors are used to calculate a set of weights $\boldsymbol{W} = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k\}$. The weight between genes $\boldsymbol{x}$ and $\boldsymbol{y}$ is calculated as $\boldsymbol{w}_{xy} = \boldsymbol{S}_{xy} / \sum_{z \in KNN(x)} \boldsymbol{S}_{xz}$, where $\boldsymbol{S}_{xy}$ represents the similarity measures between genes $\boldsymbol{x}$ and $\boldsymbol{y}$. The neighbors that have higher similarities are given higher weights and $\sum_{y \in KNN(x)} \boldsymbol{w}_{xy} = 1$. Third, we combine the KNN matrices of different datasets into a global neighborhood matrix, which includes the neighborhood information for all the genes measured in at least one dataset. If a gene has measurements in $\boldsymbol{m}$ datasets, it has $\boldsymbol{m} \times \boldsymbol{k}$ neighbors in the combined matrix with the previous weights divided by $\boldsymbol{m}$. Therefore, different genes may have different numbers of neighbors (Figure 1 Step 1). Finally, the duplicated neighbors and their weights are merged. This step ensures that the duplicated neighbors are given higher weights.

2.2 Calculation of the prior correlation probability

For each gene in the combined neighborhood matrix ($\boldsymbol{g}_1$ in Figure 1 Step 2), we construct a co-regulatory network using protein–protein interactions (PPIs), transcriptional regulatory interactions (*via* TFs or miRNAs) and KEGG pathways. This network consists of $\boldsymbol{g}_1$ and its neighbors which directly interact with $\boldsymbol{g}_1$ or share the same transcriptional regulators with $\boldsymbol{g}_1$. Second, we calculate a co-regulation score between $\boldsymbol{g}_1$ and each neighbor according to the network structure. Assuming that $\boldsymbol{g}_1$ binds to $\boldsymbol{n}_1$ miRNAs; $\boldsymbol{n}_2$ TFs and directly interacts with $\boldsymbol{n}_3$ genes through PPIs, the edges directly

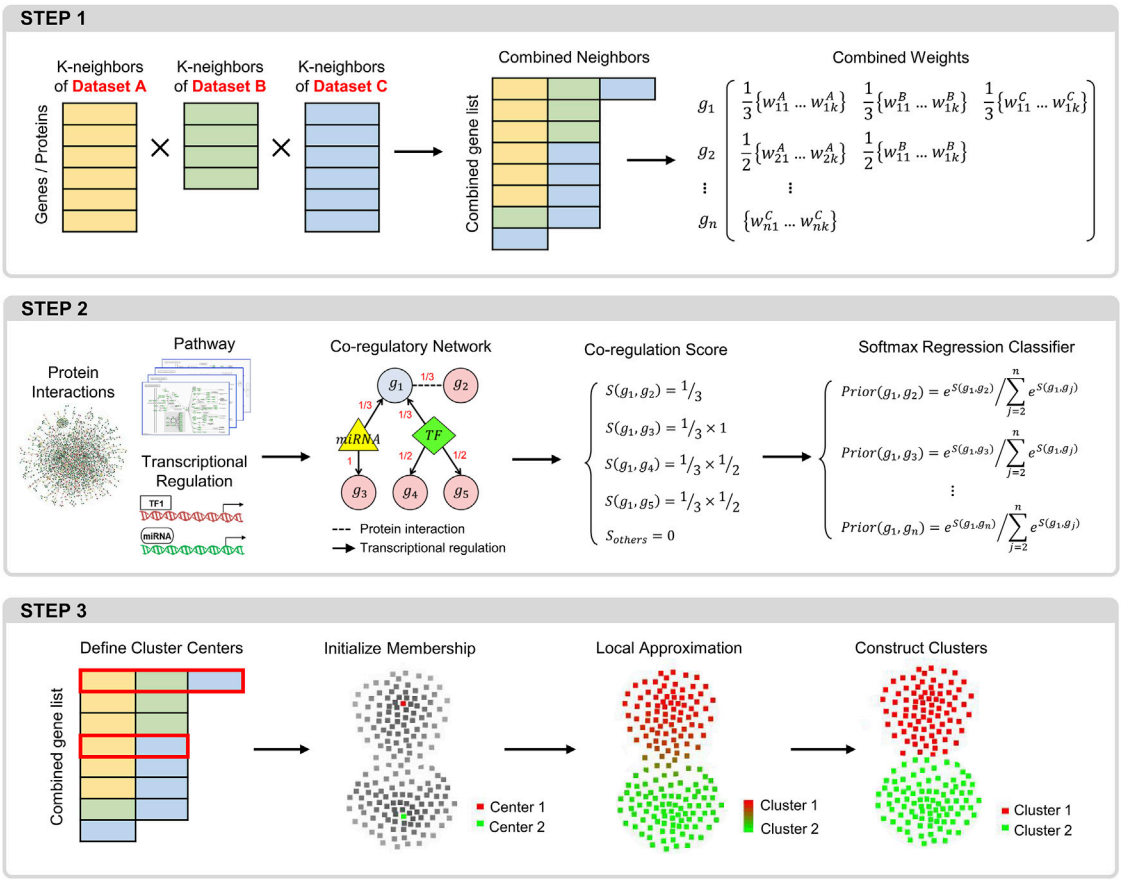

FIGURE 1
Overview of the CLAM workflow. Step 1: procedure for integrating the KNN matrices of different datasets into a global neighborhood matrix. Step 2: procedure for calculating the prior correlation probabilities between each gene and its neighbors in the neighborhood matrix. Step 3: the local approximation procedure for identifying gene modules.

connected to g_1 have the same weight $1/(n_1 + n_2 + n_3)$. Given a miRNA (or TF) that binds to n_4 neighbors of g_1, the weight between the miRNA and each target $g_{i\in[1,n_4]}$ is $1/n_4$. Finally, the co-regulation score between g_1 and g_i is $1/(n_1 + n_2 + n_3) \times 1/n_4$. Neighbors not included in the co-regulatory network are scored zero. Third, we calculate the prior probability between g_1 and each neighbor using softmax regression (a generalization of logistic regression that we can use for multi-class classification). Finally, the weights between gene x and its neighbors are transformed to $(w_{xy} \times prior_{xy})$, where $y \in KNN(x)$. This step assures that functionally related or co-regulated genes are given relatively high weights.

2.3 Identification of gene modules

We followed the local approximation process proposed by Fu et al. (Fu and Medico, 2007) to define gene modules. First, the density of one object is calculated as the average similarity measure between this object and its k-nearest neighbors. Second, the densities of all objects are used to identify cluster centers and outliers: 1) one object is defined as a cluster center when its density is higher than that of all objects in its neighborhood and 2) one object is defined as an outlier when its density is lower than that of all objects in its neighborhood. The higher k is, the fewer cluster centers will be identified; as a consequence, fewer clusters will be generated. Third, we define a membership vector for each object. Assuming that we have identified M cluster centers, the membership vector of each object x is represented as $p(x) = \{p_1(x), \ldots, p_{M+1}(x)\}$, in which each element $p_i(x)$ indicates the membership degree of x to cluster i and the last element indicates the probability that x is an outlier.

Next, we initialize the membership vector of each object. First, each cluster center is assigned a unique membership vector, where only the element corresponding to its own cluster is 1 and the other elements are 0. Second, all outliers are assigned the same membership vector, in which the last element is 1 and the other elements are 0. Third, for all other objects, the elements in each vector are set to the same value $1/(M + 1)$. Subsequently, through an iteration process, we update the membership vector of each object (except for cluster centers and outliers) using its linear approximation, which is calculated by combining its nearest neighbors' membership vectors, namely, $p(x) \approx \sum_{y\in KNN(x)} w_{xy}p(y)$, where w is the weight matrix produced by integrating multi-omics data and known relationships.

The iteration process is terminated when the overall difference between all membership vectors and their approximations is minimized, which is calculated as follows:

$$E = \sum_{x \in X} \left\| p(x) - \sum_{y \in KNN(x)} w_{xy} p(y) \right\|^2$$

where each term is the difference between the membership vector $p(x)$ and the linear approximation of $p(x)$ by its neighbors $\sum_{y \in KNN(x)} w_{xy} p(y)$. Finally, each object is assigned to the cluster with the highest score in the final membership vector.

2.4 Data collection and preprocessing

Mouse and human multi-omics data were used for the evaluation study. The mouse datasets include RNA-seq data from GEO (GSE75417) and MS data from PRIDE (PXD003263). Both datasets share the same samples collected at six time points during the differentiation process of mouse pre-B-cells. The human data include RNA-seq and MS data of CRC patients obtained from TCGA and CPTAC. The RNA-seq and MS data include 497 and 90 samples, respectively, in which 47 samples are collected from the same patients. Because some existing methods (e.g., iNMF and jNMF) require the input data to share the same set of samples, only the 47 samples from the same patients were used in the evaluation study. Among the initial genes or proteins measured in different omics datasets, we removed those with more than 20% missing values. The remaining missing values were filled using the KNN imputation method. The processed expression matrices are included in the CLAM toolkit as test data.

2.5 Evaluation metrics

We followed the evaluation pipeline proposed by Saelens et al. (Saelens et al., 2018) to compare the performance of CLAM and existing methods. First, we collected various types of known modules, including 1599 human and 1078 mouse miRNA modules extracted from known miRNA–target interactions (Chen and Wang, 2020), 795 human and 1349 mouse TF modules extracted from known TF–target interactions (Cahan et al., 2014; Han et al., 2018; Zhang et al., 2020), and 335 pathway modules from the KEGG database (Kanehisa et al., 2017). Second, we calculated the recovery, relevance, recall, and precision metrics by comparing the known modules with a set of detected modules. These scores have been previously used in several evaluation studies (Prelic et al., 2006; Amigo et al., 2009; Eren et al., 2013; Saelens et al., 2018). If G represents all genes, M a set of known modules, M' a set of observed modules, $M(g)$ the modules that contain gene g, and $E(g, M)$ the set of genes that are included with g in at least one module of M (including g itself), the precision is defined as follows:

$$Precision = \frac{1}{|G|} \sum_{g \in G} \left[\frac{1}{|E(g, M')|} \sum_{g' \in E(g, M')} \frac{min\left(|M'(g) \cap M'(g')|, |M(g) \cap M(g')|\right) \times \Phi(g, g')}{|M'(g) \cap M'(g')|} \right]$$

where $\Phi(g, g') = \frac{1}{|M'(g, g')|} \sum_{m' \in M'(g, g')} \max_{m \in M(g, g')} Jaccard(m', m)$. Recall is calculated in the same way but with M' and M switched. In addition, relevance is defined as $Relevance = \frac{1}{|M'|} \sum_{m' \in M'} \max_{m \in M} Jaccard(m', m)$, and recovery is calculated in the same way but with M' and M switched. Third, before combining the four scores, we normalized every score by dividing it by an average score of 500 permuted versions of the known modules. Finally, we calculated the harmonic mean between the normalized versions of all four scores to obtain an overall score.

2.6 Module-based survival analysis

In traditional survival analysis, patients are ranked according to the expression of a specific gene. The log-rank test is then applied to determine whether there is a significant survival difference between the top and bottom half (or 1/4) of the ranked patients. However, the differential expression of a single gene is not the only factor that affects patient survival, and the traditional approach ignores the potential effects of differential regulation between multiple genes. In this study, we present a module-based survival analysis approach to identify the sets of genes whose co-expression levels are significantly correlated with patient survival.

Given a gene module of $M = (G, S, v)$, where $G = \{g_1, g_2, g_3\}$ represents the genes included in M, $S = \{s_1, \ldots, s_N\}$ represents all patient samples, and v: $3 \times N$ represents the expression-value matrix. Assuming that the expression profile of g_1 ($\{v_{11}, \ldots, v_{1N}\}$) is positively correlated with that of g_2 ($\{v_{21}, \ldots, v_{2N}\}$) and negatively correlated with that of g_3 ($\{v_{31}, \ldots, v_{3N}\}$). First, z-score normalization is applied to $\{v_{11}, \ldots, v_{1N}\}$, $\{v_{21}, \ldots, v_{2N}\}$, and $\{v_{31}, \ldots, v_{3N}\}$, which ensures the same weight of the three genes. Second, the normalized expression of g_3 is transformed to $\{-v_{31}, \ldots, -v_{3N}\}$, which ensures that the three genes show theoretically similar expression patterns. Third, for each patient, s_i ($i \in [1, N]$), we calculate the standard deviation (σ_i) of the transformed expression values of the three genes (v_{1i}; v_{2i};; $-v_{3i}$). The standard deviation σ_i can represent the co-expression level of the three genes in patient s_i. Genes are highly co-expressed in patients with low standard deviations and present lower co-expression in patients with high standard deviations. Finally, the log-rank test is applied to compare the survival curves between patients whose standard deviations were greater than the median *versus* those whose standard deviations were less than or equal to the median. A significant *p*-value indicates that the three genes affect patient survival based on their pairwise expression similarities.

3 Results

3.1 Method evaluation

Using RNA-seq and MS data of human CRC and mouse B-cell differentiation (see 'Data collection and preprocessing'), we conducted a comprehensive evaluation of 12 module detection methods, including 7 integrative clustering methods and 5 individual clustering methods. The integrative clustering methods included CLAM1 (using known molecular interactions to assist module detection), CLAM2 (without using known interactions), jNMF (Zhang et al., 2012), iNMF (Yang and Michailidis, 2016), Lemon-tree (Bonnet et al., 2015), moCluster (Meng et al., 2016) and iCluster (Shen et al., 2009). And the individual clustering methods included CLAM3 (applying the CLAM algorithm to individual datasets), independent component analysis (ICA) (Im et al., 2022), FLAME (Fu and Medico, 2007), K-means (Lin et al., 2004) and WGCNA (Langfelder and Horvath, 2008). Among these methods, CLAM1 is the only one that addresses three critical challenges, including allowing

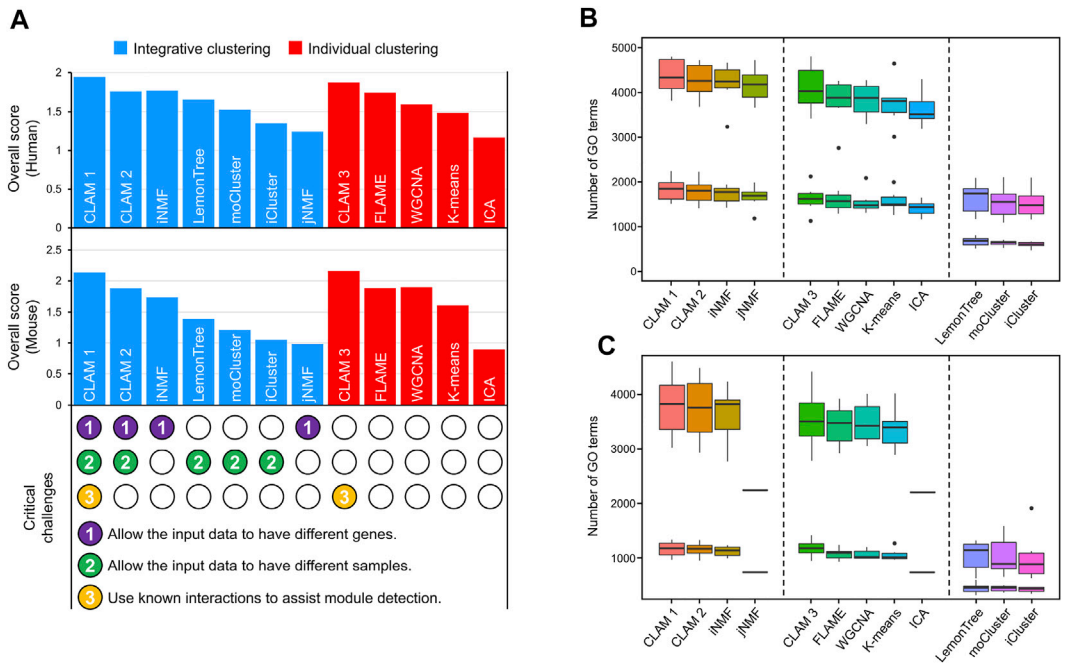

FIGURE 2
Comparison of 12 module detection methods, including 7 integrative clustering methods and 5 individual clustering methods. **(A)** The histograms show the agreement between the observed and known modules. Scores in the upper and middle histograms are generated by applying the tested methods to human and mouse data, respectively. The bottom shows whether the tested methods address three critical challenges of integrative clustering. **(B–C)** The number of significantly enriched GO terms in **(B)** human and **(C)** mouse modules with *p*-value thresholds of 0.01 and 0.05.

different samples in different datasets, allowing different genes in different datasets, and utilizing known molecular interactions to assist module detection (see the bottom of Figure 2A).

According to the evaluation pipeline described in MATERIALS AND METHODS, we scored the different methods by comparing their observed modules with sets of known modules obtained from TF/miRNA–target interaction networks and KEGG pathway database. Notably, because CLAM1 and CLAM3 use known interactions to assist model training, we used 5-fold cross-validation to differentiate the training and validation sets. First, the known modules were randomly divided into five parts. Second, CLAM1 and CLAM3 used four of those parts for training and reserved one-fifth for evaluation, while the other methods used each of the five folds to evaluate their performance. Finally, step 2 was repeated five times to calculate the average evaluation score for each method. The evaluation results revealed several interesting conclusions. First, CLAM1 and CLAM3 achieved the highest overall scores among all tested methods (Figure 2A), indicating that the utilization of known interactions significantly improved the consistency between the observed and known modules (although the interactions used were not included in the known modules). Second, CLAM2 outperformed most of the existing integrative clustering methods, but showed no significant advantage over FLAME or WGCNA. This indicated that the integration of datasets from different sources contributed little to the overall score. However, using another evaluation method, gene ontology (GO) enrichment analysis, we reached a different conclusion.

According to the number of significantly enriched GO terms, the tested methods could be divided into three categories (Figures 2B, C): methods that allow the input data to have different genes (CLAM1, CLAM2, iNMF and jNMF) performed best; methods that process each input dataset separately (CLAM3, FLAME, WGCNA and k-means) took second place; and methods that cluster the overlapping genes in the input data (Lemon-tree, moCluster and iCluster) identified the fewest GO terms. These results suggest that the number of significantly enriched GO terms is positively correlated with the total number of genes in the final modules. Therefore, the integrative clustering methods that allow the input data to have different genes are well suited for GO enrichment analysis because they cluster the union of genes in the input data. Additionally, we found a significant reduction in the number of GO terms identified by jNMF and ICA from the mouse data (Figure 2C). This is because both jNMF and ICA have a limitation that the number of modules must be less than the number of samples. However, there are only 18 samples in the mouse RNA-seq and MS data, which results in jNMF and ICA generating far fewer modules than the other methods.

In summary, the utilization of known molecular interactions can improve the agreement with known modules, while data integration can improve the discovery of functional annotations. This is why CLAM1, which integrates multi-omics data and known molecular interactions, performs best on both evaluation metrics.

3.2 Investigation of the resulting modules

CRC is the third most common malignant cancer with the second highest mortality rate (Ayerden et al., 2021; Rydbeck et al., 2021; Peng et al., 2022). By applying CLAM to multi-omics data of CRC (see 'Data collection and preprocessing') with KNN ranging from 5 to 15, we obtained 88, 71, 59, 49, 42, 37, 35, 28, 27, 25 and 23 gene modules. TF and KEGG pathway enrichment analyses performed on these modules

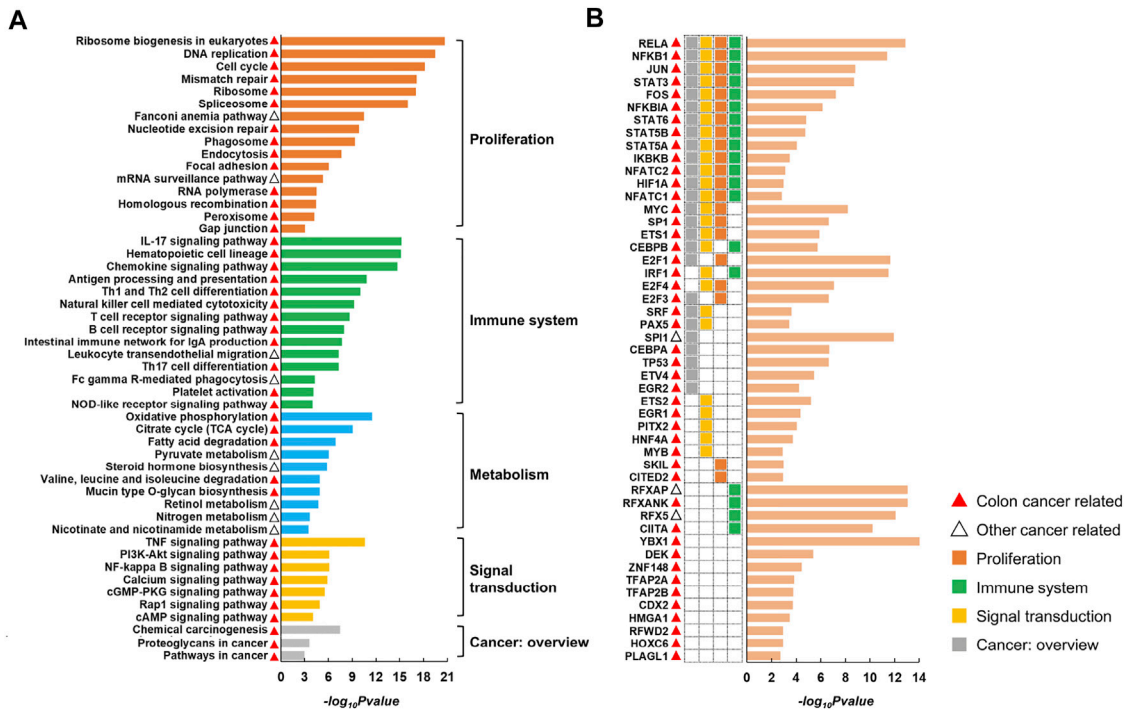

FIGURE 3
Representative pathways **(A)** and TFs **(B)** identified by enrichment analysis of CRC modules. The histogram displays the negative log10 of the enrichment *p*-values. The red and black triangles indicate that the corresponding entry has been reported to be associated with colon cancer or other cancer types. The pathways can be divided into five classes according to KEGG Orthology.

produced 77 pathways (Supplementary Table S1) and 49 TFs (Supplementary Table S2) shared by different parameters. Representative pathways and TFs are displayed in Figure 3.

According to KEGG Orthology, we divided the pathways into five classes, including proliferation (e.g., cell cycle and DNA replication), immune system (e.g., natural killer cell-mediated cytotoxicity and chemokine signaling pathway), signal transduction (e.g., PI3K/Akt signaling pathway and NF-κB signaling pathway), cancer overview (e.g., proteoglycans in cancer and chemical carcinogenesis) and metabolism (e.g., oxidative phosphorylation and pyruvate metabolism). Additionally, we found that among the 49 TFs, 39 were involved in at least one of the above pathways (see heatmap in Figure 3B). Particularly, 13 TFs were involved in four types of pathways associated with proliferation, immune system, signal transduction and cancer overview, indicating a close correlation between the resulting TFs and pathways.

Following a broad literature exploration, we observed that 67 out of the 77 overlapping pathways were previously reported to be cancer related, in which 55 pathways had been reported in colon cancer and 12 pathways had been reported in other cancer types (Supplementary Table S1). This was reasonable because the overlapping results of different parameters would have been more likely to be studied in previous studies. Additionally, 46 out of the 49 TFs are known to play roles in CRC and the remaining three (RFXAP, SPI1 and RFX5) are related to other cancers (Supplementary Table S2). For example, RELA, NFKB1, NFKBIA and IKBKB are involved in the synthesis and activation of NF-κB, which supports tumorigenesis by promoting cell proliferation, invasion and metastasis and inhibiting apoptosis (Patel et al., 2018). Dysregulation of E2F family (E2F1, E2F3 and E2F4) expression activates or silences oncogenes or tumor suppressors at multiple levels of gene regulation and is involved in CRC progression (Kent and Leone, 2019; Xu et al., 2021). RFXAP, RFX5, RFXANK and CIITA are all associated with MHC II expression, and mutations in any of them lead to MHC II deficiency, which may result in immune evasion in CRC (Michel et al., 2010; Axelrod et al., 2019). Overall, the series of results demonstrated the superior ability of CLAM in reconstructing the modular structure of complex biological systems.

3.3 Gene networks associated with CRC survival

Survival analysis is a cornerstone of medical research, enabling the assessment of clinical outcomes for disease progression and treatment efficiency (Lanczky and Gyorffy, 2021). In traditional survival analysis, patients are divided into low- and high-expression groups based on the expression of a specific gene (Figure 4A) (Figure 5). However, genes are rarely regulated independently and are instead interconnected. A relatively simple way to study the synergistic effects of multiple genes in prognosis is to divide patients based on the average expression of multiple genes. Figure 4B shows two examples of this approach, where genes in Module 1 divide patients into Groups 1 and 2, and genes in Module 2 divide patients into Groups 3 and 4. However, this approach

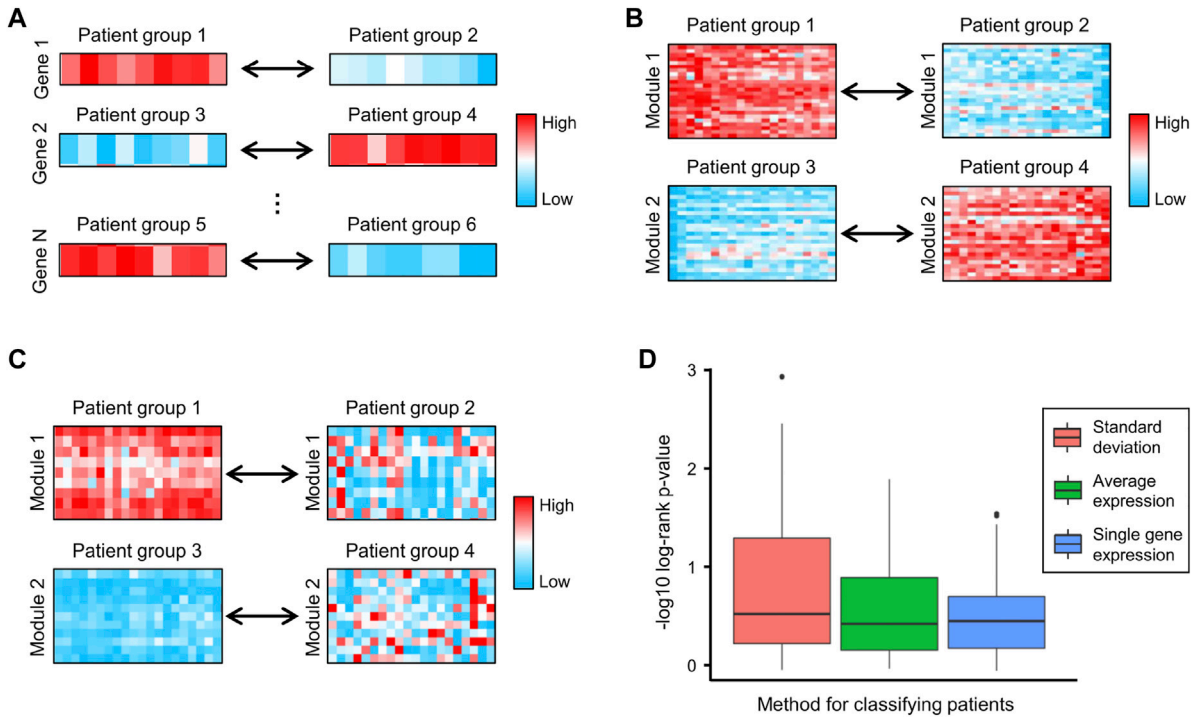

FIGURE 4
(A–C) Examples of classifying patients based on **(A)** the expression of single genes, **(B)** the average expression of genes in each module, and **(C)** the standard deviation of genes in each module, respectively. **(D)** The negative log10 of the log-rank test *p*-values generated by using single gene expression, average expression and standard deviation to classify patients.

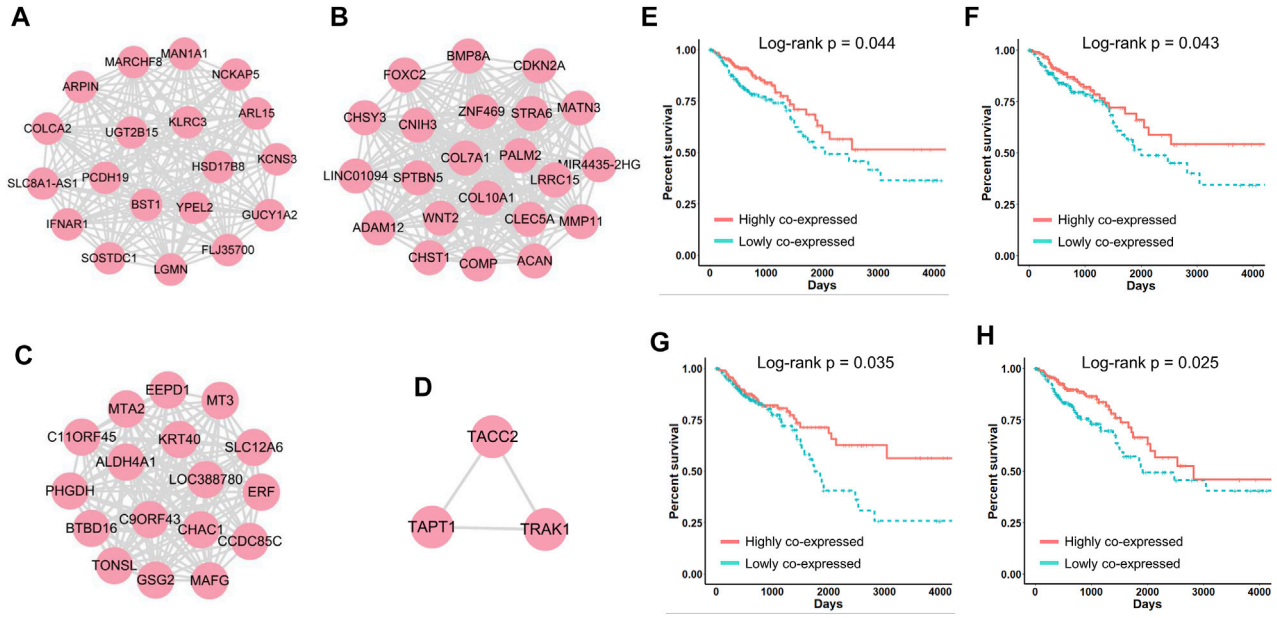

FIGURE 5
(A–D) Four networks correlated with CRC patient survival. **(E–H)** Kaplan–Meier survival curves generated by performing module-based survival analyses on networks **(A–D)**.

focuses on the differences in gene expression and ignores the correlations between genes. For example, genes in Module 1 are highly expressed in all patients in Group 1 and lowly expressed in all patients in Group 2, suggesting that these genes exhibit significant pairwise expression correlations in both groups of patients. Since many expression correlations arise from functional relationships, the functional relationships in Module 1 are preserved in both Group 1 and Group 2, which limits the difference in survival between Groups 1 and 2. To address this issue, we used the standard deviation of gene expression levels to classify patients. Figure 4C shows two examples of this approach, where genes are highly co-expressed (functionally related) in the left groups and present lower co-expression (dysfunctional) in the right groups.

To compare the above classification criteria, we used the CRC gene modules to classify CRC patients based on individual gene expression, the average expression of genes in each module, and the standard deviation of genes in each module, respectively, and used the log-rank test to assess survival differences. The results showed that classifying patients using the standard deviation yielded the lowest log-rank *p*-values (Figure 4D), indicating that gene co-expression levels were highly correlated with CRC progression and patient survival. We further investigated four survival-related modules (Supplementary Table S3) in which gene–gene co-expression levels were significantly correlated with the overall survival of CRC patients. The network maps and survival curves of these modules are shown in Figure 4. Because co-expressed genes often present functional consistency (Ghazalpour et al., 2006; Kakati et al., 2019), these modules are likely to involve critical gene regulatory or functional relationships affecting the prognosis of CRC.

Evidence has shown that many genes in these modules play crucial roles in cancer progression and can serve as potential prognostic biomarkers. In Figure 4A, HSD17B8 is a good candidate for advanced tumor stages (Luque-García et al., 2010), and COLCA2 is recognized as a colorectal cancer-associated gene (Yin et al., 2022). In Figure 4B, overexpression of COL10A1 enhances the proliferation, migration, and invasion of CRC cells (Huang et al., 2018); MMP11 expression affects the immune response in CRC (Buttacavoli et al., 2021); ADAM12 may play vital roles in the recruitment of immune cells in CRC (Huang et al., 2021); and COMP promotes CRC cell proliferation partially through the activation of the PI3K/Akt/mTOR/p70S6K pathway (Liu et al., 2018). In Figure 4C, ALDH4A1 deficiency leads to the accumulation of proline, which sustains the proliferation and survival of CRC cells (Alaqbi et al., 2022), and MT3 plays a pivotal role in tumor formation, progression, and drug resistance (Si and Lang, 2018). In Figure 4D, TRAK1 is a prognostic biomarker of CRC (An et al., 2011).

Notably, when we performed traditional survival analysis on every individual gene in these networks, only the expression of five genes (CHST1, CHSY3, COMP, MATN3 and PALM2) was significantly correlated with the survival of CRC patients (Supplementary Figure 1). This indicated that in many cases, patient prognosis is not decided by the expression of a single gene but by the synergistic effects of multiple co-regulated genes, which are often neglected by traditional approaches. In summary, with CLAM we defined four networks closely correlated with CRC patient survival, which provided numerous known and novel biomarkers that played critical roles in CRC progression.

4 Discussion

With the accumulation of multi-omics expression data, researchers have continuously improved data integration approaches for decades. Nevertheless, most methods were aimed at discovering patient subtypes, and no substantial progress has been made in gene module detection. To address this issue, we developed CLAM, which addressed several critical limitations of data integration and achieved considerable progress in discovering gene regulatory and functional relationships from multi-omics data. However, two issues are worthy of further discussion and research.

First, integration of datasets from different sources is quite time-consuming and requires bulk memory space. Suppose we apply CLAM to three expression matrices (N_1, M_1), (N_2, M_2) and (N_3, M_3), where N represent sets of genes, M represents sets of samples, and K represents the KNN parameter. For $M \ll N$ and $K \ll N$, the time complexity approximately equals to $O({N_1}^2 + {N_2}^2 + {N_3}^2)$, which is the sum of the time spent by an individual clustering algorithm on multiple datasets. We further compared the running time of CLAM with that of other integrative clustering algorithms. The results showed that CLAM took 15.4 s to perform integrative clustering on the CRC datasets, second only to moCluster (Supplementary Table S4). Considering that moCluster clusters the overlapping genes in the input data (2293 genes), CLAM is the fastest algorithm for clustering the total genes in the input data (12,847 genes). Additionally, algorithm optimization and parallelization can help save computing time. For example, moCluster saves a lot of time by using the consensus PCA algorithm to replace the EM-algorithm used by iCluster (Meng et al., 2016), and parallelization saves nearly half the time for the CLAM algorithm.

Second, methods for integrating gene mutation and expression profiles can be divided into two categories. The first category can predict cancer genes by integrating copy number variations (CNVs) and expression data, such as iPAC (Aure et al., 2013) and NetICS (Dimitrakopoulos et al., 2018). The second category aims to identify cancer driver pathways (or modules) by integrating somatic mutations, CNVs and gene expressions, such as iMCMC (Zhang et al., 2013) and ModulOmics (Silverbush et al., 2019). Different from the integrative clustering algorithms that identify all potential co-expression modules, these methods focus on genes or modules associated with cancer mutations. However, not all cancer genes are associated with cancer mutations. One possible solution is to introduce gene mutation information in the co-expression modules identified by CLAM. In this way, we can identify modules associated with cancer mutations. In subsequent iterations of CLAM, we will explore this algorithm and test its performance.

In addition to improving module detection, this study provides a module-based analysis pipeline to investigate module–disease relationships. With this pipeline we constructed four gene networks significantly correlated with CRC patient survival. Through an extensive literature exploration, we demonstrated that genes in these networks played crucial roles in tumor progression and metastasis. More importantly, we have shown that these results may be missed by traditional survival analysis. With the accumulation of multi-omics data, we believe that module detection and subsequent analysis will attract increasing attention, significantly promoting biomarker discovery in complex diseases.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

YZ conceived and supervised the project. MH designed and implemented the CLAM software. XC and JZ. applied CLAM to multi-omics data of human CRC and mouse B-cell differentiation. XL participated in the evaluation of CLAM. MH, XC, XL, and JZ. participated in writing and revising the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the National Key Research and Development Program of China (Grant No. 2021YFA1301603).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1082032/full#supplementary-material

References

Akavia, U. D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H. C., et al. (2010). An integrated approach to uncover drivers of cancer. *Cell* 143, 1005–1017. doi:10.1016/j.cell.2010.11.013

Alaqbi, S. S., Burke, L., Guterman, I., Green, C., West, K., Palacios-Gallego, R., et al. (2022). Increased mitochondrial proline metabolism sustains proliferation and survival of colorectal cancer cells. *PLoS One* 17, e0262364. doi:10.1371/journal.pone.0262364

Amigo, E., Gonzalo, J., Artiles, J., and Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr. J.* 12, 613–486. doi:10.1007/s10791-009-9106-z

An, Y., Zhou, Y., Ren, G., Tian, Q., Lu, Y., Li, H., et al. (2011). Elevated expression of MGb2-Ag/TRAK1 is correlated with poor prognosis in patients with colorectal cancer. *Int. J. Colorectal Dis.* 26, 1397–1404. doi:10.1007/s00384-011-1237-1

Aure, M. R., Steinfeld, I., Baumbusch, L. O., Liestol, K., Lipson, D., Nyberg, S., et al. (2013). Identifying in-trans process associated genes in breast cancer by integrated analysis of copy number and expression data. *PLoS One* 8, e53014. doi:10.1371/journal.pone.0053014

Axelrod, M. L., Cook, R. S., Johnson, D. B., and Balko, J. M. (2019). Biological consequences of MHC-II expression by tumor cells in cancer. *Clin. Cancer Res.* 25, 2392–2402. doi:10.1158/1078-0432.CCR-18-3200

Ayerden, D., Tayfur, M., and Balci, M. G. (2021). Comparison of histopathological findings of the colon adenomas and adenocarcinomas with cyclin D1 and Ki-67 expression. *Niger. J. Clin. Pract.* 24, 1737–1741. doi:10.4103/njcp.njcp_68_21

Bonnet, E., Calzone, L., and Michoel, T. (2015). Integrative multi-omics module network inference with Lemon-Tree. *PLoS Comput. Biol.* 11, e1003983. doi:10.1371/journal.pcbi.1003983

Buttacavoli, M., Di Cara, G., Roz, E., Pucci-Minafra, I., Feo, S., and Cancemi, P. (2021). Integrated multi-omics investigations of metalloproteinases in colon cancer: Focus on MMP2 and MMP9. *Int. J. Mol. Sci.* 22, 12389. doi:10.3390/ijms222212389

Cahan, P., Li, H., Morris, S. A., Lummertz Da Rocha, E., Daley, G. Q., and Collins, J. J. (2014). CellNet: Network biology applied to stem cell engineering. *Cell* 158, 903–915. doi:10.1016/j.cell.2014.07.020

Chen, Y., and Wang, X. (2020). miRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Res.* 48, D127–D131. doi:10.1093/nar/gkz757

Dimitrakopoulos, C., Hindupur, S. K., HäFLIGER, L., Behr, J., Montazeri, H., Hall, M. N., et al. (2018). Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics* 34, 2441–2448. doi:10.1093/bioinformatics/bty148

Eren, K., Deveci, M., Kucuktunc, O., and Catalyurek, U. V. (2013). A comparative analysis of biclustering algorithms for gene expression data. *Brief. Bioinform* 14, 279–292. doi:10.1093/bib/bbs032

Fu, L., and Medico, E. (2007). FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. *BMC Bioinforma.* 8, 3. doi:10.1186/1471-2105-8-3

Ghazalpour, A., Doss, S., Zhang, B., Wang, S., Plaisier, C., Castellanos, R., et al. (2006). Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet.* 2, e130. doi:10.1371/journal.pgen.0020130

Han, H., Cho, J. W., Lee, S., Yun, A., Kim, H., Bae, D., et al. (2018). TRRUST v2: An expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.* 46, D380–D386. doi:10.1093/nar/gkx1013

Huang, H., Li, T., Ye, G., Zhao, L., Zhang, Z., Mo, D., et al. (2018). High expression of COL10A1 is associated with poor prognosis in colorectal cancer. *Onco Targets Ther.* 11, 1571–1581. doi:10.2147/OTT.S160196

Huang, S., Chaudhary, K., and Garmire, L. X. (2017). More is better: Recent progress in multi-omics data integration methods. *Front. Genet.* 8, 84. doi:10.3389/fgene.2017.00084

Huang, Z., Lai, H., Liao, J., Cai, J., Li, B., Meng, L., et al. (2021). Upregulation of ADAM12 is associated with a poor survival and immune cell infiltration in colon adenocarcinoma. *Front. Oncol.* 11, 729230. doi:10.3389/fonc.2021.729230

Im, H., Lee, J. H., and Choi, S. H. (2022). Independent component analysis identifies the modulons expanding the transcriptional regulatory networks of enterohemorrhagic *Escherichia coli*. *Front. Microbiol.* 13, 953404. doi:10.3389/fmicb.2022.953404

Kakati, T., Bhattacharyya, D. K., Barah, P., and Kalita, J. K. (2019). Comparison of methods for differential Co-expression analysis for disease biomarker prediction. *Comput. Biol. Med.* 113, 103380. doi:10.1016/j.compbiomed.2019.103380

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). Kegg: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361. doi:10.1093/nar/gkw1092

Kent, L. N., and Leone, G. (2019). The broken cycle: E2F dysfunction in cancer. *Nat. Rev. Cancer* 19, 326–338. doi:10.1038/s41568-019-0143-7

Kirk, P., Griffin, J. E., Savage, R. S., Ghahramani, Z., and Wild, D. L. (2012). Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* 28, 3290–3297. doi:10.1093/bioinformatics/bts595

Lanczky, A., and Gyorffy, B. (2021). Web-based survival analysis tool tailored for medical research (KMplot): Development and implementation. *J. Med. Internet Res.* 23, e27633. doi:10.2196/27633

Langfelder, P., and Horvath, S. (2008). Wgcna: an R package for weighted correlation network analysis. *BMC Bioinforma.* 9, 559. doi:10.1186/1471-2105-9-559

Li, X., Ma, J., Leng, L., Han, M., Li, M., He, F., et al. (2022). MoGCN: A multi-omics integration method based on graph convolutional network for cancer subtype analysis. *Front. Genet.* 13, 806842. doi:10.3389/fgene.2022.806842

Lin, T. H., Li, H. T., and Tsai, K. C. (2004). Implementing the Fisher's discriminant ratio in a k-means clustering algorithm for feature selection and data set trimming. *J. Chem. Inf. Comput. Sci.* 44, 76–87. doi:10.1021/ci030295a

Liu, T. T., Liu, X. S., Zhang, M., Liu, X. N., Zhu, F. X., Zhu, F. M., et al. (2018). Cartilage oligomeric matrix protein is a prognostic factor and biomarker of colon cancer and promotes cell proliferation by activating the Akt pathway. *J. Cancer Res. Clin. Oncol.* 144, 1049–1063. doi:10.1007/s00432-018-2626-4

Lock, E. F., and Dunson, D. B. (2013). Bayesian consensus clustering. *Bioinformatics* 29, 2610–2616. doi:10.1093/bioinformatics/btt425

Louhimo, R., and Hautaniemi, S. (2011). CNAmet: an R package for integrating copy number, methylation and expression data. *Bioinformatics* 27, 887–888. doi:10.1093/bioinformatics/btr019

Luque-GarcíA, J. L., MartíNEZ-Torrecuadrada, J. L., Epifano, C., CañAMERO, M., Babel, I., and Casal, J. I. (2010). Differential protein expression on the cell surface of colorectal cancer cells associated to tumor metastasis. *Proteomics* 10, 940–952. doi:10.1002/pmic.200900441

Meng, C., Helm, D., Frejno, M., and Kuster, B. (2016). moCluster: Identifying joint patterns across multiple omics data sets. *J. Proteome Res.* 15, 755–765. doi:10.1021/acs.jproteome.5b00824

Michel, S., Linnebacher, M., Alcaniz, J., Voss, M., Wagner, R., Dippold, W., et al. (2010). Lack of HLA class II antigen expression in microsatellite unstable colorectal carcinomas is caused by mutations in HLA class II regulatory genes. *Int. J. Cancer* 127, 889–898. doi:10.1002/ijc.25106

Patel, M., Horgan, P. G., Mcmillan, D. C., and Edwards, J. (2018). NF-κB pathways in the development and progression of colorectal cancer. *Transl. Res.* 197, 43–56. doi:10.1016/j.trsl.2018.02.002

Peng, X., Chen, G., Lv, B., and Lv, J. (2022). MicroRNA-148a/152 cluster restrains tumor stem cell phenotype of colon cancer via modulating CCT6A. *Anticancer Drugs* 33, e610–e621. doi:10.1097/CAD.0000000000001198

Prelic, A., Bleuler, S., Zimmermann, P., Wille, A., Buhlmann, P., Gruissem, W., et al. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* 22, 1122–1129. doi:10.1093/bioinformatics/btl060

Rydbeck, D., Asplund, D., Bock, D., Haglind, E., Park, J., Rosenberg, J., et al. (2021). Younger age at onset of colorectal cancer is associated with increased patient's delay. *Eur. J. Cancer* 154, 269–276. doi:10.1016/j.ejca.2021.06.020

Saelens, W., Cannoodt, R., and Saeys, Y. (2018). A comprehensive evaluation of module detection methods for gene expression data. *Nat. Commun.* 9, 1090. doi:10.1038/s41467-018-03424-4

Schulte-Sasse, R., Budach, S., Hnisz, D., and Marsico, A. (2021). Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. *Nat. Mach. Intell.* 3, 513–526. doi:10.1038/s42256-021-00325-y

Shen, R., Mo, Q., Schultz, N., Seshan, V. E., Olshen, A. B., Huse, J., et al. (2012). Integrative subtype discovery in glioblastoma using iCluster. *PLoS One* 7, e35236. doi:10.1371/journal.pone.0035236

Shen, R., Olshen, A. B., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 25, 2906–2912. doi:10.1093/bioinformatics/btp543

Si, M., and Lang, J. (2018). The roles of metallothioneins in carcinogenesis. *J. Hematol. Oncol.* 11, 107. doi:10.1186/s13045-018-0645-x

Silverbush, D., Cristea, S., Yanovich-Arad, G., Geiger, T., Beerenwinkel, N., and Sharan, R. (2019). Simultaneous integration of multi-omics data improves the identification of cancer driver modules. *Cell Syst.* 8, 456–466. doi:10.1016/j.cels.2019.04.005

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11, 333–337. doi:10.1038/nmeth.2810

Xu, Z., Qu, H., Ren, Y., Gong, Z., Ri, H. J., and Chen, X. (2021). An update on the potential roles of E2F family members in colorectal cancer. *Cancer Manag. Res.* 13, 5509–5521. doi:10.2147/CMAR.S320193

Yang, Z., and Michailidis, G. (2016). A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics* 32, 1–8. doi:10.1093/bioinformatics/btv544

Yin, R., Song, B., Wang, J., Shao, C., Xu, Y., and Jiang, H. (2022). Genome-wide association and transcriptome-wide association studies identify novel susceptibility genes contributing to colorectal cancer. *J. Immunol. Res.* 2022, 5794055. doi:10.1155/2022/5794055

Zhang, J., Zhang, S., Wang, Y., and Zhang, X. S. (2013). Identification of mutated core cancer modules by integrating somatic mutation, copy number variation, and gene expression data. *BMC Syst. Biol.* 7 (2), S4. doi:10.1186/1752-0509-7-S2-S4

Zhang, Q., Liu, W., Zhang, H. M., Xie, G. Y., Miao, Y. R., Xia, M., et al. (2020). hTFtarget: A comprehensive database for regulations of human transcription factors and their targets. *Genomics Proteomics Bioinforma.* 18, 120–128. doi:10.1016/j.gpb.2019.09.006

Zhang, S., Li, Q., Liu, J., and Zhou, X. J. (2011). A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. *Bioinformatics* 27, i401–i409. doi:10.1093/bioinformatics/btr206

Zhang, S., Liu, C. C., Li, W., Shen, H., Laird, P. W., and Zhou, X. J. (2012). Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res.* 40, 9379–9391. doi:10.1093/nar/gks725