SCIENTIFIC REPORTS
nature research

OPEN

Identification of Key Components in Colon Adenocarcinoma Using Transcriptome to Interactome Multilayer Framework

Ehsan Pournoor[1], Zaynab Mousavian[2], Abbas Nowzari Dalini[2] & Ali Masoudi-Nejad[1*]

Complexity of cascading interrelations between molecular cell components at different levels from genome to metabolome ordains a massive difficulty in comprehending biological happenings. However, considering these complications in the systematic modelings will result in realistic and reliable outputs. The multilayer networks approach is a relatively innovative concept that could be applied for multiple omics datasets as an integrative methodology to overcome heterogeneity difficulties. Herein, we employed the multilayer framework to rehabilitate colon adenocarcinoma network by observing co-expression correlations, regulatory relations, and physical binding interactions. Hub nodes in this three-layer network were selected using a heterogeneous random walk with random jump procedure. We exploited local composite modules around the hub nodes having high overlay with cancer-specific pathways, and investigated their genes showing a different expressional pattern in the tumor progression. These genes were examined for survival effects on the patient's lifespan, and those with significant impacts were selected as potential candidate biomarkers. Results suggest that identified genes indicate noteworthy importance in the carcinogenesis of the colon.

One of the most prominent aspects of computational biology is network biology in which network science plays an essential role in discovering biological occurrences. Networks are reconstructed from different biological phenomena and deduced from mathematical and topological aspects. Protein-protein interaction (PPI), metabolic, and regulatory networks are examples of biological networks that have frequently been employed in recent years. However, nature comes with a lot of complexity. Inside a cell, there is a broad spectrum of entities contributing to create different cascading interactions that define the cell life and function. Considering these complexities in the modeling will gain comprehensive and valuable insights on biological events.

Currently, thanks to progress in high-throughput sequencing technologies and large projects such as TCGA[1], there is a booming amount of omics data that could be used for wide-ranging analyses. Curated databases contain information for regulatory interactions between biomolecules and their targets, single nucleotide polymorphisms (SNPs), biological pathways, and gene expression profiles of various phenotypes. Moreover, tools and applications, developed by scientific societies, are increasingly accessible, resulted in an easier exploration of biological happenings[2]. However, these data are heterogeneous, inconsistent, and provider technologies may come with a bias[3]. To overcome these problems, data integration would be an asset. Recently, integrative systems biology has become a popular area in which more than just a single type of biologic data is incorporated[4–7]. Bearing these considerations in mind, the results will be more truthful and reliable. In an excellent review article written by Koyel *et al.*[8], they described different approaches in integrative biological networks. Also, Peng *et al.*[9] proposed a new multi-omics approach for bladder cancer-related genes discovery.

Systematically, multi-omics datasets could be regarded as multilayer networks. Based on scientific definitions[10,11], a multilayer network contains multiple layers (different layers for different types of interactions) and considering the topology of layers, various kinds of multilayer networks are characterized. As Hmimida *et al.*[12] mentioned, multilayer networks exploration can be executed in three ways. First, Layer Aggregation, in which all layers are aggregated to make a single network and traditional single layer (monoplex) analysis could be applied

[1]Laboratory of Systems Biology and Bioinformatics (LBB), Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran. [2]School of Mathematics, Statistics, and Computer Science, College of Science, University of Tehran, Tehran, Iran. *email: amasoudin@ut.ac.ir

to explore it. Second, Ensemble (Consensus) approaches, in which each layer is individually evaluated; then, the results are combined to create the final consequence. Third, methods extended for multilayer networks (briefly called extended approaches), in which the analysis process is simultaneously conducted on all layers. Didier *et al.*[13] compared these three approaches in terms of community detection and found that the extended modularity function has superiority over the other two methods.

The extension of topological attributes from monolayer to multilayer is a critical and challenging topic in this area[12,14–16]. Hmimida *et al.*[12] have defined metrics (such as degree, shortest-path, neighbor set) for multiplex networks using an entropy-like aggregate function. Domenico *et al.*[16] proposed reducibility methods for multilayer networks to eliminate redundant interactions and layers. In this context, community detection for multilayer networks is considered one of the most challenging topics. Given the topological perspective, a community is a cluster of densely connected nodes, which are far from other clusters. Communities may be either local or global and may have overlap with each other. Recently, various extended multilayer community detection algorithms have been proposed to seek modules in layers simultaneously[17–21]. A specific type of community detection method is based on seed-centric approach, in which communities are localized around predefined (manual or computational) seed nodes[12,22].

Extended approaches for multilayer networks were recently used in biological and medical sciences. Berenstein *et al.*[23] have taken benefit from these methods for the application of drug repositioning in neglected diseases. Furthermore, Rai *et al.*[24] found a similar structure in the PPI network of seven types of cancers using spectral graph theory and the multilayer framework. Although these kinds of integrative methods were used recently in some contexts of biology, there is still a gap in their usage in prognosis and diagnosis of human disorders such as cancer even with the high availability of omics datasets. Also, because of the complexity of this kind of research with high dimensional data, previous multilayer-based works have relied mostly on the usage of two types of data and missed complementary interactions such as regulatory links in transcriptional and post-transcriptional phases.

Here, we utilize an extended approach to find functional communities in colon adenocarcinoma (COAD). To model such a multifaceted system in a realistic and comprehensive way, three levels of abstraction are declared. First, at the transcriptional level, gene correlations can be defined to represent the co-expression patterns among the genes. Second, at the post-transcriptional level, biomolecules such as RNAs and proteins have regulatory interrelations. Regulatory interactions indicate direct or indirect control of gene expression. Third, physical interactions show bindings of molecules such as proteins or RNAs to other molecules (such as protein-protein and RNA-protein bindings). In this arrangement, it is possible that a regulatory interaction may also be a physical binding interaction. Accordingly, we construct a three-layer network with co-expression, regulatory and physical interaction layers. One of the novelties addressed in this study is utilizing the most diverse types of interactions including co-expression, signaling, kinase-substrate, metabolic enzyme-coupled, nucleoproteins, protein complexes, RNA-RNA, and regulatory in a multi-layer framework to get a holistic view of the genes involved in carcinogenesis. The analysis program in this research contains the following steps to achieve potential biomarkers derived from raw datasets. We employ a local seed-centric community detection algorithm to explore modules in the multilayer network. In this process, to select seed items computationally, we propose an innovative multilayer heterogeneous random walker to score nodes. The top-ranked nodes are applied as seeds, and local communities around these seed nodes are computed as modules of interest. Out of the identified modules, those with a high overlap with COAD, based on validated databases, are selected as candidate modules for further steps. For every module, differential expression (DE) analysis is performed to extract differentially expressed genes (DEGs). We conduct the survival analysis for DEGs in the final step, to find genes that their differential expression influences the survival of patients with COAD. Modules containing a large amount of these genes are selected as final modules, and the functional enrichment was carried out on them. Finally, we discussed the upstream regulators of candidate genes to specify their role in the differential expression of target candidates and the related biological pathways. This kind of novel evaluation led us to identify a list of new potential targets in colorectal cancer[25].

Materials and methods

Data collection and preprocessing. In the process of data collection, the information needed to build a multilayer network was gathered from multiple sources. As presented in the research workflow (Fig. 1), in this step, three types of data were used: gene expression profiles, regulatory relationships, and physically binding interactions. First, to build the co-expression network (layer 1), RNAseq data for COAD were exploited from the TCGA data portal[26]. In the preprocessing of expression data, we performed sample and gene filtering and then used the FPKM-UQ (Fragments Per Kilobase of transcript per Million mapped reads upper quartile) normalized data to generate the network. In the gene filtering step, genes with the following conditions were excluded: (1) missing values in any samples, (2) the expression count value of zero in more than 80% of all samples[27], (3) genes that possessed the expression rates with zero standard deviation across all samples, and (4) genes with average CPM (count per million) lower than 1.

Second, for the generation of regulatory layer, we utilized the experimentally curated regulatory interactions using the NPInter (non-coding RNA interactions with biomolecules)[28], the RegNetwork[29] (Regulatory Network Repository of Transcription Factor and microRNA Mediated Gene Regulations), and the TRRUST (database of TF-target regulatory information)[30] databases. Third, the network of physical interaction layer was assembled using the interactome data, as previously published[31]. These data comprise different interactions including (1) regulatory interaction of transcription factors, (2) yeast two-hybrid (Y2H) binary interactions, (3) metabolic enzyme-coupled interactions, (4) signaling interactions, (5) protein complexes, (6) kinase-substrate interactions, and (7) low-throughput manually curated interactions in the literature. In addition to the co-expression layer, in other layers (regulatory and physical interaction), genes possessing the average expression value of zero in

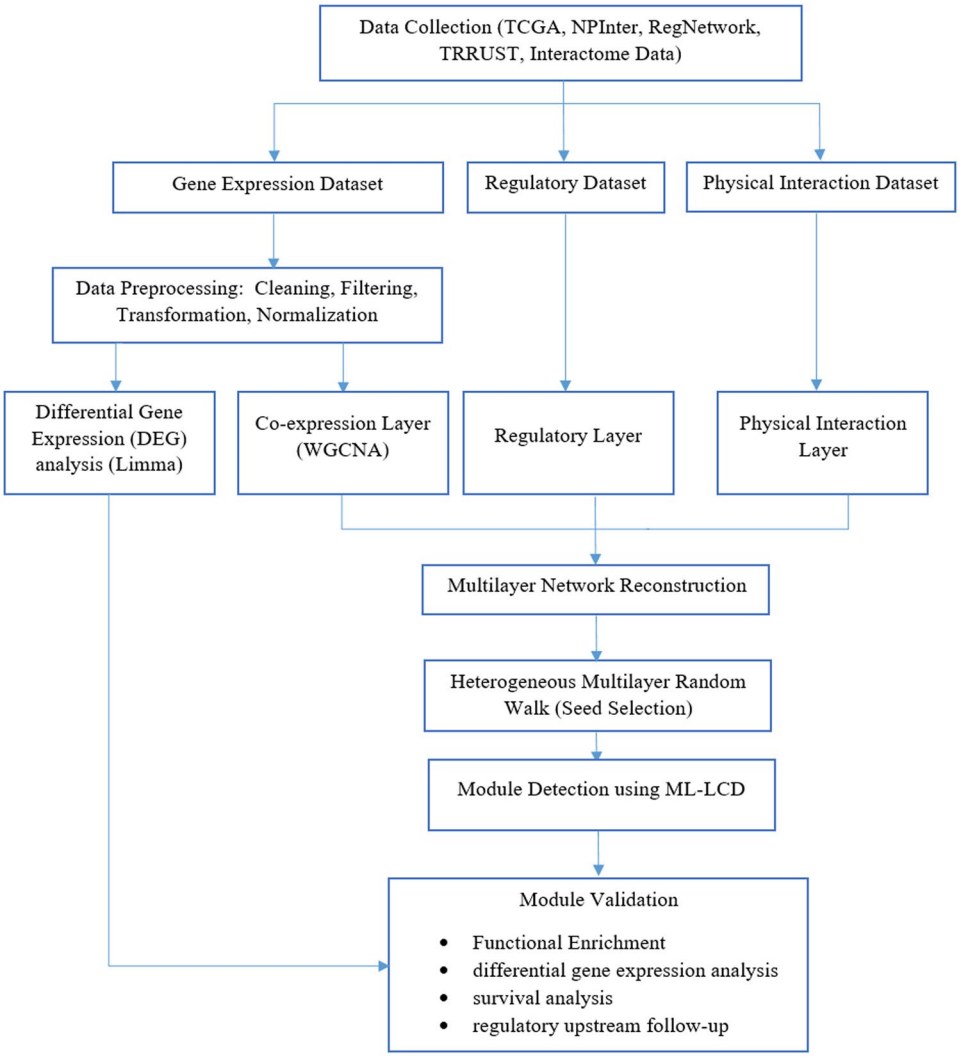

Figure 1. The Workflow of the research.

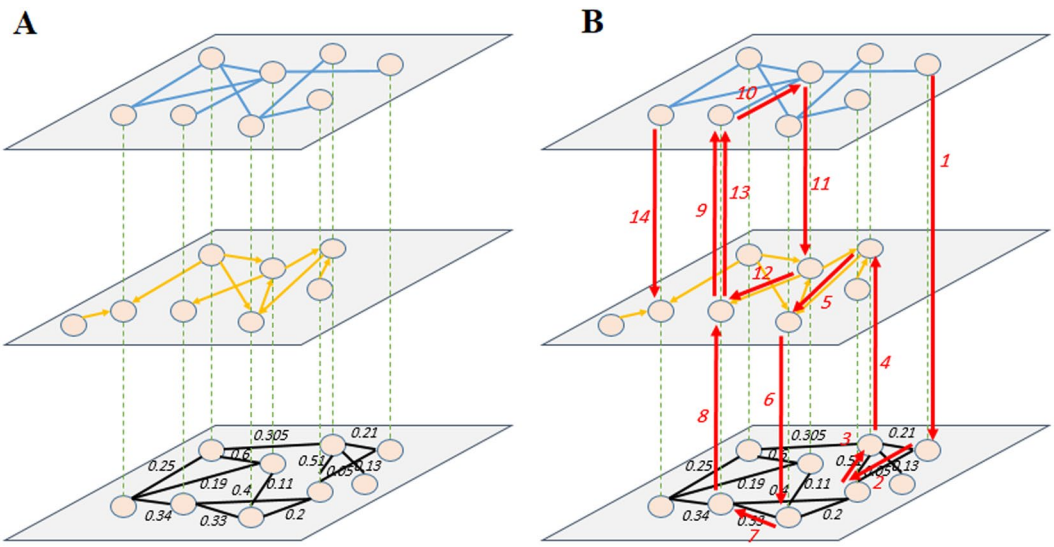

Figure 2. (**A**) A sample heterogeneous multilayer network of the transcriptome to interactome. (**B**) A typical random walk with the random jump.

transcriptomic data were also removed. Our reason is that genes with no expression will never be translated into proteins, and they have no regulatory function or physical binding interactions.

Multilayer network construction. Herein, we used FPKM-UQ normalized data and established the co-expression network using the WGCNA[27] package in the R. In this layer; we selected those correlations having the values higher than 0.75 (considering scale-freeness of the network). Since the gene-set is made of both coding and non-coding RNAs and contains regulators and their targets, in the WGCNA "adjacency" function, we set the network type to "unsigned," and correlations were calculated using the Pearson Correlation Coefficient (PCC) measure. The output is a weighted and undirected network in which the edge weights address correlations between the genes. We built other layers of the multilayer network (regulatory and physical binding layers) using interactions deduced by the experimentally curated databases. The regulatory network is directed, while the physical interaction network is undirected; however, both of them are unweighted networks. Figure 2A depicts such a heterogeneous multilayer network.

Personalized seed selection. In most of the seed-centric community detection approaches, the seeds are either selected manually (such as predefined nodes) or based on their specific properties (such as being a hub). Here, to select seeds in a sophisticated way, nodes were ranked using a random walk with a random jump procedure. To attain this objective, we customized the random walker to walk in all three layers considering the heterogeneity. Since the network topology (node set, edge set, directedness, and edge weighting) is different in each layer (Fig. 2A), we proposed a walker algorithm (algorithm 1) to resolve this kind of ambiguity.

Algorithm 1. Proposed heterogeneous multilayer random walker.

```
Input: personalization_dict, multilayer_network, iteration_count, random_jump_prob
Output: score_dict

for node in all layers:
    score_dict[node] = 0
node = choose_node(personalization_dict)
for 1 → iteration_count:
    layer = choose_layer_having_node(multilayer_network, node)
    score_dict [node] += 1
    jump = jump_or_stay(random_jump_prob)
    if jump:
        node = choose_node(personalization_dict)
    else:
        node = move_to_a_neighbor(node, layer)
normalize(score_dict)
return score_dict
```

The walker must be able to move in all three layers with the ability to change layers and nodes. Figure 2B shows a sample walk with node and layer exchange. In this procedure, moving to a neighbor follows the probability function $P_l(i, j)$ defined in Eq. (1) for unweighted and Eq. (2) for weighted networks.

As defined by Kivelä *et al.*[10], given a set of nodes $\mathcal{V}$ and a set of layers $\mathcal{L} = \{L_1, L_2, \ldots, L_s\}, V_{\mathcal{L}} \subseteq \mathcal{V} \times \mathcal{L}$ specifies nodes and $E_{\mathcal{L}} \subseteq V_{\mathcal{L}} \times V_{\mathcal{L}}$ denotes the edge list (including inter-layer and intra-layer edges) in a multilayer framework. We indicated the multilayer network graph with notation $G_{\mathcal{L}} = (V_{\mathcal{L}}, E_{\mathcal{L}}, \mathcal{V}, \mathcal{L})$. In this context, the probability of proceeding to a next neighbor in unweighted networks (regulatory and physical interaction layers) is defined as:

$$P_l(i, j) = \frac{1}{\sum_{k \in m_l} A_l(i, k)}, \quad A_l(i, j) = \begin{cases} 1 & \text{if there is an edge from node } i \text{ to } j \\ 0 & \text{else} \end{cases} \tag{1}$$

where $P_l(i, j)$ is the probability of proceeding from node i to any other neighbor (out-neighbors) j in layer l, m_l indicates the neighbors (or out-neighbors if the network is directed) of node i in layer l, and A_l demonstrates the adjacency matrix of the layer l. However, in the co-expression layer, we set the walk probabilities as:

$$P_l(i, j) = \frac{A_l(i, j)}{\sum_{k \in m_l} A_l(i, k)}, \quad A_l(i, j) = \begin{cases} Corr_l(i, j) & \text{if there is an edge from node } i \text{ to } j \\ 0 & \text{else} \end{cases} \tag{2}$$

where $Corr(i, j)$ implies the co-expression correlation between nodes (genes) i and j. Also, the walker must be able to go from a node to its counterparts in other layers. For instance, when the walker is at node x inside layer L_1, for any layer $L_r \in \mathcal{L}$, $r \neq 1$, if L_r contains node x, the walker can jump to node x in L_r, or stay in L_1 itself. In each step, the walker should choose to go to one of its counterparts in other layers (line 7 in algorithm 1) or move to one of its neighbors in the same layer (lines 13 in algorithm 1). We set the probability of moving to a neighbor in the same layer as same as moving to its counterpart in a different layer to enable the walker to score nodes based on shared properties of nodes in all layers in an unbiased manner. Moreover, to avoid tangle in traps, we set the walker to jump with a predefined probability βor continue to walk with a probability of $1 - \beta$ (line 10 in algorithm 1). To be fair in the selection of jump destination and ascribe it to the biological gene expression, we set the personalized random jump probability as:

$$pr(i) = \frac{\exp(i)}{\sum_{k=1}^{N} \exp(k)} \tag{3}$$

where $pr(i)$ specifies the probability to jump to node i, $\exp(i)$ indicates the average expression of gene i across the samples of patients with cancer and N is the total number of genes (nodes) in the multilayer framework.

Local seed-centric module detection. We used a local community detection approach to detect the functional composite modules around the seed nodes in our multilayer network. The idea underlying this approach is to find items that contribute to the development of carcinogenesis. Items are scattered in layers, and their nature might be a gene, RNA, or protein; however, we consider the equivalent gene names in the final module. Since currently there is not an applicable community detection algorithm about large heterogeneous multilayer networks compatible with our network; in this phase, we considered layers in an unweighted and undirected manner. We utilized ML-LCD local community detection method[22] to discover modules for every seed node. It is a modularity expansion-based community detection method, which has better performance in large multilayer networks. Given a seed node $v_0 \in \mathcal{V}$, ML-LCD algorithm attempts to find a subgraph $G_{\mathcal{L}}^{v_0} \subseteq G_{\mathcal{L}}$ that consisted of the seed node v_0 and maximizes the local community function (LC) in Eq. (4). For simplifications, the letter C is used to denote local community subgraph. In this regard, $E^C \subseteq E_{\mathcal{L}}$ demonstrates the edge set of the subgraph.

$$G_{\mathcal{L}}^{v_0} = \underset{C = (V, E, \mathcal{V}, \mathcal{L}) \subseteq G_{\mathcal{L}} \wedge v_0 \in V}{argmax\ LC(C)} \tag{4}$$

where:

$$LC(C) = \frac{LC^{int}(C)}{LC^{ext}(C)} \tag{5}$$

In Eq. (5), $LC^{int}(C)/LC^{ext}(C)$ indicates the density of links inside C, over the density of links between nodes inside and outside of the C. To formulate local community function LC, terms *Shell Nodes* and *Boundary Nodes* should be defined. In this regard, to specialize the edges inside the community C in a layer L_i, symbol $E_i^C = \{(u, v) | \exists ((u, L_i), (v, L_i)) \in E^C\}$ will be used. For a local community being constructed, the shell nodes refer to nodes outside of the community that is a neighbor of nodes inside the community (displayed by symbol S), and these within-community neighbors of shell nodes are also called boundary nodes (depicted by symbol B):

$$S = \{v \in \mathcal{V} \backslash C | \exists ((u, L_i), (v, L_j)) \in E_{\mathcal{L}} \wedge u \in C\}$$

$$B = \{v \in C | \exists ((u, L_i), (v, L_j)) \in E_{\mathcal{L}} \wedge v \in S\}$$

Furthermore, $E^B = \{(u, v) | \exists ((u, L_i), (v, L_j)) \in E_{\mathcal{L}} \wedge u \in B \wedge v \in S\}$ indicates the set of edges outgoing from the C, and for any layer L_i, $E_i^B = \{(u, v) | \exists ((u, L_i), (v, L_i)) \in E^B\}$ is the portion of E^B corresponding to edges of layer L_i. As reported by[22], $LC^{int}(C)$ and $LC^{ext}(C)$ are defined as:

$$LC^{int}(C) = \frac{1}{|C|} \sum_{v \in C} \sum_{L_i \in \mathcal{L}} \omega_i \left| E_i^C(v) \right| \tag{6}$$

$$LC^{ext}(C) = \frac{1}{|B|} \sum_{v \in B} \sum_{L_i \in \mathcal{L}} \omega_i \left| E_i^B(v) \right| \tag{7}$$

where ω_i (for every $L_i \in \mathcal{L}$) is non-negative coefficient, with $\sum_{L_i \in \mathcal{L}} \omega_{L_i} = 1$, demonstrating layer weights.

Module validation and biomarker extraction. Communities detected by the mentioned approach should be validated by a biological viewpoint. Our goal was to seek modules (communities) that their genes were involved in the malignant tumor of the colon. We chose modules possessing considerable counts of genes associated with COAD according to the DisGeNet[32] and containing a significant number of DEGs. For these disease-related modules, functional enrichment analysis is a strategy to check the role of module genes in pathways and processes. To enrich the discovered modules, we employed two popular tools, the Enrichr[33,34] and

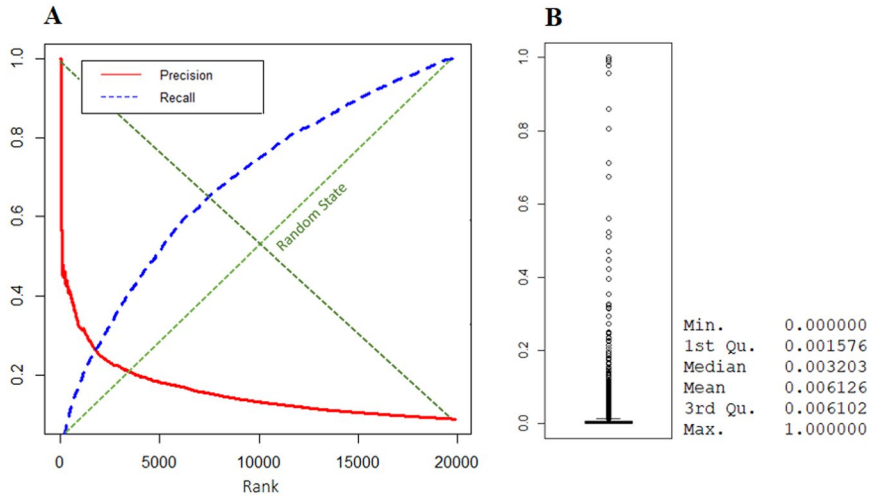

Figure 3. (**A**) Evaluation of walker in terms of precision and recall for COAD. (**B**) Distribution of nodes score.

#	Seed	Seed RW score	number of module genes	Involvement percentage	p-value	FDR B&H
1	SP1	0.977729191	239	0.414	2.667E-33	3.54E-30
2	EGR1	0.858748157	250	0.464	3.344E-44	3.31E-41
3	USF1	0.805867697	153	0.614	5.507E-48	2.97E-44
4	YY1	0.674073321	140	0.364	4.201E-15	7.65E-13
5	E2F1	0.559109168	156	0.25	6.14E-07	3.27E-05
6	MXI1	0.523567397	63	0.302	0.00003692	0.001946
7	JUN	0.511059135	140	0.457	2.984E-25	3.84E-22

Table 1. Explored modules around seed nodes with an overlap percentage higher than 0.25 in COAD.

ToppFun enrichment portal of the ToppGene[35], and studied the presence of modules genes in tumor-associated pathways and biological processes. We selected the KEGG[36] and Reactome[37] as reference pathway databases.

For the genes inside the discovered modules, that were not previously reported as COAD-associated genes in the DisGeNet database, differential gene expression analysis was performed. Herein, we focused on genes that their expression pattern statistically differs in the malignant tumor. We did DE analysis using the Limma[38] and the edgeR[39] R packages based on workflow presented in[40] and selected genes with larger fold changes ($|\log(FC)| > 1$) and smaller p-value (adj. p-value < 0.01).

To observe the effects of the expression level of discovered DEGs on the survival of patients, the survival analysis was carried out. This examination was applied to compare the lifespan of people when the expression of a gene differs from the normal. Kaplan-Meier (KM)[41] curve is a worthy choice when the data are censored, and there is not complete information on subjects. It estimates the lifespan of a group of people having a low gene expression rate in comparison to another group with a high expression rate. The SurvExpress[42] utility was used for the log-rank test and Kaplan-Meier survival analysis. In this phase, we chose those genes that a change in their expression rates influences the survival rate of patients with COAD and considered them as potential biomarkers. Although the impact of variations on the expression of upstream regulators is not the only reason for changes in the expression of downstream genes, it provides some explanations for these alterations. Finally, to validate changes in the transcription levels of biomarkers, in the regulatory layer, we performed a regulation follow-up on upstream items of biomarkers.

Results

Data preparation and network generation. The obtained dataset from the TCGA includes total RNAseq expression data for 60483 coding/non-coding RNAs in 424 samples that consisted of samples collected from healthy subjects and patients with colon cancer. Thus, we selected the paired samples that comprised of 49 samples obtained from healthy individuals and their cancer counterparts (49 normal specimens and 49 cancerous samples). Then, we carried out gene filtering as described in the "Materials and Methods" section. The cleaned data include the expression count value for 14515 genes in 98 samples. The 49-paired samples were used in the differential gene expression analysis. However, only the expression data of 49 cancerous samples were applied for the co-expression network construction. The generated co-expression layer comprises 5993 nodes and 75121 edges. On the other side, the regulatory and physical interaction layers were directly generated from source datasets. The physical interaction layer has 12751 nodes and 135712 edges, and the regulatory layer encompasses 17640 nodes with 133180 edges.

Module Seed	Min	Max	Avg.	Median	Involvement %
USF1	0.0061	0.80587	0.05692	0.03458	61%
YY1	0.00422	0.67407	0.03989	0.02677	46%
JUN	0.00158	0.51106	0.03635	0.0257	45%
EGR1	0.00219	0.85875	0.03331	0.02212	41%
SP1	0.00214	0.97773	0.03095	0.01978	36%
MXI1	0.00422	0.52357	0.02937	0.01693	30%
E2F1	0.00417	0.55911	0.01669	0.0091	25%
All Genes	0	1	0.00613	0.0032	—

Table 2. Module relevance to COAD in terms of Random walk scores.

	Module Seed	USF1	YY1	JUN	EGR1	SP1	MXI1	E2F1
1	# module genes	153	140	140	250	239	63	156
2	# COAD genes in 1	94 (61%)	51 (36%)	64 (45%)	116 (46%)	99 (41%)	19 (30%)	39 (25%)
3	# DEGs in 1	21 (13%)	8 (5%)	11 (7%)	40 (16%)	35 (14%)	1 (1%)	130 (83%)
4	# DEGs in (1–2)	3 (5%)	4 (4%)	3 (3%)	12 (8%)	14 (10%)	1 (2%)	100 (85%)
5	# Survival Candidate in 4	1	0	0	0	0	0	9

Table 3. Module comparison and evaluation. The first row demonstrates the number of genes inside each module. For each module, its overlap with COAD has been presented in the second row. Values in the third row are the count of DEGs inside each module. However, the fourth row shows the number of DEGs in module genes, which were not previously reported as COAD-related genes in databases. Additionally, the number of genes with significant effects on patients' survival (according to the fourth row) are presented in the fifth row.

Personalized seed selection. We ranked nodes in the network using the proposed heterogeneous multilayer random walker. The walker was set to jump with a probability of $\beta = 0.2$ or continue to walk with a probability of 0.8. We ran the walker for three million moves and ranked nodes based on scores (Min. = 0.0, 1st Qu. = 0.0015, Median = 0.0032, Mean = 0.00612, 3rd Qu. = 0.00610, Max = 1.0). Although the goal of this project was far from the gene prioritization, to test the accuracy of our personalized seed selection algorithm against a random selection of genes, we examined the result of the walker in terms of precision and recall (Fig. 3). Precision (Eq. 8) measures the percentages of retrieved items that are relevant. However, recall (Eq. 9) evaluates the percentage of relevant items that have been retrieved. Results show that the proposed ranking system performs much better than the random assortment (diagonals of the chart). The complete result of random walker scores is provided in Supplementary File S1. We designated nodes with score larger than 0.5 ($rw\_score > 0.5$) as seed nodes. As reported by the DisGeNET database, 11 out of 12 chosen seed nodes are involved in COAD, as previously mentioned, indicating that our computationally explored centroid seeds are biologically meaningful.

$$precision = \frac{|\{relevant\ items\} \cap \{retrieved\ items\}|}{|\{retrieved\ items\}|} \tag{8}$$

$$recall = \frac{|\{relevant\ items\} \cap \{retrieved\ items\}|}{|\{relevant\ items\}|} \tag{9}$$

Module detection. For the selected seed nodes, we calculated local communities using the ML-LCD method. Since there is not any precedence on the three layers, we set layer weight $\omega_{L_i} = 1/3$ for all layers equally. We selected modules whose genes (at least 25%) are known as predisposing genes in the development of COAD according to the reports by curated databases. Out of 12 modules found in this step, seven were related to colon cancer with an overlap more than the specified threshold (Table 1). The detailed information of 12 modules and their genes are available in Supplementary File S2.

Also, we checked random walk scores of module items separately to observe if there is a relationship between the module score and the percentage of its overlap with COAD. In all validated modules (seven modules), the minimum, maximum, median, and average scores were calculated (Table 2). It was observed that the average module score was highly associated with the involvement of module in COAD. For example, for the module with the seed gene *USF*1 (called module *USF1*), the average score of module items was the highest (avg. score = 0.05692) and it was the most related module to cancer (overlap = 61%). It emphasizes that COAD-related modules contain nodes with a high random walk score (hub nodes in the multilayer framework).

DEG analysis and module evaluation. To assess each detected module, we applied differential gene expression analysis, with limitations adj. p-value < 0.01 and $|\log(FC)| > 1$. As presented in Table 3, the module *E2F1* has the highest value of differentially expressed genes (100 DEGs from 117 genes). We started with DEGs

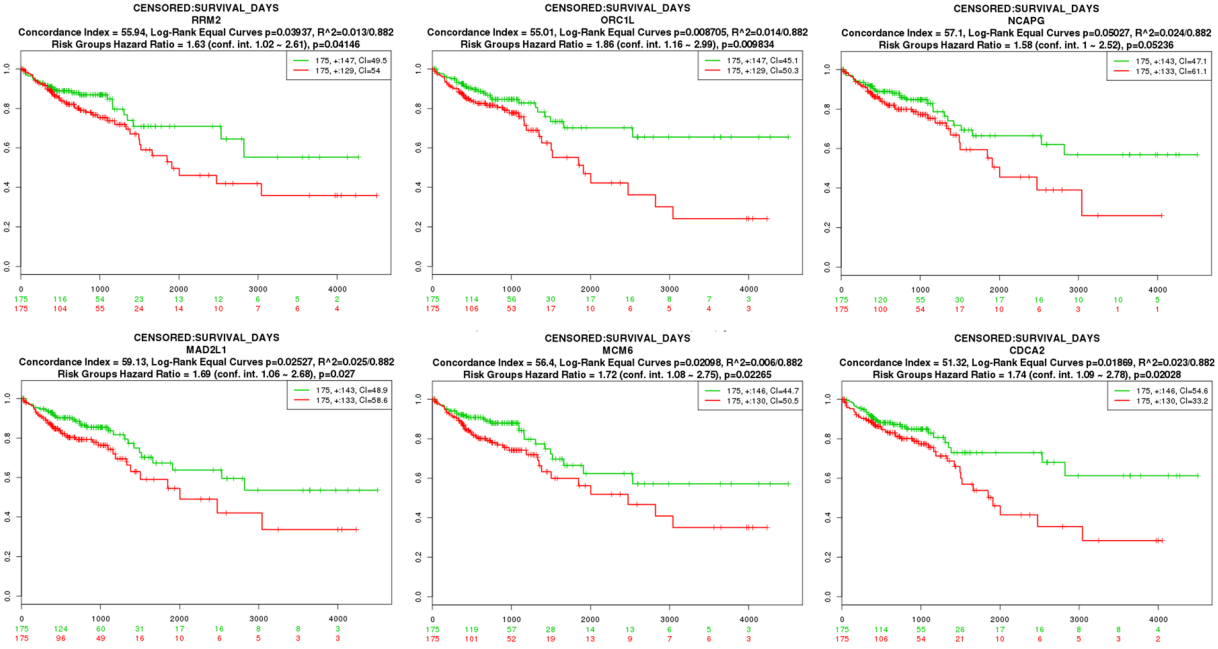

Figure 4. Kaplan-Mier curves (from SurvExpress tools) for candidate biomarkers *RRM2*, *ORC1*, *NCAPG*, *MAD2L1*, *MCM6*, and *CDCA2* selected with log-rank test p-Value < 0.05.

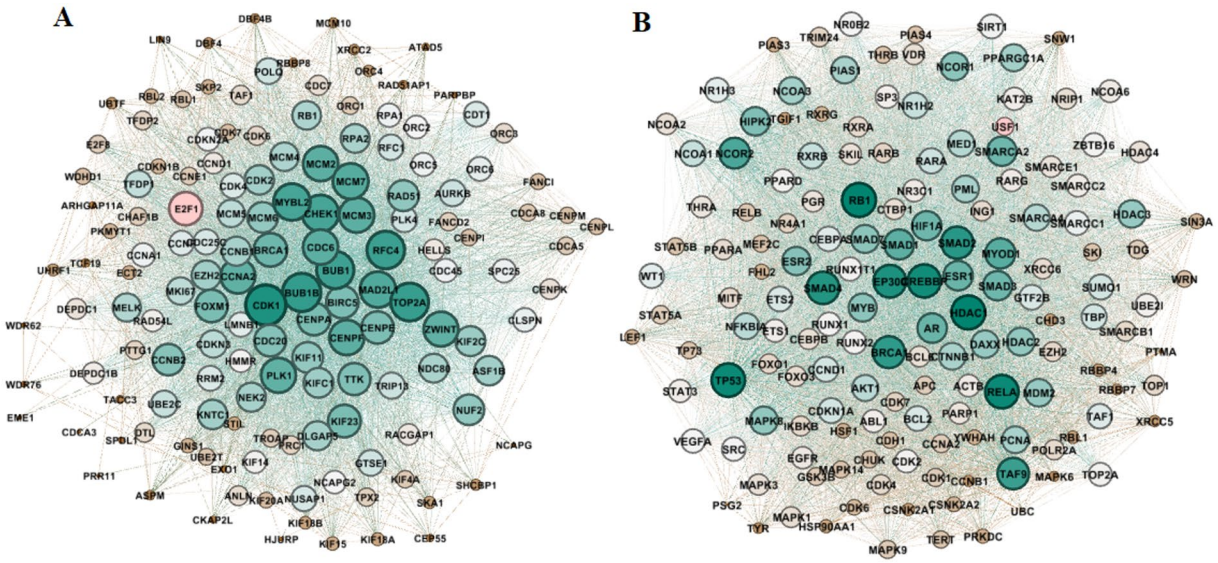

Figure 5. FLN subnetworks for modules with seed nodes (**A**) *E2F1* and (**B**) *USF1*. Nodes' sizes are based on their degree.

that their expression rate was changed in the transition from healthy to cancer state, and continued to survival analysis. Among the identified DEGs that were not reported in databases, 10 genes had different expression patterns (modules with seed genes *E2F1* and *USF1*), resulting in a different survival rate with log-rank test p-value $<$ 0.05 (Table 3, Fig. 4). The genes *CDC6, RRM2, ORC1, NCAPG, MAD2L1, MCM6, CCNF, CDCA2, ECT2*, and *DEPDC1B*, are the output candidate genes identified in explored modules, which are differentially expressed between normal and cancer states, and their differential expression levels have noteworthy effects on the lifespan of patients. Since we looked for novel unreported genes, we ignored modules with low coverage of DEGs and survival biomarkers and proceed with two significant modules, the module *E2F1* (high DEG and survival rate) and the module *USF1* (high overlap) for functional enrichment analysis.

In another parallel process, we performed another tryout to validate our approach in finding differentially expressed genes inside detected modules. For this purpose, the proportional test was used to check the accuracy of our method in distinguishing genes that their expression varies between healthy samples and patients with COAD. In other words, this test was designed to answer the question of whether the proposed approach would be

Term	P-value	Adj. P-value	Z-score
Androgen receptor signaling pathway WP138	6.79E-58	2.17E-55	−1.38815
TGF-beta Signaling Pathway WP366	7.66E-38	1.23E-35	−1.2601
Nuclear Receptors WP170	6.06E-31	4.85E-29	−1.53161
Adipogenesis WP236	8.28E-30	5.3E-28	−1.04883
Cell Cycle WP179	1.46E-27	7.77E-26	−1.11229
AGE/RAGE pathway WP2324	2.48E-25	1.13E-23	−1.9584
Non-small cell lung cancer WP4255	1.44E-23	4.62E-22	−1.53125
RAC1/PAK1/p38/MMP2 Pathway WP3303	2.65E-23	7.06E-22	−1.39478
Oncostatin M Signaling Pathway WP2374	5.77E-22	1.35E-20	−1.85675
Circadian rhythm related genes WP3594	5.89E-22	1.35E-20	−0.88343
TGF-beta Receptor Signaling WP560	1.01E-21	2.16E-20	−2.33059
VEGFA-VEGFR2 Signaling Pathway WP3888	1.54E-21	3.09E-20	−0.79477
DNA Damage Response (only ATM dependent) WP710	1.04E-20	1.97E-19	−1.06194
TNF related weak inducer of apoptosis (TWEAK) Signaling Pathway WP2036	5.64E-20	9.49E-19	−2.38891
Brain-Derived Neurotrophic Factor (BDNF) signaling pathway WP2380	8.93E-20	1.43E-18	−0.86385
Chromosomal and microsatellite instability in colorectal cancer WP4216	2.24E-19	3.41E-18	−1.18043
Energy Metabolism WP1541	3.53E-19	5.13E-18	−1.95291
Leptin signaling pathway WP2034	4.51E-19	6.28E-18	−1.01867
IL-4 Signaling Pathway WP395	3.2E-18	4.1E-17	−1.68254
RANKL/RANK (Receptor activator of NFKB (ligand)) Signaling Pathway WP2018	4.26E-18	5.25E-17	−1.67297
ErbB Signaling Pathway WP673	9.84E-18	1.12E-16	−1.27747
Aryl Hydrocarbon Receptor WP2586	1.5E-17	1.65E-16	−1.90374
Thymic Stromal LymphoPoietin (TSLP) Signaling Pathway WP2203	2.06E-17	2.2E-16	−2.11
Wnt/beta-catenin Signaling Pathway in Leukemia WP3658	2.55E-17	2.64E-16	−2.43499
Nuclear Receptors Meta-Pathway WP2882	3.15E-17	3.15E-16	−0.67084
DNA Damage Response WP707	1.11E-16	1.08E-15	−1.09236
miRNA Regulation of DNA Damage Response WP1530	2.12E-16	1.94E-15	−1.33999
Non-genomic actions of 1,25 dihydroxyvitamin D3 WP4341	2.12E-16	1.94E-15	−0.88709
Vitamin D in inflammatory diseases WP4482	3.05E-16	2.71E-15	−2.36581
Corticotropin-releasing hormone signaling pathway WP2355	4.19E-16	3.62E-15	−1.07055
Integrated Cancer Pathway WP1971	4.43E-16	3.64E-15	−2.03412
Interleukin-11 Signaling Pathway WP2332	4.43E-16	3.64E-15	−2.02649

Table 4. Significant pathways for the module *USF1*.

sufficiently precise in discovering DEGs inside the identified modules. For this aim, the "One-sample proportions test with continuity correction" method was applied. This was executed through the *prop.test* function in the R. The *prop.test* can be used for examination of the null hypothesis if the proportions (probabilities of success) in several groups are the same, or they have certain equal values. Pearson's chi-square test statistic 16.87184 with degree freedom of 1 and a p-value of 3.99906e-05 demonstrates relatively high precision in the detection of DEGs in our proposed module detection approach.

Functional enrichment. The module *E2F1* has the highest number of differentially expressed genes that have survival impact; however, it has a low coverage with respect to the predefined COAD genes. On the side, the module *USF1* covers more percentage of COAD-related genes but contains fewer DEGs and effects on the survival rate. To study these two modules deeply, we utilized the human Functional Linkage Network (FLN), in which each node is a protein, and there is an edge between two nodes if there is evidence that nodes have a degree of the functional similarity. In this network, edges are weighted and predicted based on PPI interactions, gene expression profiles, literature mining, experimental techniques, and computational approaches[43,44]. A common methodology for predicting function based on FLNs applies a simple local threshold rule (mentioned as 'guilt-by-association')[45]. For the genes inside modules that were not specified as COAD related genes in databases, to explore their role in neoplastic cellular processes, we used functional linkage similarities of those genes with previously annotated genes. In the full FLN network, edge weights demonstrate the functional similarity between proteins. We extracted the subnetworks of the two modules from the human FLN network provided that the edge weights were higher than the average of the whole network. The *E2F1* FLN (Fig. 5A) is a dense subnetwork with an average path length of 1.18 and a density of 0.41. Out of 156 genes inside the module, 148 genes have functional similarities above the average values with respect to other module members. We surveyed the direct neighbors (containing 12111 nodes) of the module *E2F1* (module boundary); of them, 1496 were specified as cancer-related genes (coverage = 0.78). The module *USF1* (Fig. 5B) has a high intersection with the COAD; however, its boundary neighbors also have a tight coverage with cancer-related genes (coverage = 0.84). It is a dense subnetwork with a density of 0.59 and a network diameter of 1.4. Among genes belonging to the module *USF1*,

Term	P-value	Adj. P-value	Z-score
Cell Cycle WP179	1.97965E-63	2.71E-61	−1.15315
G1 to S cell cycle control WP45	3.94348E-47	1.8E-45	−1.95642
DNA Replication WP466	2.39913E-38	8.22E-37	−2.09373
miRNA Regulation of DNA Damage Response WP1530	1.76872E-24	4.85E-23	−1.72269
DNA Damage Response WP707	3.80714E-23	8.69E-22	−1.32103
DNA IR-damage and cellular response via ATR WP4016	9.81255E-22	1.92E-20	−1.27862
ATM Signaling Pathway WP2516	1.88381E-11	2.87E-10	−2.23923
Integrated Cancer Pathway WP1971	4.75612E-11	6.52E-10	−2.23108
Regulation of sister chromatid separation at the metaphase-anaphase transition WP4240	9.65678E-10	1.1E-08	−2.3819
DNA IR-Double Strand Breaks (DSBs) and cellular response via ATM WP3959	1.02091E-08	9.99E-08	−2.07864
H19 action Rb-E2F1 signaling and CDK-Beta-catenin activity WP3969	5.12131E-08	4.68E-07	−2.84809
ID signaling pathway WP53	1.10338E-07	8.89E-07	−2.87544
Human Thyroid Stimulating Hormone (TSH) signaling pathway WP2032	8.12668E-07	5.3E-06	−1.74211
Tumor suppressor activity of SMARCB1 WP4204	3.90517E-06	2.43E-05	−2.13364
Non-small cell lung cancer WP4255	1.2629E-05	6.92E-05	−1.38824
DNA Mismatch Repair WP531	3.77788E-05	0.000185	−2.89841
Photodynamic therapy-induced AP-1 survival signaling. WP3611	4.32184E-05	0.000197	−1.95428
Wnt Signaling Pathway WP363	5.23527E-05	0.000231	−2.18808
TGF-beta Signaling Pathway WP366	8.00653E-05	0.000343	−0.93161
LncRNA involvement in canonical Wnt signaling and colorectal cancer WP4258	9.45935E-05	0.000393	−1.14694
Chromosomal and microsatellite instability in colorectal cancer WP4216	0.000265291	0.001069	−1.07313
Regulation of Wnt/B-catenin Signaling by Small Molecule Compounds WP3664	0.000292112	0.001143	−1.89391
Aryl Hydrocarbon Receptor WP2586	0.000450398	0.001668	−1.83167
Androgen receptor signaling pathway WP138	0.000698091	0.002391	−1.02597
Wnt/beta-catenin Signaling Pathway in Leukemia WP3658	0.001060871	0.00338	−2.25473
PPAR Alpha Pathway WP2878	0.001060871	0.00338	−1.32492
Wnt Signaling Pathway and Pluripotency WP399	0.001227751	0.003823	−0.87218
Extracellular vesicle-mediated signaling in recipient cells WP2870	0.001619214	0.004822	−1.99363
Association Between Physico-Chemical Features and Toxicity Associated Pathways WP3680	0.001763243	0.00514	−1.61192
ATR Signaling WP3875	0.002099432	0.005981	−2.88868
Adipogenesis WP236	0.00354635	0.009717	−0.68844
Homologous recombination WP186	0.004456552	0.011972	−2.24022
Endoderm Differentiation WP2853	0.005003582	0.013183	−0.7774
PI3K-Akt Signaling Pathway WP4172	0.005302493	0.013706	−0.26886
Regulation of Microtubule Cytoskeleton WP2038	0.005527788	0.014024	−1.48329
Senescence and Autophagy in Cancer WP615	0.009259877	0.023066	−0.91906
Wnt Signaling WP428	0.012618554	0.029806	−0.8459
Vitamin D Receptor Pathway WP2877	0.014158639	0.032877	−0.63668
IL-7 Signaling Pathway WP205	0.016124042	0.036213	−2.13899
RAC1/PAK1/p38/MMP2 Pathway WP3303	0.016118811	0.036213	−0.70315
AMP-activated Protein Kinase (AMPK) Signaling WP1403	0.016757067	0.037028	−0.97432
Apoptosis WP254	0.028018802	0.04767	−0.75476

Table 5. Significant pathways for the module *E2F1*.

153 genes were in its FLN subnetwork, demonstrating the biological and chemical similarities between them. The functional linkage subnetworks are biological evidence on the modularity and resemblance of their genes and suggest a strong possibility to constitute a module and act together in biological pathways.

Additionally, to evaluate the relevance of selected modules to the development of carcinogenesis, we performed functional enrichment analysis. Results showed a significant correlation between discovered modules and pathways involved in the malignant tumors of the colon. Therefore, we set a threshold of 0.05 for enrichment adj. p-value. The extracted significant pathways for modules *USF1 and E2F1* are listed in Tables 4 and 5, respectively. The detailed results for functional enrichment of the two modules are accessible in Supplementary File S3. It was found that both modules were involved in pathways such as *Adipogenesis*, *Cell Cycle*, *Chromosomal and microsatellite instability in colorectal cancer*, and *TGF-beta Signaling Pathway* that they play crucial roles in colorectal cancer. However, the module *E2F1* seems to be involved in a specific way. Its genes have roles in key pathways involved in colon cancer such as *Wnt Signaling Pathway*, *Apoptosis*, *Tumor suppressor activity of SMARCB1*, *G1 to S cell cycle control*, *TGF-beta Signaling Pathway*, *H19 action Rb-E2F1 signaling, and CDK-Beta-catenin activity*, *Regulation of Wnt/B-catenin Signaling by Small Molecule Compounds* and *lncRNA involvement in canonical Wnt*

Method/ Framework	Min.	1st quart.	Med.	3rd quart.	Mean	Max.
Gene4x	0.2580	0.2583	0.2586	0.2589	0.3183	0.3750
mPageRank	0.0	0.0016	0.0032	0.0048	0.1865	0.3838
Our Approach	0.2692	0.2697	0.2702	0.2706	0.4121	0.6143

Table 6. The overlap of communities resulted from multi-network methods with DisGeNet.

signaling and colorectal cancer. Moreover, Gene Ontology (GO) that is associated with *E2F1* modules includes biological processes such as *mitotic cell cycle, cell division*, *chromosome organization*, *G1/S transition of the mitotic cell cycle*, *DNA metabolic process*, and *DNA repair* which are linked to cell proliferation (Supplementary File S3).

Method evaluation. In order to evaluate the framework of this study against other community detection methods, we defined a two-step comparative analysis.

First, to assess our results in contrast with other multilayer methods, two well-known frameworks, Gene4x[46] and mPageRank[19], were selected and measures of central tendency were used to examine the distribution of each method overlap with disease-related genes (Table 6). We applied the mPageRank on the two layers of co-expression and PPI, by choosing random seeds from COAD-specific genes, as defined in the original article. However, the Gene4x is not a seed-centric approach, and from its multiple output modules, we selected top-ranked ones with considerable size for the valuation task. Here, due to the unavailability of differential expression data for the Gene4x method, evaluation has continued with the finding of the similarity of modules to gene sets involved in colorectal cancer. The results of this evaluation reveal the accuracy of our method in finding relevant communities.

Second, to test the performance of the prepared framework with other single-layer methods, another validation on detected modules was accomplished using DEG-based criteria defined by Cantini *et al.*[46]. We selected two single-layer community detection algorithms, Loavin[47] and Label-propagation[48], which were applicable to large networks. In comparing the proposed multi-layer framework with the module identification methods in single-layer networks, our aim is to investigate whether adding more layers to the proposed method has led to the identification of better modules. The two methods were executed on the co-expression layer and their output modules having an acceptable size were selected to be assessed. Here, we examined the fold change and p-value of differentially expressed genes inside modules. Data distribution for the logarithm of fold change values, student's t-test p-value and standard deviation of fold change values for the modules are depicted in Fig. 6. The comparison results denote that the approach presented here outperforms the other two methods in terms of unfolding homogenous modules containing genes with a higher change in their expression with low p-values.

Discussion

In the previous sections, we investigated the relevance of the modules with COAD in terms of functional linkage networks and pathway analyses, and results suggest that the two modules were highly correlated with the disease causality. Therefore, we discussed the candidate biomarkers of the two modules extracted from the survival analysis: *CDC6, RRM2, ORC1, NCAPG, MAD2L1, MCM6, CCNF, CDCA2, ECT2*, and *DEPDC1B*.

ORC1 is a protein-coding gene, which is overexpressed in colon cancer. *ORC1* is involved in some critical pathways: *Cell Cycle, DNA Replication, E2F transcription factor network, G1 to S cell cycle control*, and *Retinoblastoma (RB) in* cancer. Gene ontology annotations that are related to this gene include *chromatin binding*. *MCM2* and *MCM6* are the upstream transcription factors regulating the *ORC1* protein factor and are upregulated in the case of cancer; therefore, it could provide some explanations for the upregulation of *ORC1* (Fig. 7). *ORC1* is an important paralog of *CDC6*. The *CDC6* protein is essential for the initiation of DNA replication, which is a crucial phase during the cell division. This protein acts as a regulator at the early steps of DNA replication. It could be localized within the cell nucleus during cell cycle G1 but translocated to the cytoplasm at the initiation of S phase. Among its related pathways, namely *Cell cycle_Role of APC in cell cycle regulation* and *CDK-mediated phosphorylation and removal of CDC6* are more characterized compared with others. Gene ontology annotations that are associated with this gene comprise nucleotide and kinase binding. Among its regulators, *MCM3, E2F2, MCM2, E2F1, E2F7, FOXM1, ARID3A* and *MCM7* are all overexpressed in cancer. Correspondingly, *MYC* and *AR* are the other regulators, down-regulated when the cells are transformed into cancer cells.

CCNF (cyclin F) is another protein-coding gene, which encodes a member of the cyclin family. Cyclins are regulators of cell cycle transitions through their capability to bind and activate cyclin-dependent protein kinases. It is introduced as a key gene in carcinogenic pathways associated with colorectal adenoma-to-cancer progression[49]. In the same way, *MAD2L1 (Mitotic Arrest Deficient 2 like 1)* is a component of the mitotic spindle assembly checkpoint that prevents the onset of anaphase until all chromosomes are suitably aligned at the metaphase plate. Among its associated pathways, *Mitotic Metaphase and Anaphase* and *Cell cycle_Role of APC in cell cycle regulation* are well-defined and investigated. In a study conducted by Abal *et al.*[50], they mentioned that *APC* inactivation is associated with abnormal mitosis and concomitant *BUB1B*/*MAD2L1* up-regulation. However, the overexpression of these two genes was correlated with tumor metastasis and poor prognosis of patients with breast cancer[51]. Its expression is amplified in colon cancer as well.

The *CDCA2* gene encodes a subunit of *protein phosphatase 1*, which is associated with cell cycle and has relatively higher expression in cancer conditions. *POU2F2* is a regulator of *CDCA2*, which is down-regulated in colon cancer, and it is a negative regulator of *CDCA2*. Conversely, *POU5F1* is another controller of *CDCA2, which* is overexpressed in cancer and a positive regulator.

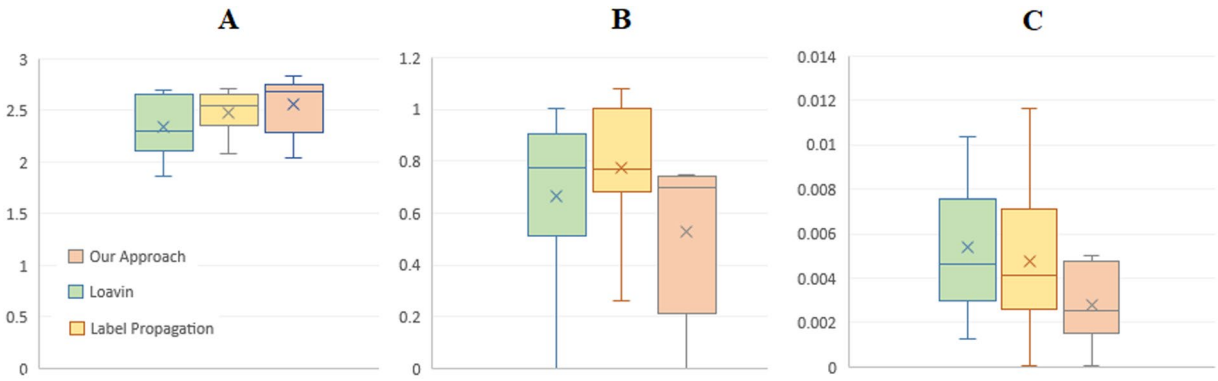

Figure 6. Comparison of single-layer methods, Loavin and Label-propagation, with our approach using DEG-based criteria defined by Cantini *et al.*[46]. (**A**) $|\text{mean}_{i\in C}(\log 2\ (\text{fold change})_i)|$; (**B**) $\text{sd}_{i\in C}(\log 2\ (\text{fold change})_i)$; (**C**) Student's t-test p-value. The framework employed here unfolds homogenous communities (low standard deviation of expression change) containing genes with higher changes in their expression and less p-value.

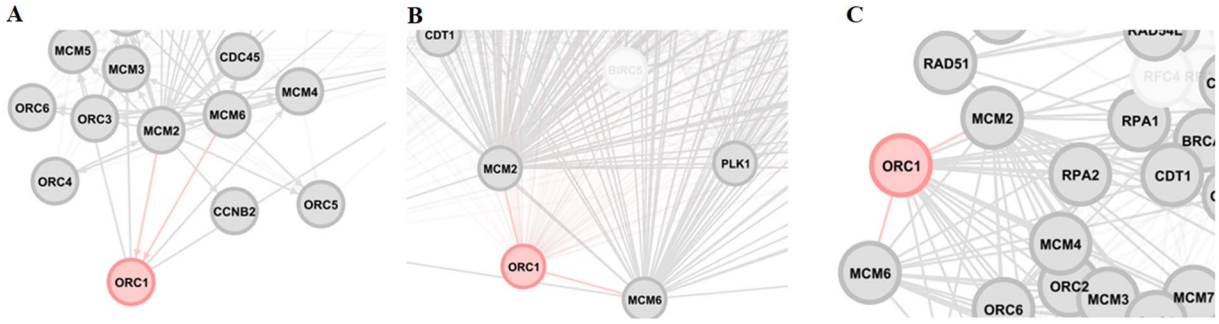

Figure 7. Three-layer overview of genes, namely *ORC1*, *MCM2*, and *MCM6*. (**A**) Regulatory layer. *MCM2* and *MCM6* (from *MCM* complex family) both regulate the transcription of the gene *ORC1*. (**B**) Co-expression layer. All three genes have expression correlations (negative and positive). (**C**) Physical binding layer.

Among the proposed gene list, *ECT2 (Epithelial Cell Transforming 2)* is another biomarker previously mentioned that it is involved in breast cancer[52]. Its expression is increased in tumor tissues of patients and involved in the following pathways *RET signaling* and *G-protein signaling_RhoA regulation*. Annotations that are associated with this gene include *protein homodimerization activity* and *GTPase activator activity*. Its role in colorectal cancer is discussed earlier by Chen *et al.*[53] and Luo *et al.*[54].

Condensin complex is responsible for the condensation and stabilization of chromosomes during mitosis and meiosis. *NCAPG (Non-SMC Condensin I Complex Subunit G)* encodes a subunit of *Condensin* complex, and phosphorylation of the encoded protein activates the complex. *Mitotic Prometaphase* and *Cell cycle Chromosome condensation in prometaphase* are biological pathways in which *NCAPG* plays a role in them. *RRM2 (Ribonucleotide Reductase Regulatory Subunit M2)* is another DEG that it has a crucial impact on the survival rate of patients with cancer. It has been indicated that *RRM2* is involved in the pathogenesis of pancreas adenocarcinoma. However, its role in colorectal cancer has been addressed in several studies[55,56]. *KRAS*-mediated upregulation of *RRM2* is vital for the proliferation of colorectal cancer cell lines[57].

The protein encoded by *MCM6* is one of the extremely conserved mini-chromosome maintenance proteins (*MCM*) that are necessary for the beginning of eukaryotic genome replication. The MCM complex consisted of this protein, as well as *MCM2*, *MCM*4, and *MCM7*. This complex has been shown to have DNA helicase activity and act as a DNA unwinding enzyme. The phosphorylation of the complex by *CDC2* kinase decreases the helicase activity, signifying a role in the regulation of DNA replication. Huang *et al.* mentioned that the interaction between *RAD51* and *MCM* complex is indispensable for the formation of *RAD51* foci in colon cancer *HCT116* cells[58]. Also, its significance has been shown in other types of cancers[59,60]. Accordingly, the protein *DEPDC1B* has some significant roles in pathways *p75 NTR receptor-mediated signaling* and *Signaling by Rho GTP*ases. An important paralog of this gene is *DEPDC1*, and it is overexpressed in the case of tumor progression. Our computations suggest that the genes mentioned above have a critical role in colon carcinoma; however, additional experimental validations are needed to approve these markers.

Conclusion

In this research, we employed multilayer networks to model colon adenocarcinoma and investigated its functional core components. We used the transcriptome-to-interactome integrative approach, and different omics data sources to reconstruct disease network and study its foundations. Then, we utilized an extended community detection algorithm and achieved two modules with centroid genes *USF1* and *E2F1*, which are dense

subnetworks, and their genes have functional linkage similarities. Both modules showed a significant overlay with COAD-related genes and pathways; however, the module *E2F1* contained statistically meaningful differentially expressed genes, and it had a marked effect on the survival of patients with COAD. We selected suitable DEGs as potential biomarkers and examined their regulatory cascade flow. Results of the literature mining suggest that the candidate genes play roles in critical pathways associated with cell cycle, apoptosis, and COAD progression; however, their role in the development and pathogenesis of cancer should be approved by experimental approaches in the future. The approach employed here is a general framework applicable to other problems in this context; however, the construction of the multilayer network is the core part of the procedure that must be constructed based on the phenotype-specific transcriptomic dataset. Modules extracted in this study are dedicated to colon adenocarcinoma, which should be confirmed experimentally.

Received: 6 August 2019; Accepted: 31 January 2020;
Published online: 19 March 2020

References

1. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology* **19**, A68 (2015).
2. Motieghader, H., Najafi, A., Sadeghi, B. & Masoudi-Nejad, A. A hybrid gene selection algorithm for microarray cancer classification using genetic algorithm and learning automata. *Informatics in Medicine Unlocked* **9**, 246–254 (2017).
3. Masoudi-Sobhanzadeh, Y., Omidi, Y., Amanlou, M. & Masoudi-Nejad, A. DrugR+: A comprehensive relational database for drug repurposing, combination therapy, and replacement therapy. *Computers in biology and medicine* **109**, 254–262 (2019).
4. Yugi, K., Kubota, H., Hatano, A. & Kuroda, S. Trans-omics: how to reconstruct biochemical networks across multiple 'omic'layers. *Trends in biotechnology* **34**, 276–290 (2016).
5. Yan, J., Risacher, S. L., Shen, L. & Saykin, A. J. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Briefings in bioinformatics* **19**, 1370–1381 (2017).
6. Bonnet, E., Calzone, L. & Michoel, T. Integrative multi-omics module network inference with Lemon-Tree. *PLoS computational biology* **11**, e1003983 (2015).
7. Kuo, T.-C., Tian, T.-F. & Tseng, Y. J. 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. *BMC systems biology* **7**, 64 (2013).
8. Mitra, K., Carvunis, A.-R., Ramesh, S. K. & Ideker, T. Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics* **14**, 719–732 (2013).
9. Peng, C., Li, A. & Wang, M. Discovery of bladder Cancer-related genes using integrative heterogeneous network modeling of multi-omics data. *Scientific reports* **7**, 15639 (2017).
10. Kivelä, M. *et al*. Multilayer networks. *Journal of complex networks* **2**, 203–271 (2014).
11. Boccaletti, S. *et al*. The structure and dynamics of multilayer networks. *Physics Reports* **544**, 1–122 (2014).
12. Hmimida, M. & Kanawati, R. Community detection in multiplex networks: A seed-centric approach. *NHM* **10**, 71–85 (2015).
13. Didier, G., Brun, C. & Baudot, A. Identifying communities from multiplex biological networks. *PeerJ* **3**, e1525 (2015).
14. Gomez, S. *et al*. Diffusion dynamics on multiplex networks. *Physical review letters* **110**, 028701 (2013).
15. Sánchez-García, R. J., Cozzo, E. & Moreno, Y. Dimensionality reduction and spectral properties of multilayer networks. *Physical Review E* **89**, 052815 (2014).
16. De Domenico, M., Nicosia, V., Arenas, A. & Latora, V. Structural reducibility of multilayer networks. *Nature communications* **6**, 6864 (2015).
17. Bennett, L., Kittas, A., Muirhead, G., Papageorgiou, L. G. & Tsoka, S. Detection of composite communities in multiplex biological networks. *Scientific reports* **5** (2015).
18. Mucha, P. J. & Porter, M. A. Communities in multislice voting networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **20**, 041108 (2010).
19. Li, J. & Zhao, P. X. Mining functional modules in heterogeneous biological networks using multiplex PageRank approach. *Frontiers in plant science* **7** (2016).
20. Rocklin, M. & Pinar, A. In *International Workshop on Algorithms and Models for the Web-Graph*. 38–49 (Springer).
21. Jeub, L. G., Mahoney, M. W., Mucha, P. J. & Porter, M. A. A local perspective on community structure in multilayer networks. *Network Science* **5**, 144–163 (2017).
22. Interdonato, R., Tagarelli, A., Ienco, D., Sallaberry, A. & Poncelet, P. Local community detection in multilayer networks. *Data Mining and Knowledge Discovery* **31**, 1444–1479 (2017).
23. Berenstein, A. J., Magariños, M. P., Chernomoretz, A. & Agüero, F. A multilayer network approach for guiding drug repositioning in neglected diseases. *PLoS neglected tropical diseases* **10**, e0004300 (2016).
24. Rai, A. *et al*. Understanding cancer complexome using networks, spectral graph theory and multilayer framework. *Scientific reports* **7**, 41676 (2017).
25. Motieghader, H., Kouhsar, M., Najafi, A., Sadeghi, B. & Masoudi-Nejad, A. mRNA–miRNA bipartite network reconstruction to predict prognostic module biomarkers in colorectal cancer stage differentiation. *Molecular BioSystems* **13**, 2168–2180 (2017).
26. Goldman, M. *et al*. The UCSC Xena Platform for cancer genomics data visualization and interpretation. *BioRxiv*, 326470 (2018).
27. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* **9**, 559 (2008).
28. Wu, T. *et al*. NPInter: the noncoding RNAs and protein related biomacromolecules interaction database. *Nucleic acids research* **34**, D150–D152 (2006).
29. Liu, Z.-P., Wu, C., Miao, H. & Wu, H. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database* **2015** (2015).
30. Han, H. *et al*. TRRUST: a reference database of human transcriptional regulatory interactions. *Scientific reports* **5**, 11432 (2015).
31. Menche, J. *et al*. Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601 (2015).
32. Piñero, J. *et al*. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research*, gkw943 (2016).
33. Chen, E. Y. *et al*. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics* **14**, 128 (2013).
34. Kuleshov, M. V. *et al*. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research* **44**, W90–W97 (2016).
35. Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research* **37**, W305-W311 (2009).
36. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27–30 (2000).
37. Croft, D. *et al*. The Reactome pathway knowledgebase. *Nucleic acids research* **42**, D472–D477 (2013).
38. Smyth, G. K. In *Bioinformatics and computational biology solutions using R and Bioconductor*. 397–420 (Springer, 2005).

39. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
40. Law, C. W., Alhamdoosh, M., Su, S., Smyth, G. K. & Ritchie, M. E. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *F1000Research* **5** (2016).
41. Bland, J. M. & Altman, D. G. Survival probabilities (the Kaplan-Meier method). *Bmj* **317**, 1572–1580 (1998).
42. Aguirre-Gamboa, R. *et al*. SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis. *PloS one* **8**, e74250 (2013).
43. Murali, T., Wu, C.-J. & Kasif, S. The art of gene function prediction. *Nature biotechnology* **24**, 1474 (2006).
44. Mousavian, Z., Khakabimamaghani, S., Kavousi, K. & Masoudi-Nejad, A. Drug–target interaction prediction from PSSM based evolutionary information. *Journal of pharmacological and toxicological methods* **78**, 42–51 (2016).
45. Schwikowski, B., Uetz, P. & Fields, S. A network of protein–protein interactions in yeast. *Nature biotechnology* **18**, 1257 (2000).
46. Cantini, L., Medico, E., Fortunato, S. & Caselle, M. Detection of gene communities in multi-networks reveals cancer drivers. *Scientific reports* **5**, srep17386 (2015).
47. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* **2008**, P10008 (2008).
48. Cordasco, G. & Gargano, L. In 2010 IEEE International Workshop on: Business Applications of Social Network Analysis (BASNA). 1–8 (IEEE).
49. Sillars-Hardebol, A. H. *et al*. Identification of key genes for carcinogenic pathways associated with colorectal adenoma-to-carcinoma progression. *Tumor Biology* **31**, 89–96 (2010).
50. Abal, M. *et al*. APC inactivation associates with abnormal mitosis completion and concomitant BUB1B/MAD2L1 up-regulation. *Gastroenterology* **132**, 2448–2458 (2007).
51. Wang, Z. *et al*. Biological and clinical significance of MAD2L1 and BUB1, genes frequently appearing in expression signatures for breast cancer prognosis. *PloS one* **10**, e0136246 (2015).
52. Wang, H.-k, Liang, J.-f, Zheng, H.-x & Xiao, H. Expression and prognostic significance of ECT2 in invasive breast cancer. *Journal of clinical pathology* **71**, 442–445 (2018).
53. Chen, C.-J. *et al*. Early assessment of colorectal cancer by quantifying circulating tumor cells in peripheral blood: ECT2 in diagnosis of colorectal cancer. *International journal of molecular sciences* **18**, 743 (2017).
54. Luo, Y. *et al*. Elevated expression of ECT2 predicts unfavorable prognosis in patients with colorectal cancer. *Biomedicine & Pharmacotherapy* **73**, 135–139 (2015).
55. Lu, A.-G., Feng, H., Pu-Xiong-Zhi Wang, D.-P., Han, X.-H. C. & Zheng, M.-H. Emerging roles of the ribonucleotide reductase M2 in colorectal cancer and ultraviolet-induced DNA damage repair. *World Journal of Gastroenterology: WJG* **18**, 4704 (2012).
56. Liu, X. *et al*. Ribonucleotide reductase small subunit M2 serves as a prognostic biomarker and predicts poor survival of colorectal cancers. *Clinical science* **124**, 567–579 (2013).
57. Yoshida, Y. *et al*. KRAS-mediated up-regulation of RRM2 expression is essential for the proliferation of colorectal cancer cell lines. *Anticancer research* **31**, 2535–2539 (2011).
58. Huang, J. *et al*. Interaction between RAD51 and MCM complex is essential for RAD51 foci forming in colon cancer HCT116 cells. *Biochemistry (Moscow)* **83**, 69–75 (2018).
59. Liu, Y.-Z. *et al*. MCMs expression in lung cancer: implication of prognostic significance. *Journal of Cancer* **8**, 3641 (2017).
60. Kwok, H. F. *et al*. Prognostic significance of minichromosome maintenance proteins in breast cancer. *American journal of cancer research* **5**, 52 (2015).

Author contributions

Ehsan Pournoor: conceptualization, implementation, data analysis, investigation, writing, editing, and revising the manuscript. Zaynab Mousavian: results analysis, validation, interpretation, writing, editing, and revising the manuscript. Abbas Nowzari Dalini: supervision, conceptualization, interpretation, editing and revising the manuscript. Ali Masoudi-Nejad: supervision, conceptualization, project administration, editing and revising the manuscript. All authors have read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-59605-z.

Correspondence and requests for materials should be addressed to A.M.-N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.