Mathematical Biosciences and Engineering

MBE, 18(6): 8997–9015.
DOI: 10.3934/mbe.2021443
Received: 11 August 2021
Accepted: 25 September 2021
Published: 19 October 2021

http://www.aimspress.com/journal/MBE

Research article

Identification of potential biomarkers with colorectal cancer based on bioinformatics analysis and machine learning

Ahmed Hammad[1,2], Mohamed Elshaer[1,3] and Xiuwen Tang[1,*]

[1] Department of Biochemistry and Department of Thoracic Surgery of the First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310003, China
[2] Radiation Biology Department, National Center for Radiation Research and Technology, Egyptian Atomic Energy Authority, Cairo 13759, Egypt
[3] Labeled Compounds Department, Hot Labs Center, Egyptian Atomic Energy Authority, Cairo 13759, Egypt

* **Correspondence:** E-mail: xiuwentang@zju.edu.cn; Tel: +0086(0)57188981270; Fax: +0086(0)57188208266

Abstract: Colorectal cancer (CRC) is one of the most common malignancies worldwide. Biomarker discovery is critical to improve CRC diagnosis, however, machine learning offers a new platform to study the etiology of CRC for this purpose. Therefore, the current study aimed to perform an integrated bioinformatics and machine learning analyses to explore novel biomarkers for CRC prognosis. In this study, we acquired gene expression microarray data from Gene Expression Omnibus (GEO) database. The microarray expressions GSE103512 dataset was downloaded and integrated. Subsequently, differentially expressed genes (DEGs) were identified and functionally analyzed via Gene Ontology (GO) and Kyoto Enrichment of Genes and Genomes (KEGG). Furthermore, protein protein interaction (PPI) network analysis was conducted using the STRING database and Cytoscape software to identify hub genes; however, the hub genes were subjected to Support Vector Machine (SVM), Receiver operating characteristic curve (ROC) and survival analyses to explore their diagnostic values. Meanwhile, TCGA transcriptomics data in Gene Expression Profiling Interactive Analysis (GEPIA) database and the pathology data presented by in the human protein atlas (HPA) database were used to verify our transcriptomic analyses. A total of 105 DEGs were identified in this study. Functional enrichment analysis showed that these genes were significantly enriched in biological processes related to cancer progression. Thereafter, PPI network explored a total of 10 significant hub genes. The ROC curve was used to predict the potential application of biomarkers in CRC diagnosis, with an area under ROC curve (AUC) of these genes exceeding 0.92 suggesting that this risk classifier can discriminate between CRC patients and normal controls. Moreover, the prognostic values of these hub genes were confirmed by survival analyses using different CRC patient cohorts. Our results demonstrated that these 10 differentially expressed hub genes could be used as potential biomarkers for CRC diagnosis.

Keywords: hub genes; PPI; gene microarray; colorectal cancer; AUC; biomarkers

Abbreviations: CRC: Colorectal cancer; GEO: Gene expression omnibus; DEGs: Differentially expressed genes; GO: Gene ontology; KEGG: Kyoto enrichment of genes and genomes; PPI: Protein protein interaction; SVM: Support vector machine; ROC: Receiver operating characteristic curve; GEPIA: Gene expression profiling interactive analysis; AUC: Area under ROC curve; DAVID: The NIH database for annotation, visualization and integrated discovery; HPA: The human protein atlas; IHC: Immunohistochemistry; MYH11: Myosin-11; IGF1: Insulin-like growth factor 1; CLU: Clusterin; FOS: FBJ murine osteosarcoma viral oncogene homolog; MYL9: Myosin regulatory light polypeptide 9; CXCL12: Chemokine (CXC motif) ligand12; LMOD1: Leiomodin 1; CNN1: Calponin 1; HIST1H2BO: Histone cluster 1 H2B family member O; C3: Complement component 3

1. Introduction

Colorectal cancer is one of the most common malignancies worldwide [1], and the third common cancer worldwide, with more than 1.2 million new cases diagnosed annually [2]. In 2015 in China, there were about 376,000 of new CRC patients and 191,000 one of CRC death, accounting for the fifth of malignant tumor incidence and mortality [3]. While there has been significant progress in improving the diagnosis of CRC, the disease is often diagnosed at an advanced stage. Furthermore, several biomarkers including KRAS and BRAF that can be used to detect CRC, but these biomarkers are not sufficiently sensitive or specific [4]. Consequently, there is an urgent need to explore efficacious biomarkers for an early diagnosis of CRC.

Transcriptome analysis of high-throughput sequencing such as microarrays and RNA sequencing has been considered as a promising tool in cancer research to identify pathways and genes for candidate prognostic and diagnostic biomarkers [5–8]. Moreover, these biomarkers may bring a breakthrough in improving the prevention and treatment of CRC [9–13]. Recently, bioinformatic analysis of gene expression data explored potential gene biomarkers for CRC [8,9,14–16], however sometimes the results of bioinformatics are inconsistent in behavior [17,18]. In this context, the integration of machine learning techniques with bioinformatic methods can provide consistent results and enhance training and validation of CRC biomarkers [17–20]. Moreover, large lists of DEGs have been identified in CRC from microarray datasets; however, the involvement of the DEGs in the molecular mechanisms and signaling networks related to CRC progression are not fully understood [21]. More recently, machine learning tools such as ROC and SVM are recently used to evaluate the diagnostic efficacy of newly discovered biomarkers in different types of diseases, including cancers [22].

Support vector machine (SVM), a kernel algorithm, is widely applied in bioinformatics due to its high accuracy, and has the ability to identify the multivariate statistical properties of data that discriminate between two different groups [23]. SVM can draw an optimal hyper-plane in a high dimensional feature space that defines a boundary that maximizes the margin between data samples in two classes. Recently, several biomarkers were predicted using SVM model that was able to distinguish normal samples from those of CRC patients [24–26]. We have been used these tools in our study to explore the diagnostic value of biomarkers for CRC.

Recent studies have highlighted the prognostic value of different DEGs in CRC [27,28]. However, these studies have produced varied results, potentially due to the different analysis methods used. In addition, evaluation of the prognostic value of the DEGs using machine learning tools in CRC is still lacking. Furthermore, the enrichment pathways, Gene Ontology (GO) functions and the interaction network of the DEGs remain to be clarified. Therefore, in this study, we used bioinformatics and

machine learning methods to analyze key genes for CRC based on publicly available databases and verified the diagnostic values of the candidate genes.

2. Material and methods

2.1. The GEO dataset and data processing

The gene expression microarray GSE103512 dataset was downloaded from GEO and analyzed as described in Figure 1. The GSE103512 series (GPL13158 platform, [HT_HG-U133_Plus_PM] Affymetrix HT HG-U133+ PM Array Plate) included a total of 69 samples (57 colorectal cancers with 12 normal samples). Probe symbols were converted into gene symbols using the R statistical software package (www.r-project.org). When multiple probes corresponded to a single gene, mean expression was used. Samples were extracted from tumor (CRC adenocarcinomas) or adjacent normal tissue, and then formalin-fixed and paraffin embedded. Total RNA was extracted using High Pure FFPET RNA Isolation Kits. Hybridization was performed using GeneTitan Hybridization, Wash and Stain Kit. During data processing, the *CEL* files were normalized and summarized into probe-set values using RMA normalization.

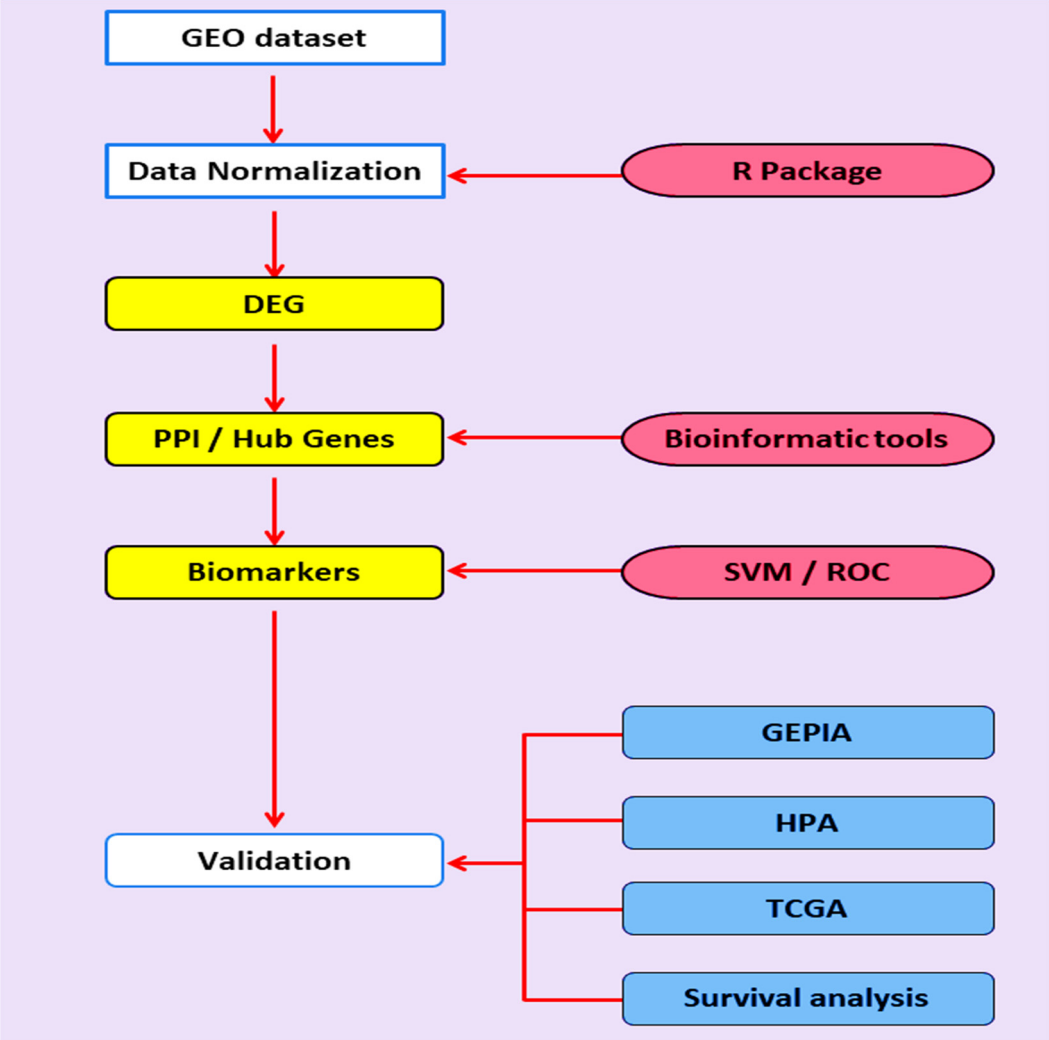


Figure 1. Flow chart for bioinformatics analysis of colorectal cancer (CRC) samples.

2.2. DEG identification

Analyses of the dataset (GSE103512) was performed via the R statistical software package [29]. The cutoff criteria were |log2 fold change (log2 FC) | > 1.5 and P-value < 0.05. A fold change (FC) of gene describes the ratio between the gene expression values for cancer and normal.

2.3. DEG enrichment analyses

GO terms (http://www.geneontology.org/) for gene sets were collected using the Database for Annotation, Visualization and Integrated Discovery (DAVID) web tool [30]. Kyoto Enrichment of Genes and Genomes (KEGG) analysis was performed using the KOBAS (Version 3.0) web tool (http://kobas.cbi.pku.edu.cn/) [31]. These tools provide a functional interpretation for large gene lists derived from genomic studies. A Benjamini P-value < 0.05 was used in the analysis.

2.4. Protein-protein interaction networks

To further investigate the molecular mechanism of DEGs in the CRC and the interactive associations between DEGs, the genes were used to construct a PPI network with the biological online database tool Search Tool for the Retrieval of Interacting Genes, (STRING, http://stringdb.org) [32]**.** A combined score > 0.4 (high confidence score) was considered significant, and then the PPI network was visualized via Cytoscape software (Version 3.5.1) [33]. The hub genes/proteins, a small number of crucial nodes for the protein interactions in the PPI network, were chosen with a centrality degree > 5. Degree centrality quantifies the number of neighbours to which a node directly connects

2.5. Clustering analyses

The expression profiles of DEGs and hub genes in all cases (cancer and normal) were determined using heatmaper [34].

2.6. Evaluation of diagnostic biomarkers

SVM is a supervised learning algorithm capable of solving complex classification problems. SVM is based on the structural risk minimization principle from statistical learning theory [35]. A set of positive (CRC) and negative (normal) examples can be represented by the feature vectors x i (i = 1, 2, ... N) with corresponding labels y i $\in$ {+1,-1}. To classify the data, the SVM trains a classifier by mapping the input samples onto a high-dimensional space using a kernel function, followed by seeking a separating hyperplane that differentiates the two different classes with maximal margin and minimal error [36–38]. In this study, we applied the re-sampling technique to solve the class imbalance problem and to enhance classification performance of our model. Generally, re-sampling of the data can be performed through adding data to the minority class (over-sampling) and deleting some of the data from the majority class (under-sampling), however oversampling was preferred over under-sampling to minimize information loss [39]. Therefore, we applied oversampling via the *smotefamily* package in R in this study to make a balance between the number of tumors and normal samples.

Furthermore, SVM analysis was performed via the *e1071* package in R to explore diagnostic biomarkers of CRC [40,41]. In brief, the gene expression data is split into two sets: Training set and Test set. Training set is used to train the classifier. Test set is used to estimate the performance of the developed system. The SVM classifier was subsequently established to predict cancer based on the expression levels.

The other dataset (GSE128435) was used to further verify the results of the SVM classifier and the predictive value of these biomarkers. Subsequently, ROC curve analysis was applied to evaluate the specificity and sensitivity of the SVM prediction model. The AUC values were determined to evaluate the performance of the established SVM classifier. Thereafter, we used easyROC: a web-tool for ROC curve analysis (Version 1.3.1) to identify the AUC value for each gene and separately determine the diagnostic accuracy of each marker [42]. The discrimination power of biomarker panels is assessed by ROC curve analysis which combines sensitivity and specificity of a given marker for diagnostic test evaluation which ranges from 0.5 (no discriminating power) to 1.0 (complete separation).

2.7. Validation of biomarkers gene expression

The expression levels of hub genes in CRC and normal cases were verified via using GEPIA (http://gepia.cancer-pku.cn/), a database of data retrieved from the UCSC Xena server, which includes 9736 tumor samples and 8587 normal samples. P-value < 0.05 indicated statistically significant differences [43]. The human protein atlas (HPA) database (https://www.proteinatlas.org/) was used to confirm the protein expression level of biomarkers in CRC tissues using the immunohistochemistry (IHC) staining data. Furthermore, the correlation between hub gene expression and CRC clinical stages was performed via GEPIA datasets.

2.8. Survival analysis

The survival analyses of biomarkers were performed by PROGgeneV2 Prognostic Database with different CRC cohorts [44].

2.9. Statistical analysis

Statistical analyses were performed with the statistical software R [29]. Student's t-test was used to compare two groups. A value of P-value < 0.05 was considered statistically significant.

3. Results

3.1. Identification of DEGs

The gene expression microarray provides an opportunity to simultaneously analyze large sets of genes and identify differences between molecular pathways of CRC tissue and normal tissue. We analyzed the profiles of DEGs in diseased and normal samples following the cutoff criteria fold-change $|logFC| > 1.5$ and P-value < 0.05. After processing the gene expression profile, which contained 2,740 genes, we identified a total of 105 DEGs with 17 upregulated and 88 downregulated genes (Figure 2). A full list of genes and fold changes for DEGs is reported in Supplementary Table S1

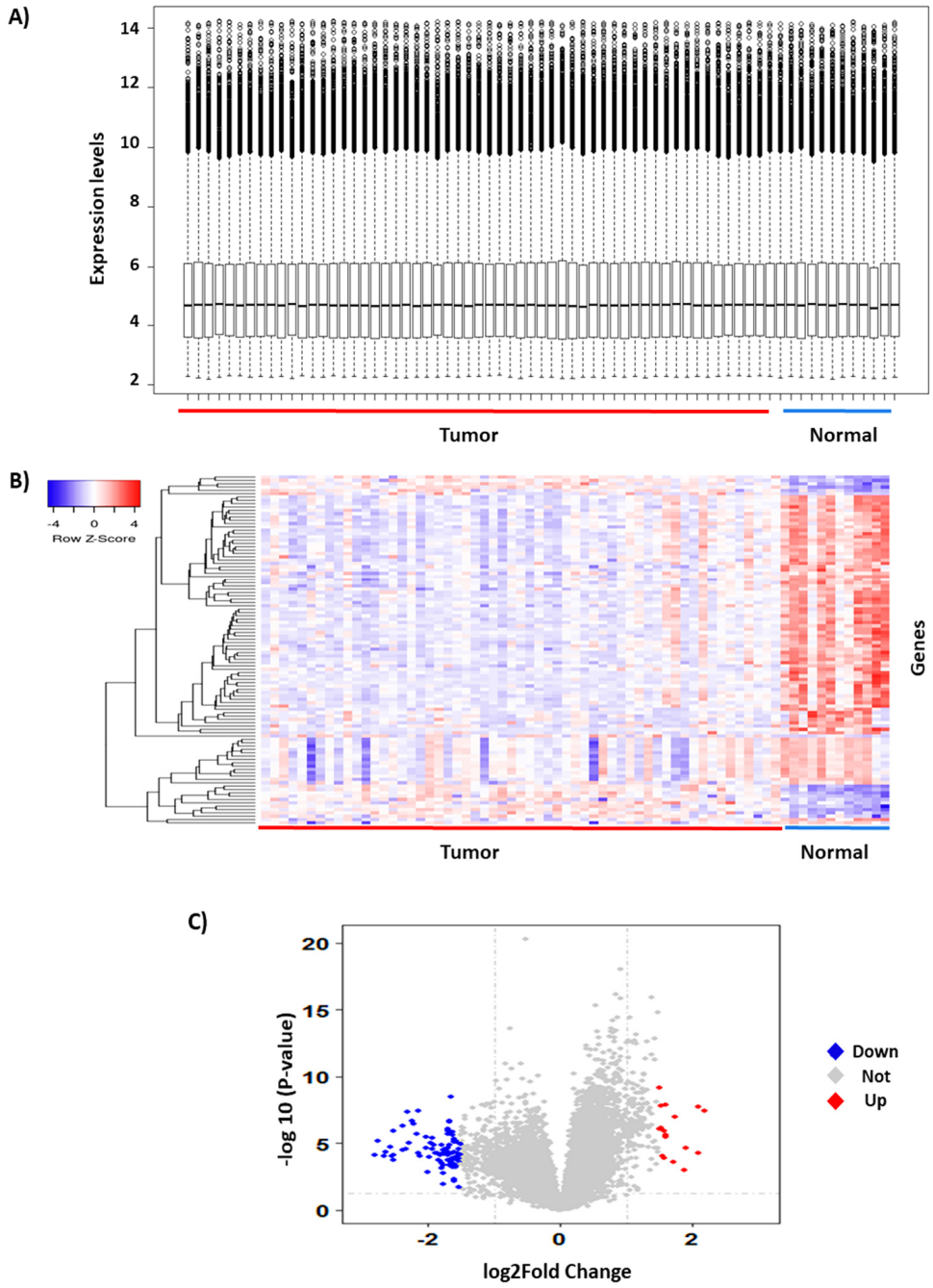


Figure 2. Identification of differentially expressed genes (DEGs). A: Box plots of the mean expression for each case in the dataset. Box plots generated from normalized microarray measurements. X-axis: individual samples grouped into tumor (red) and normal (blue); Y-axis: expression intensity values. B: Expression heatmap for DEGs. Relative gene expression values are presented from low relative expression (blue) to high relative expression (red) on a color scale (-4.0 to 4.0). Color intensity is proportional to the relative expression value of each gene. Rows contain gene expression data, and columns contain samples (normal vs. tumor). C: Volcano plot showing the magnitude of differential expression between tumor and normal samples. Each dot represents one gene that had detectable expression in both tissues. Blue dots represent downregulated genes with log2FC < -1.5 and P-value of 0.05. Red dots represent upregulated genes with log2FC > 1.5 and P-value of 0.05.

3.2. *Molecular pathways associated with DEGs*

The GO analysis data indicated that the DEGs were markedly enriched in biological processes, including complement activation, positive regulation of B cell activation, phagocytosis, B cell receptor signaling pathway, immune response, defense response to bacterium, proteolysis, and innate immune response (Figure 3A), while, the KEGG analysis exhibited that the DEGs were significantly enriched in vascular smooth muscle contraction, systemic lupus erythematosus, arachidonic acid metabolism, complement and coagulation cascades, alcoholism, transcriptional misregulation in cancer, oxytocin signaling pathway, pertussis, focal adhesion, rheumatoid arthritis, and IL-17 signaling pathways (Figure 3B).

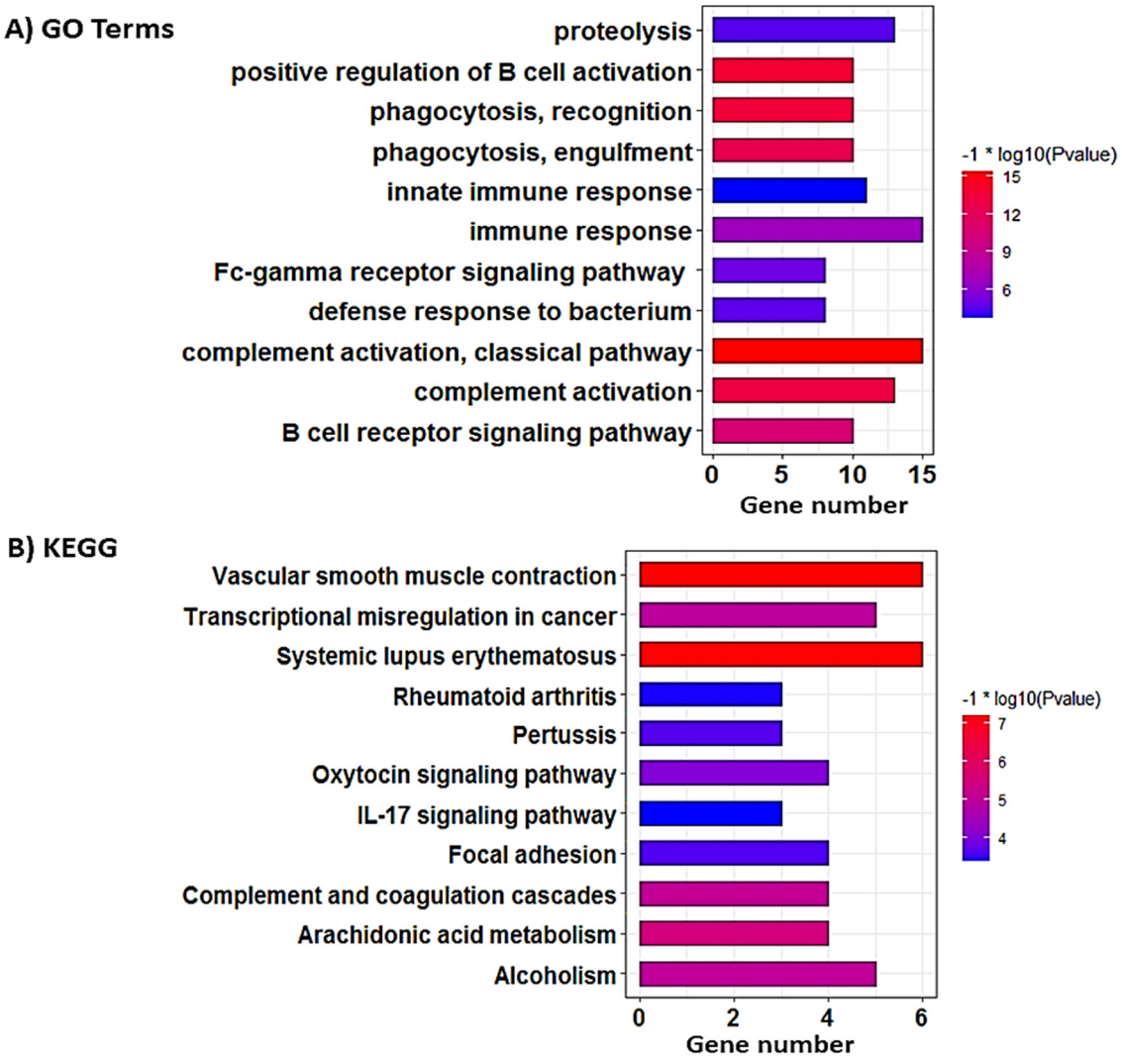


Figure 3. Functional enrichment analyses. A: GO terms for differentially expressed genes (DEGs) in this study. B: KEGG pathways for DEGs.

3.3. *Construction of the PPI network for DEGs*

The PPI network for 105 DEGs via STRING database exhibited that the PPI network was consisted of 82 nodes and 91 interactions. Subsequently, hub genes of PPI network were screened via Cytoscape software (Figure 4). A total of 10 candidate genes with high degrees of interaction

were selected, including insulin-like growth factor 1 (IGF1), myosin-11 (MYH11), clusterin (CLU), FBJ murine osteosarcoma viral oncogene homolog (FOS), low myosin regulatory light polypeptide 9 (MYL9), chemokine (CXC motif) ligand 12 (CXCL12), leiomodin 1 (LMOD1), calponin 1 (CNN1), complement component 3 (C3), and histone cluster 1 H2B family member O (HIST1H2BO) (Figure 4 and Table S2).

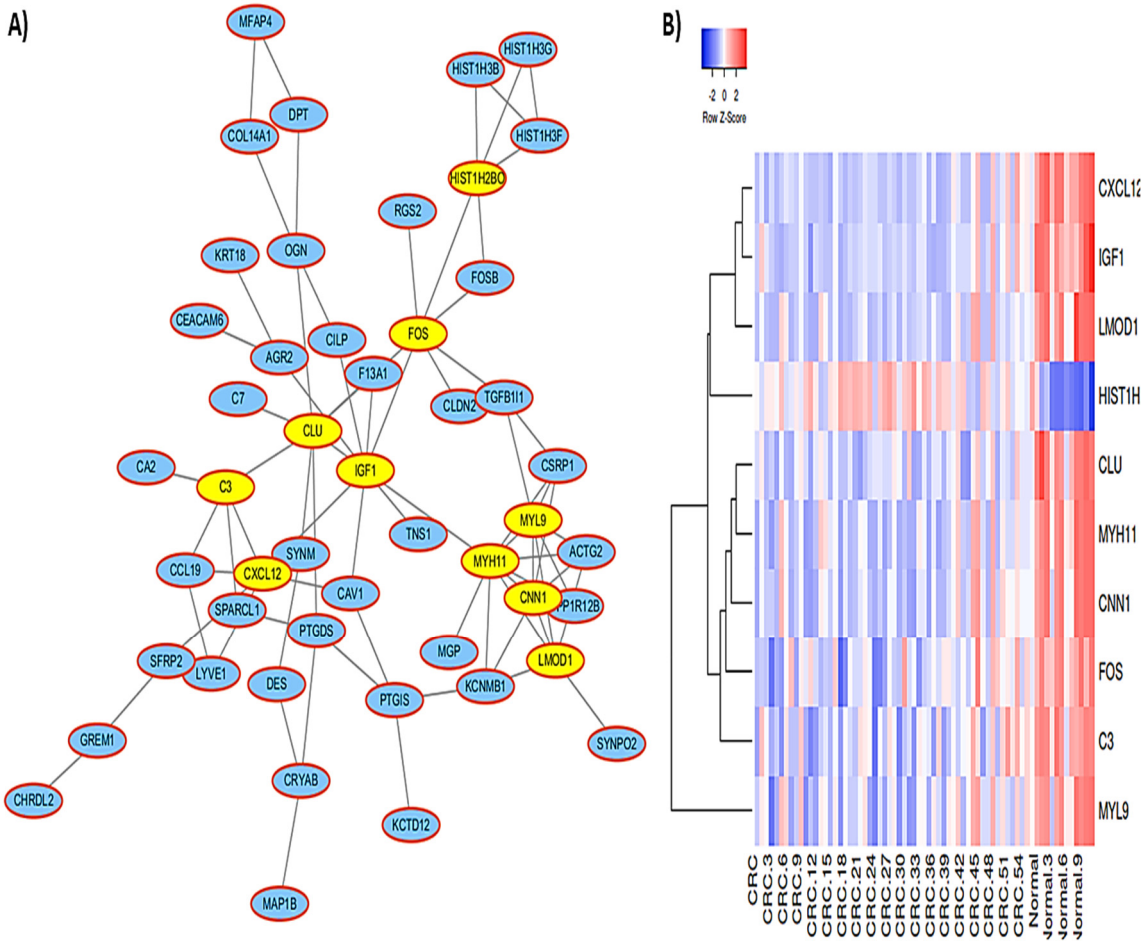


Figure 4. Identification of hub genes. A: PPI network analysis of differentially expressed genes (DEGs) in tumors from colorectal cancer (CRC) patients. The protein–protein interaction (PPI) network was constructed with STRING. This network contained 82 nodes and 91 interaction pairs. Yellow-highlighted rectangles show genes with higher connectivity (hub genes). Blue indicates genes with lower degrees of interactions. B: Expression heatmap for discovered biomarkers. Relative gene expression is reported from low relative expression (blue) to high relative expression (red) on a color scale (-2.0 to 2.0). Color intensity is proportional to the relative expression value of gene. Rows contain gene expression data, and columns contain the samples (normal vs. tumor).

3.4. Diagnostic value of biomarkers

To investigate the diagnostic value of the hub genes, we prepared a gene expression heatmap for all samples. The heatmap revealed differential expression patterns between the CRC and control samples (Figure 4). Subsequently, we developed an SVM model based on gene expression to identify the diagnostic value of these 10 hub genes in CRC. where the input of an SVM is a training set $S = (x_1,$

y_1), ..., (x_n, y_n) of vector of features $x_i \in X$ together with their known classes $y_i \in \{-1, +1\}$. The output of an SVM is a model f: $X \rightarrow \{-1, +1\}$ that predicts the class $f(x)$ of any new object $x \in X$.

Sensitivity and Specificity were used to judge the performance of the classification system. Computing of these measurements is based on true positives (TP), which mean the samples are predicted positive, true negatives (TN), which mean the samples are correctly predicted negative, false positives (FP), which mean samples are wrongly predicted positive, and false negatives (FN), which mean the samples are wrongly predicted negative. Moreover, false positive rate is considered as (1–Specificity).

1- Sensitivity (true positive rate) is calculated as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \text{ x } 100$$

2- Specificity (true negative rate) is calculated as follows:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \text{ x } 100$$

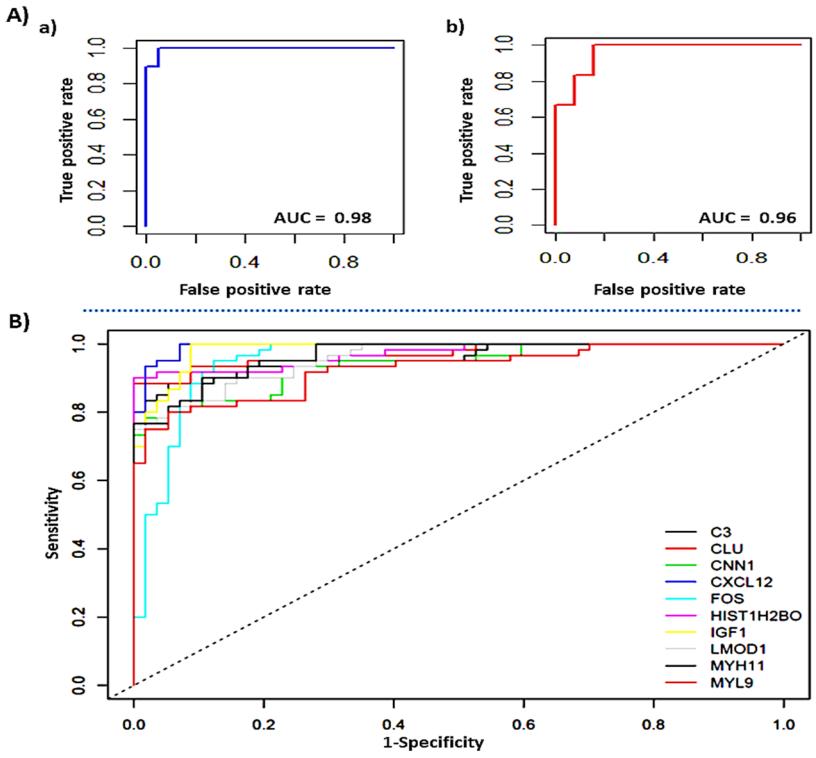
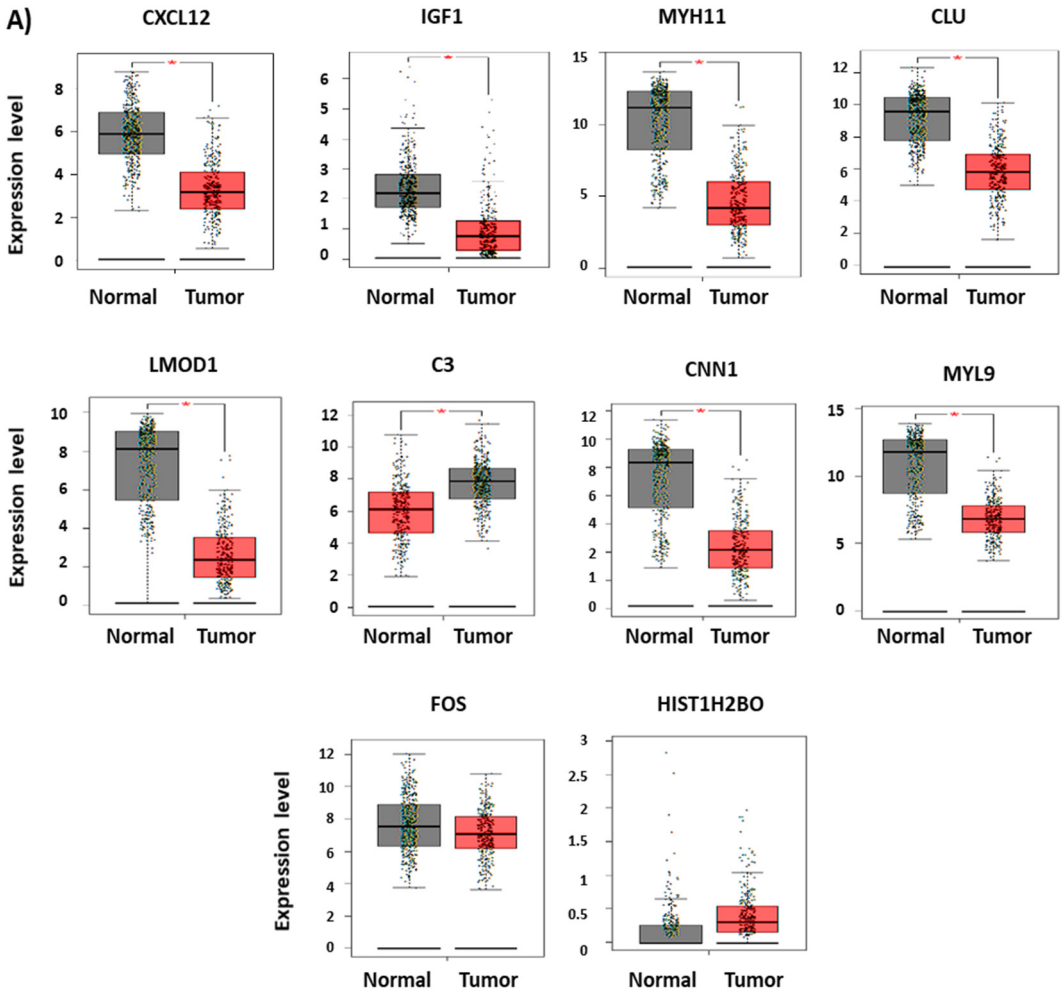


Figure 5. Diagnostic value of the newly developed biomarkers. A: Receiver operating characteristic curve (ROC) curves of support vector machine (SVM)-based hub gene risk classifier in the training set (a) for differentiating colorectal cancer (CRC) patients from the healthy controls and for validation of the results (b). B: ROC curves of 10 genes are shown for differentiating CRC patients from healthy controls. ROC curves were generated for each biomarker. The dashed reference line represents the ROC curve for a test with no discriminatory ability. Area under the ROC curve (AUC) is displayed for each marker in the Table S3.

We achieved AUC of 0.98 for the training set and AUC of 0.96 for validation set. Furthermore, ROC analysis was performed to estimate the diagnostic efficiency of these 10 candidate hub genes in CRC patients based on gene expression. The results showed that AUC values for all gene were > 0.92, suggesting that the hub gene risk classifier had good discrimination between the CRC and control samples with high sensitivity and specificity for CRC diagnosis (Figure 5 and Table S3). These results suggested that the 10 differentially expressed hub genes could be used as potential biomarkers for the diagnosis of CRC.

3.5. Biomarker validation

The GEPIA database was used to verify the mRNA expression of hub genes with P-value < 0.05 and Log2FC > 1 as the threshold. GEPIA box plots exhibited that the expressions of all hub genes except HIST1H2BO and FOS were significantly downregulated CRC patients. (Figure 6A). Furthermore, Correlation analysis explored that the mRNA expressions of MYH11, LMOD1, CNN1, and MYL9 genes were significantly correlated with CRC clinical stages (Figure 6B). Subsequently, immunohistochemical analysis of the Human Protein Atlas (HPA) database revealed that the protein expression of HIST1H2BO was significantly upregulated in CRC tissue while the protein expression of CXCL12 and CNN1 were significantly downregulated in CRC tissues (Supplementary Figure S1).



Continued on next page

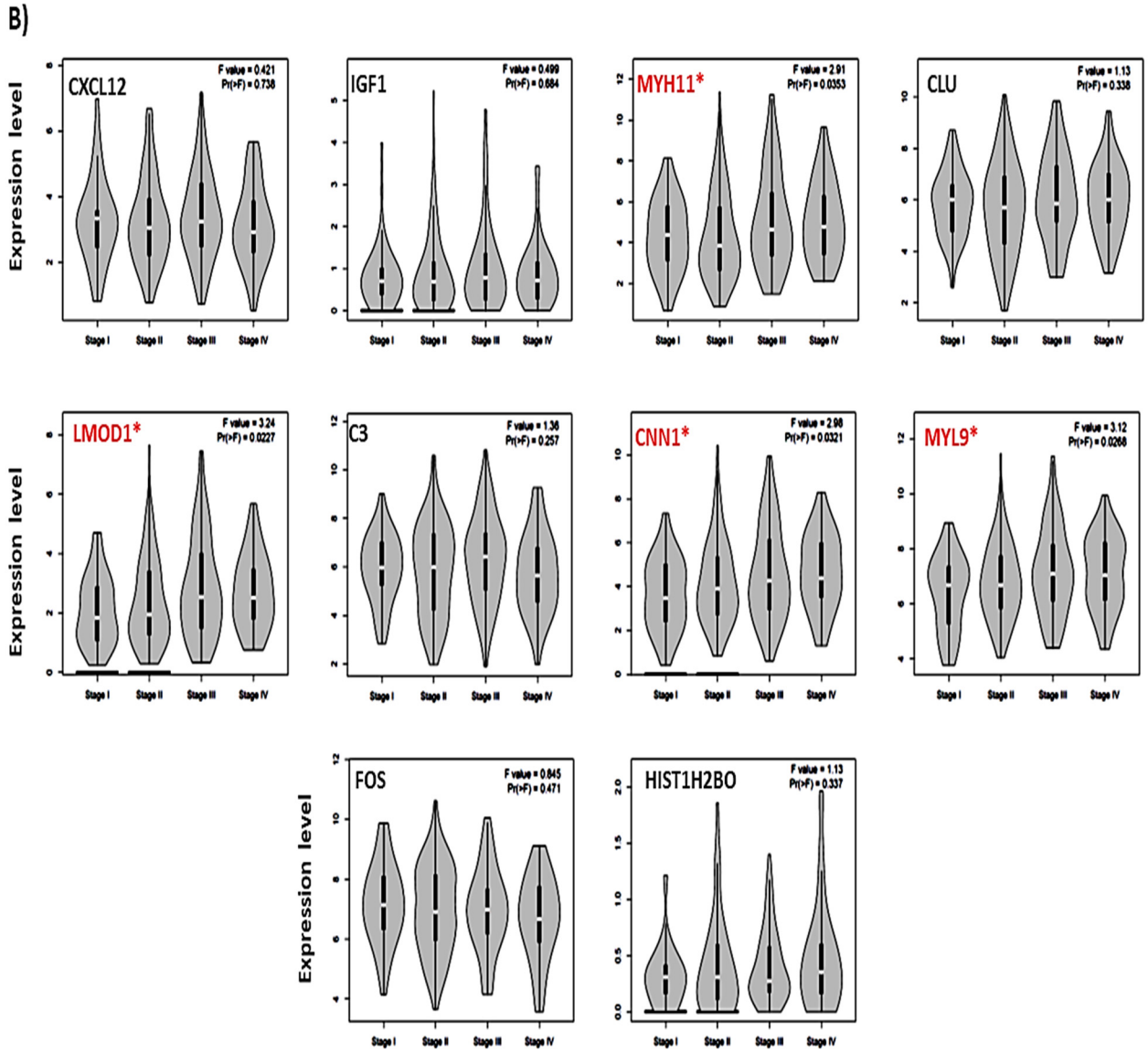


Figure 6. Validation of bioinformatic analysis. A: Boxplots are showing the expression of hub genes in cancer patients and normal controls via GEPIA datasets. B: The correlation analysis between the hub gene expressions and clinical stages via GEPIA datasets.

We further screened the DEGs profile for TCGA dataset with 275 cancer patients and 41 normal tissues, through the cutoff criteria fold-change |log2FC| > 1.5 and P-value < 0.05, to evaluate GEO dataset results. The results showed that a total of 61 genes were common between TCGA cohort and GEO cohort; however, a total of 7 genes (C3, CLU, CNN1, CXCL12, LMOD1, MYH11 and MYL9) of hub genes were persistent with DEGs in TCGA (Figure 7A). Altogether, these results were coincided with transcriptomic analysis for GEO dataset of CRC patients.

3.6. Survival analysis:

Kaplan-Meier survival analysis revealed that CRC patients are separated into two groups (high-risk group in red and low-risk group in green) according to the expression profiles of biomarker genes.

Meanwhile, these genes provided the best split between patients with high and low risks based on their expression. Therefore, these genes could be used in the prognosis of CRC (Figure 7B).

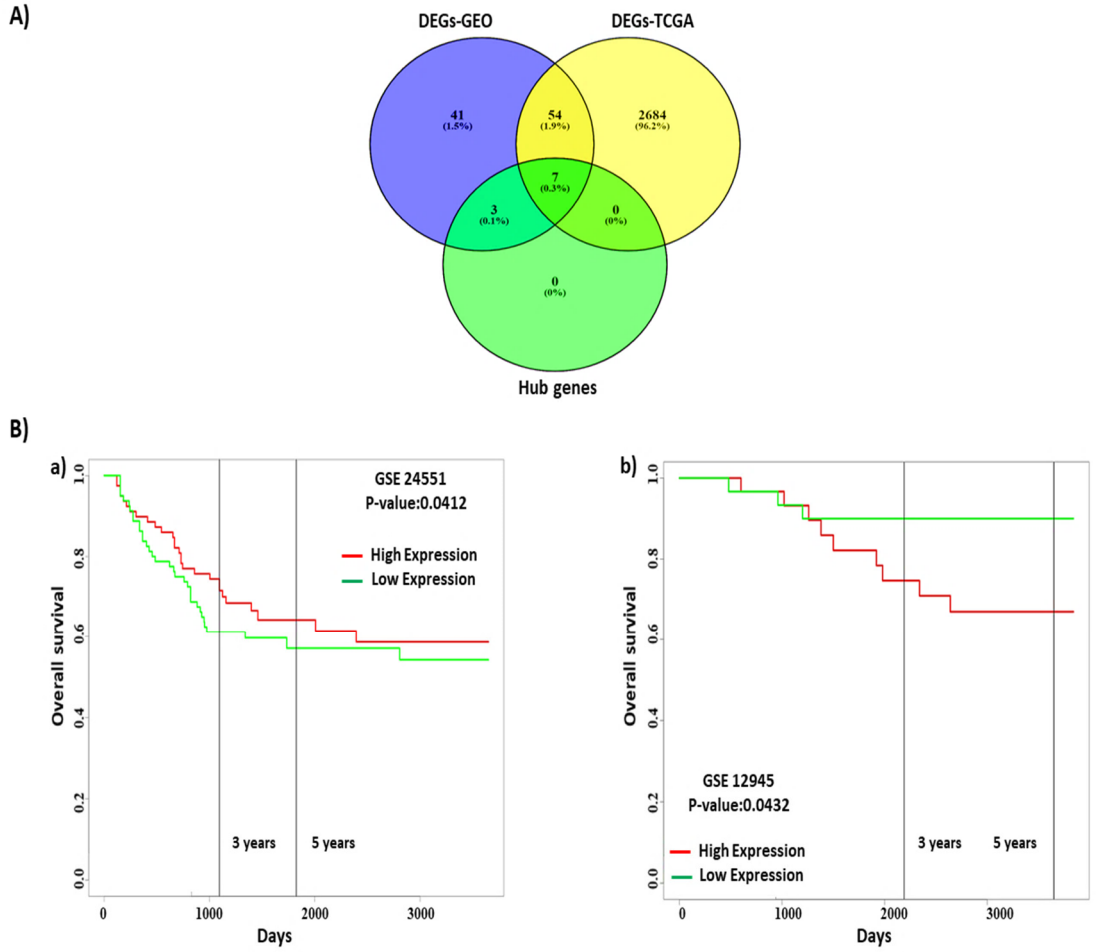


Figure 7. Prognostic value of suggested biomarkers. A: Venn plot is showing the common genes between GEO dataset, TCGA cohort, and Hub genes. B) the survival analysis of the biomarkers. a) Kaplan-Meier plot is showing that lower expression of nine genes (IGF1, MYH11, CLU, MYL9, CXCL12, LMOD1, C3, CNN1, and FOS) as a signature was correlated with poor survival in CRC patients for GSE24551 cohort with P-value:0.0412 and hazard ratio (HR): 0.64(0.41-0.98). b) higher expression of HIST1H2BO was correlated with poor survival in CRC patients for GSE12945 with P-value:0.0432 and HR: 21.84.

4. Discussion

CRC is one of the common malignant tumors of the digestive tract in the world. Dietary and environmental factors, as well as genetic mutations are the main causes of CRC [16,45]. While there have been advances in the diagnosis and treatment of CRC, mortality rate ranks second among all types of cancer because lacking of the early detection [16,46]. Therefore, novel diagnostic and prognostic biomarkers are critically needed.

In this study, we performed an integrative bioinformatic analysis of gene expression microarray data to identify hub genes as diagnostic biomarkers of CRC. A total of 105 DEGs were identified, including 17 upregulated and 88 downregulated DEGs. GO and KEGG pathway enrichment analyses suggested that DEGs were significantly enriched in biological processes related to immune responses

and cancer progression. Subsequently, 10 hub genes with the highest degrees of interactions were explored using the PPI network. The SVM model and ROC curves were used to predict the utility of these hub genes as biomarkers in CRC diagnosis.

Based on functional enrichment analyses, the DEGs were significantly enriched in biological functions including immune response, defense response to bacterium, proteolysis, and innate immune response, systemic lupus erythematosus, arachidonic acid metabolism, complement and coagulation cascades, alcoholism, transcriptional misregulation in cancer, oxytocin signaling pathway, pertussis, focal adhesion, rheumatoid arthritis, and IL-17 signaling. A previous study reported immune destruction, which induces chronic inflammation, as an important cause of CRC; thus, our transcriptomic results are coincided with previous studies that demonstrated that inflammation is a major feature of the tumor microenvironment in CRC [47–49].

Based on the construction of PPI network, *IGF1, MYH11, CLU, MYL9, CXCL12, LMOD1, C3, CNN1, FOS,* and *HIST1H2BO* with a high degree of connectivity were identified as hub genes. Nine of them were significantly downregulated in CRC tissues compared with normal tissues, while the *HIST1H2BO* was significantly upregulated. *MYH11* is a smooth muscle myosin, and its functions are related to cell migration and adhesion, intracellular transport, and signal transduction [50]. A previous study identified that low *MYH11* expression contributes in poor prognosis of CRC patients [51], in addition to forming an oncogenic fusion with core-binding factor subunit beta (CBFB) [52]. Moreover, *IGF-1* could be used as biomarker for CRC patients as shown by others [53]. The expression of *CLU* , a multifunctional intra-/extra-cellular molecular chaperone, is downregulated in malignant tumors compared to the normal colorectal tissue in some cases [54,55] and may be a promising prognostic biomarker for CRC [56].

Moreover, *FOS* is involved in the environmental changes adaptation through formation of transcription factor activating protein 1(AP-1) by its interaction with Jun proto-oncogene (JUN, c-JUN) [57], in addition to it has a crucial role in many diseases including CRC through regulation of many genes related to cancer progression pathways such as proliferation and apoptosis [58]. Recently, using gene expression analysis, Chen et al. reported that FOS could be a potential therapeutic target for CRC [59]. Additionally, *MYL9* expression might be associated with cancer development and metastasis in some tumors, such as non-small-cell lung cancer [60]. Indeed, *CXCL12* has been reported as a hub gene in CRC [61,62], which coincides with our findings. Aberrant expression of *LMOD1* may be associated with cancer development in some types of tumors [50]; however, its role in CRC is unknown.

Additionally, *CNN1* is expressed at significantly higher levels in normal tissue compared to carcinoma tissues in CRC [63] and plays a tumor-suppressive role in different cancers [64,65]. *HIST1H2BO* belongs to the histone family members that are associated with the development and proliferation of multiple cancer types and has been identified as a hub gene in breast cancer [66]. *C3* plays a crucial role in the complement system and contributes to innate immunity [67] ; however, its role in CRC is unknown.

In this study, machine learning methods including SVM model and ROC were conducted to predict the potential application of biomarkers in CRC diagnosis. However, the results exhibited that the AUC values > 0.92 for all genes, demonstrating that these candidate genes could be used as potential biomarkers for CRC diagnosis. Furthermore, GEPIA database exhibited that eight of the hub genes (IGF1, MYH11, CLU, MYL9, CXCL12, LMOD1, C3, and CNN1) were differentially expressed in CRC patients. Subsequently, immunohistochemical analysis revealed that the protein expression of HIST1H2BO was significantly upregulated while the protein expression of CXCL12 and CNN1 were significantly downregulated in CRC tissues. More importantly, the low expression of IGF1, MYH11,

CLU, MYL9, CXCL12, LMOD1, C3, CNN1, and FOS and high HIST1H2BO expression were significantly correlated with poor patient prognosis, therefore, these hub genes exhibit predictive value for CRC patient survival. Collectively, a total of 105 DEGs and 10 hub genes were screened; however, the hub genes could be used as novel biomarkers for CRC patients.

In this research, we further assessed the differential expression of hub genes in various CRC clinical stages, and the results showed that the expression of four genes of hub genes (MYH11, LMOD1, CNN1, and MYL9) was significantly correlated with tumor stages. Therefore, the expression of these 4 hub genes may be used to predict the pathological stage of CRC.

In fact, the whole genome is divided into the protein-coding genes that account for approximately 1.5% of the genome and noncoding RNAs (ncRNAs), which were falsely regarded as transcriptional noise. Based on transcript lengths, ncRNAs can be further divided into small ncRNAs and long ncRNAs (lncRNAs) [68–70]. Indeed, ncRNAs including microRNAs, lncRNAs, and Circular RNAs play critical roles in many biological processes and play an important role in the development of various complex diseases [68–71], therefore the identification of ncRNAs-cancer associations could contribute to the diagnosis and treatment of CRC. Recently, computational models were applied to identify the non-coding RNA biomarker of human complex diseases including cancer to improve cancer prediction, diagnosis and treatment. Therefore, the identification of novel non-coding RNA biomarkers via computational models considers as a future direction for improving CRC prediction, diagnosis and treatment [68].

Furthermore, the necessity of using biomarkers as surrogate outcomes in large trials of major diseases, such as cancer has been widely discussed. However, using the biomarkers as surrogate endpoints and serve as true replacements for relevant clinical endpoints need constant reevaluation and clinical validation. Therefore, we explored a novel CRC biomarker in the present study via transcriptome analysis. Then we further verified these hub genes by using bioinformatics and machine learning methods, however, these biomarkers need further evaluation by biological and clinical investigations to be used as a surrogates in CRC prognosis and prevention [72]. Therefore, those differentially regulated genes may be used as promising biomarker after they have been verified and passed the rigorous examination by the Food and Drug Administration (FDA).

5. Conclusions

We have for the first time identified 10 key genes (IGF1, MYH11, CLU, MYL9, CXCL12, LMOD1, C3, CNN1, FOS, and HIST1H2BO) with diagnostic and prognostic values in CRC. We identified these genes by using comprehensive bioinformatics and machine learning technology. Upon further biological investigation, these genes have the potential to be used in CRC prognosis.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (31971188 and 31370772).

Author contributions

Ahmed Hammad conducted the study, Xiuwen Tang supervised the study, and Mohamed Elshaer assisted in data interpretation. Ahmed Hammad and Xiuwen Tang wrote the manuscript. All authors read and approved the manuscript.

Availability of data

The microarray expressions data of raw and processed files are publicly available in NCBI Gene Expression Omnibus, accession number GSE103512 and GSE128435

Conflict of interest

All the authors have declared no conflict of interests.

Reference

1. R. Siegel, D. Naishadham, A. Jemal, Cancer statistics, 2013, *CA Cancer J. Clin.*, **63** (2013), 11–30.
2. M. R. Sadeghi, F. Jeddi, N. Soozangar, M. H. Somi, N. Samadi, The role of Nrf2-Keap1 axis in colorectal cancer, progression, and chemoresistance, *Tumor. Biol.*, **39** (2017), 1010428317705510.
3. W. Chen, R. Zheng, P. D. Baade, S. Zhang, H. Zeng, F. Bray, et al., Cancer statistics in China, 2015, *CA Cancer J. Clin.*, **66** (2016), 115–132.
4. M. R. Sadeghi, F. Jeddi, N. Soozangar, M. H. Somi, N. Samadi, The role of Nrf2-Keap1 axis in colorectal cancer, progression, and chemoresistance, *Tumour. Biol.*, **39** (2017), 1010428317705510.
5. B. Raphael, R. Hruban, A. Aguirre, R. Moffitt, J. Yeh, C. Stewart, et al., Cancer Genome Atlas Research Network Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma, *Cancer Cell*, **32** (2017), 185–203.
6. I. Kinde, C. Bettegowda, Y. Wang, J. Wu, N. Agrawal, I.-M. Shih, et al., Evaluation of DNA from the Papanicolaou test to detect ovarian and endometrial cancers, *Sci. Transl. Med.*, **5** (2013), 167ra164–167ra164.
7. M. Elshaer, A. I. ElManawy, A. Hammad, A. Namani, X. J. Wang, X. Tang, Integrated data analysis reveals significant associations of KEAP1 mutations with DNA methylation alterations in lung adenocarcinomas, *Aging (Milano)*, **12** (2020), 7183–7206.
8. A. Hammad, Z. H. Zheng, A. Namani, M. Elshaer, X. J. Wang, X. Tang, Transcriptome analysis of potential candidate genes and molecular pathways in colitis-associated colorectal cancer of Mkp-1-deficient mice, *BMC Cancer*, **21** (2021), 607.
9. B. Liang, C. Li, J. Zhao, Identification of key pathways and genes in colorectal cancer using bioinformatics analysis, *Med. Oncol.*, **33** (2016), 016–0829.
10. S. A. Bustin, S. Dorudi, Gene expression profiling for molecular staging and prognosis prediction in colorectal cancer, *Expert Rev. Mol. Diagn.*, **4** (2004), 599–607.
11. V. Kulasingam, E. P. Diamandis, Strategies for discovering novel cancer biomarkers through utilization of emerging technologies, *Nat. Clin. Pract. Oncol.*, **5** (2008), 588–599.
12. M. Nannini, M. A. Pantaleo, A. Maleddu, A. Astolfi, S. Formica, G. Biasco, Gene expression profiling in colorectal cancer using microarray technologies: results and perspectives, *Cancer Treat. Rev.*, **35** (2009), 201–209.
13. M. Ernst, T. L. Putoczki, Targeting IL-11 signaling in colon cancer, *Oncotarget*, **4** (2013), 1860.
14. C. Isella, A. Terrasi, S. E. Bellomo, C. Petti, G. Galatola, A. Muratore, et al., Stromal contribution to the colorectal cancer transcriptome, *Nat. Genet.*, **47** (2015), 312–319.
15. B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, K. W. Kinzler, Cancer genome landscapes, *Science*, **339** (2013), 1546–1558.

16. A. Hammad, Z. H. Zheng, Y. Gao, A. Namani, H. F. Shi, X. Tang, Identification of novel Nrf2 target genes as prognostic biomarkers in colitis-associated colorectal cancer in Nrf2-deficient mice, *Life Sci.*, **238** (2019), 116968.
17. K. GÜÇKIRAN, İ. Cantürk, L. ÖZYILMAZ, DNA microarray gene expression data classification using SVM, MLP, and RF with feature selection methods relief and LASSO, *Süleyman Demirel Üniv. Fen Bilimleri Enst. Derg.*, **23** (2019), 126–132.
18. N. S. Maurya, S. Kushwaha, A. Chawade, A. Mani, Transcriptome profiling by combined machine learning and statistical R analysis identifies TMEM236 as a potential novel diagnostic biomarker for colorectal cancer, *Sci. Rep.*, **11** (2021), 14304.
19. N. Auslander, A. B. Gussow, E. V. Koonin, Incorporating Machine Learning into Established Bioinformatics Frameworks, *Int. J. Mol. Sci.*, **22** (2021), 2903.
20. W. Lian, H. Jin, J. Cao, X. Zhang, T. Zhu, S. Zhao, et al., Identification of novel biomarkers affecting the metastasis of colorectal cancer through bioinformatics analysis and validation through qRT-PCR, *Cancer Cell Int.*, **20** (2020), 105.
21. L. Xu, R. Wang, J. Ziegelbauer, W. W. Wu, R. F. Shen, H. Juhl, et al., Transcriptome analysis of human colorectal cancer biopsies reveals extensive expression correlations among genes related to cell proliferation, lipid metabolism, immune response and collagen catabolism, *Oncotarget*, **8** (2017), 74703–74719.
22. J. Zhou, L. Li, L. Wang, X. Li, H. Xing, L. Cheng, Establishment of a SVM classifier to predict recurrence of ovarian cancer, *Mol. Med. Rep.*, **18** (2018), 3589–3598.
23. J. Mourao-Miranda, A. A. T. S. Reinders, V. Rocha-Rego, J. Lappin, J. Rondina, C. Morgan, et al., Individualized prediction of illness course at the first psychotic episode: a support vector machine MRI study, *Psychol. Med.*, **42** (2012), 1037–1047.
24. X. Chen, Q. F. Wu, G. Y. Yan, RKNNMDA: Ranking-based KNN for MiRNA-Disease Association prediction, *RNA Biol.*, **14** (2017), 952–962.
25. J. Zhi, J. Sun, Z. Wang, W. Ding, Support vector machine classifier for prediction of the metastasis of colorectal cancer, *Int. J. Mol. Med.*, **41** (2018), 1419–1426.
26. M. N. Gabere, M. A. Hussein, M. A. Aziz, Filtered selection coupled with support vector machines generate a functionally relevant prediction model for colorectal cancer, *Oncol. Targets Ther.*, **9** (2016), 3313–3325.
27. Y. R. Liu, Y. Hu, Y. Zeng, Z. X. Li, H. B. Zhang, J. L. Deng, et al., Neurexophilin and PC-esterase domain family member 4 (NXPE4) and prostate androgen-regulated mucin-like protein 1 (PARM1) as prognostic biomarkers for colorectal cancer, *J. Cell. Biochem.*, **120** (2019), 18041–18052.
28. X. Song, T. Tang, C. Li, X. Liu, L. Zhou, CBX8 and CD96 Are Important Prognostic Biomarkers of Colorectal Cancer, *Med. Sci. Monit.*, **24** (2018), 7820–7827.
29. R. C. Team, The R project for statistical computing Available at: https://www. r-project. org, *Accessed January*, **26** (2018).
30. W. H. Da, B. T. Sherman, R. A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nat. Protoc.*, **4** (2009), 44–57.
31. S. Friedman, P. H. Rubin, C. Bodian, E. Goldstein, N. Harpaz, D. H. Present, Screening and surveillance colonoscopy in chronic Crohns colitis, *Gastroenterology*, **120** (2001), 820–826.
32. D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, et al., STRING v10: protein-protein interaction networks, integrated over the tree of life, *Nucleic Acids Res.*, **43** (2015), D447–D452.

33. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome Res.*, **13** (2003), 2498–2504.
34. S. Babicki, D. Arndt, A. Marcu, Y. Liang, J. R. Grant, A. Maciejewski, et al., Heatmapper: web-enabled heat mapping for all, *Nucleic Acids Res.*, **44** (2016), 17.
35. J. Zhou, L. Li, L. Wang, X. Li, H. Xing, L. Cheng, Establishment of a SVM classifier to predict recurrence of ovarian cancer, *Mol. Med. Rep.*, **18** (2018), 3589–3598.
36. L. J. K. Wee, D. Simarmata, Y. W. Kam, L. F. P. Ng, J. C. Tong, SVM-based prediction of linear B-cell epitopes using Bayes Feature Extraction, *BMC Genom.*, **11** (2010), S21.
37. Y. Hu, T. Hase, H. P. Li, S. Prabhakar, H. Kitano, S. K. Ng, et al., A machine learning approach for the identification of key markers involved in brain development from single-cell transcriptomic data, *BMC Genom.*, **17** (2016), 1025–1025.
38. C. D. A. Vanitha, D. Devaraj, M. Venkatesulu, Gene Expression Data Classification Using Support Vector Machine and Mutual Information-based Gene Selection, *Proc. Comput. Sci.*, **47** (2015), 13–21.
39. N. S. Maurya, S. Kushwaha, A. Chawade, A. Mani, Transcriptome profiling by combined machine learning and statistical R analysis identifies TMEM236 as a potential novel diagnostic biomarker for colorectal cancer, *Sci. Rep.*, **11** (2021), 021–92692.
40. T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, D. Haussler, Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, **16** (2000), 906–914.
41. K. Liu, Q. Fu, Y. Liu, C. Wang, An integrative bioinformatics analysis of microarray data for identifying hub genes as diagnostic biomarkers of preeclampsia, *Biosci. Rep.*, **39** (2019).
42. L. K. Boroughs, R. J. DeBerardinis, Metabolic pathways promoting cancer cell survival and growth, *Nat. Cell Biol.*, **17** (2015), 351–359.
43. Z. Tang, C. Li, B. Kang, G. Gao, C. Li, Z. Zhang, GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses, *Nucleic Acids Res.*, **45** (2017), W98–W102.
44. I. M. Copple, The Keap1-Nrf2 cell defense pathway-a promising therapeutic target?, *Adv. Pharmacol.*, **63** (2012), 43–79.
45. K. Tong, O. Pellon-Cardenas, V. R. Sirihorachai, B. N. Warder, O. A. Kothari, A. O. Perekatt, et al., Degree of Tissue Differentiation Dictates Susceptibility to BRAF-Driven Colorectal Cancer, *Cell Rep.*, **21** (2017), 3833–3845.
46. F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA Cancer J. Clin.*, **68** (2018), 394–424.
47. R. B. Sartor, Mechanisms of Disease: pathogenesis of Crohns disease and ulcerative colitis, *Nat. Clin. Pract. Gastroenterol. Hepatol.*, **3** (2006), 390–407.
48. A. J. Schottelius, H. Dinter, Cytokines, NF-κB, Microenvironment, Intestinal Inflammation and Cancer, *Cancer Treat. Res.*, **130** (2006), 67–87.
49. C. Rubie, V. O. Frick, S. Pfeil, M. Wagner, O. Kollmar, B. Kopp, et al., Correlation of IL-8 with induction, progression and metastatic potential of colorectal cancer, *World J. Gastroenterol.*, **13** (2007), 4996–5002.
50. B. Zhao, Z. Baloch, Y. Ma, Z. Wan, Y. Huo, F. Li, et al., Identification of Potential Key Genes and Pathways in Early-Onset Colorectal Cancer Through Bioinformatics Analysis, *Cancer Control*, **26** (2019), 1073274819831260.

51. R. J. Wang, P. Wu, G. X. Cai, Z. M. Wang, Y. Xu, J. J. Peng, et al., Down-regulated MYH11 expression correlates with poor prognosis in stage II and III colorectal cancer, *Asian Pac. J. Cancer Prev.*, **15** (2014), 7223–7228.
52. N. Yamamoto, T. Oshima, K. Yoshihara, T. Aoyama, T. Hayashi, T. Yamada, et al., Clinicopathological significance and impact on outcomes of the gene expression levels of IGF-1, IGF-2 and IGF-1R, IGFBP-3 in patients with colorectal cancer: Overexpression of the IGFBP-3 gene is an effective predictor of outcomes in patients with colorectal cancer, *Oncol .Lett.*, **13** (2017), 3958–3966.
53. S. Wu, F. Wu, Z. Jiang, Identification of hub genes, key miRNAs and potential molecular mechanisms of colorectal cancer, *Oncol. Rep.*, **38** (2017), 2043–2050.
54. T. Chen, J. Turner, S. McCarthy, M. Scaltriti, S. Bettuzzi, T. J. Yeatman, Clusterin-mediated apoptosis is regulated by adenomatous polyposis coli and is p21 dependent but p53 independent, *Cancer Res.*, **64 (**2004), 7412–7419.
55. W. Gomaa, M. Al-Ahwal, H. Al-Maghrabi, A. Buhmeida, M. Al-Qahtani, B. Al-Maghrabi, et al., Expression of clusterin in colorectal carcinoma in relation to clinicopathological criteria, *Malays. J. Pathol.*, **39** (2017), 243–250.
56. P. I. Artemaki, A. D. Sklirou, C. K. Kontos, A. A. Liosi, D. D. Gianniou, I. N. Papadopoulos, et al., High clusterin (CLU) mRNA expression levels in tumors of colorectal cancer patients predict a poor prognostic outcome, *Clin. Biochem.*, **75** (2020), 62–69.
57. S. Mahner, C. Baasch, J. Schwarz, S. Hein, L. Wölber, F. Jänicke, et al., C-Fos expression is a molecular predictor of progression and survival in epithelial ovarian carcinoma, *Br. J. Cancer*, **99** (2008), 1269–1275.
58. R. Ashida, K. Tominaga, E. Sasaki, T. Watanabe, Y. Fujiwara, N. Oshitani, et al., AP-1 and colorectal cancer, *Inflammopharmacology $V 13*, (2006), 113–125.
59. G. Chen, N. Han, G. Li, X. Li, Z. Li, Q. Li, Time course analysis based on gene expression profile and identification of target molecules for colorectal cancer, *Cancer Cell Int.*, **16** (2016), 016–0296.
60. X. Tan, M. Chen, MYLK and MYL9 expression in non-small cell lung cancer identified by bioinformatics analysis of public expression data, *Tumor. Biol.*, **35** (2014), 12189–12200.
61. B. Liang, C. Li, J. Zhao, Identification of key pathways and genes in colorectal cancer using bioinformatics analysis, *Med. Oncol.*, **33** (2016), 111.
62. G. Sun, Y. Li, Y. Peng, D. Lu, F. Zhang, X. Cui, et al., Identification of differentially expressed genes and biological characteristics of colorectal cancer by integrated bioinformatics analysis, *J. Cell. Physiol.*, **234** (2019), 15215–15224.
63. J. E. Drew, A. J. Farquharson, C. D. Mayer, H. F. Vase, P. J. Coates, R. J. Steele, et al., Predictive gene signatures: molecular markers distinguishing colon adenomatous polyp and carcinoma, *PLoS One*, **9** (2014).
64. T. Yamane, K. Asanoma, H. Kobayashi, G. Liu, H. Yagi, T. Ohgami, et al., Identification of the Critical Site of Calponin 1 for Suppression of Ovarian Cancer Properties, *Anticancer Res.*, **35** (2015), 5993–5999.
65. Z. Y. Lin, W. L. Chuang, Genes responsible for the characteristics of primary cultured invasive phenotype hepatocellular carcinoma cells, *Biomed. Pharmacother.*, **66** (2012), 454–458.
66. W. Xie, J. Zhang, P. Zhong, S. Qin, H. Zhang, X. Fan, et al., Expression and potential prognostic value of histone family gene signature in breast cancer, *Exp. Ther. Med.*, **18** (2019), 4893–4903.
67. V. Afshar-Kharghan, The role of the complement system in cancer, *J. Clin. Invest.*, **127** (2017), 780–789.

68. X. Chen, C. C. Yan, X. Zhang, Z. H. You, Long non-coding RNAs and complex diseases: from experimental results to computational models, *Briefings Bioinf.*, **18** (2017), 558–576.
69. X. Chen, D. Xie, Q. Zhao, Z. H. You, MicroRNAs and complex diseases: from experimental results to computational models, *Briefings Bioinf.*, **20** (2019), 515–539.
70. X. Chen, L. Wang, J. Qu, N. N. Guan, J. Q. Li, Predicting miRNA-disease association based on inductive matrix completion, *Bioinformatics*, **34** (2018), 4256–4265.
71. C. C. Wang, C. D. Han, Q. Zhao, X. Chen, Circular RNAs and complex diseases: from experimental results to computational models, *Briefings Bioinfo.*, 2021.
72. K. Strimbu, J. A. Tavel, What are biomarkers?, *Curr. Opin. HIV AIDS*, **5** (2010), 463–466.

AIMS Press