

RESEARCH

Open Access



Bioinformatics screening of colorectal-cancer causing molecular signatures through gene expression profiles to discover therapeutic targets and candidate agents

Md Abu Horaira¹, Md. Ariful Islam¹, Md. Kaderi Kibria¹, Md. Jahangir Alam¹, Syed Rashel Kabir² and Md. Nurul Haque Mollah^{1*}

Abstract

Background Detection of appropriate receptor proteins and drug agents are equally important in the case of drug discovery and development for any disease. In this study, an attempt was made to explore colorectal cancer (CRC) causing molecular signatures as receptors and drug agents as inhibitors by using integrated statistics and bioinformatics approaches.

Methods To identify the important genes that are involved in the initiation and progression of CRC, four microarray datasets (GSE9348, GSE110224, GSE23878, and GSE35279) and an RNA_Seq profiles (GSE50760) were downloaded from the Gene Expression Omnibus database. The datasets were analyzed by a statistical r-package of LIMMA to identify common differentially expressed genes (cDEGs). The key genes (KGs) of cDEGs were detected by using the five topological measures in the protein–protein interaction network analysis. Then we performed in-silico validation for CRC-causing KGs by using different web-tools and independent databases. We also disclosed the transcriptional and post-transcriptional regulatory factors of KGs by interaction network analysis of KGs with transcription factors (TFs) and micro-RNAs. Finally, we suggested our proposed KGs-guided computationally more effective candidate drug molecules compared to other published drugs by cross-validation with the state-of-the-art alternatives of top-ranked independent receptor proteins.

Results We identified 50 common differentially expressed genes (cDEGs) from five gene expression profile datasets, where 31 cDEGs were downregulated, and the rest 19 were up-regulated. Then we identified 11 cDEGs (*CXCL8*, *CEMIP*, *MMP7*, *CA4*, *ADH1C*, *GUCA2A*, *GUCA2B*, *ZG16*, *CLCA4*, *MS4A12* and *CLDN1*) as the KGs. Different pertinent bioinformatic analyses (box plot, survival probability curves, DNA methylation, correlation with immune infiltration levels, diseases-KGs interaction, GO and KEGG pathways) based on independent databases directly or indirectly showed that these KGs are significantly associated with CRC progression. We also detected four TFs proteins (FOXO1, YY1, GATA2 and NFkB) and eight microRNAs (hsa-mir-16-5p, hsa-mir-195-5p, hsa-mir-203a-3p, hsa-mir-34a-5p, hsa-mir-107, hsa-mir-27a-3p, hsa-mir-429, and hsa-mir-335-5p) as the key transcriptional and post-transcriptional regulators of KGs. Finally, our proposed 15 molecular signatures including 11 KGs and 4 key TFs-proteins guided 9 small molecules (Cyclosporin

*Correspondence:

Md. Nurul Haque Mollah
mollah.stat.bio@ru.ac.bd

Full list of author information is available at the end of the article



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

A, Manzamine A, Cardidigin, Staurosporine, Benzo[A]Pyrene, Sitosterol, Nocardiosis Sp, Troglitazone, and Riccardin D) were recommended as the top-ranked candidate therapeutic agents for the treatment against CRC.

Conclusion The findings of this study recommended that our proposed target proteins and agents might be considered as the potential diagnostic, prognostic and therapeutic signatures for CRC.

Keywords Colorectal cancer (CRC), Gene expression profiles, Receptor proteins, Drug agents, Integrated bioinformatics analyses

Introduction

Colorectal cancer (CRC) is an uncontrolled cell growth in the colon, rectum or appendix. It is the second most commonly diagnosed cancer in females and the third in males. The world health organization (WHO) reported in 2018 that over 1.8 million new cases and nearly 862,000 deaths due to CRC worldwide [1, 2]. With more than 2.2 million new cases and 1.1 million fatalities, the global incidence of CRC is projected to be increased 60% by 2030 [3]. The early stages of CRC symptoms are uncharacteristic and frequently ignored or misdiagnosed. Importantly, CRC is diagnosed at the middle or late stages of the disease. It is characteristically identified at the middle or late stages of the disease. The fecal-based examination, enteroscopy and blood-based examination are commonly considered the early detection methods for CRC [4]. However, several instrument-dependent detection methods are time-consuming, laborious and expensive. The leading treatment options for CRC are surgery, adjuvant chemotherapy (for colon cancer), neo-adjuvant radiotherapy (for rectal cancer), and molecular drugs [5, 6]. However, these types of treatments have several drawbacks. According to the previous studies, less than 15% of metastatic CRC is suitable for surgery, the spreading rate of CRC exceeds more than 80% within 3 years after surgery, and the spreading rate exceeds more than 95% within 5 years after surgery [7]. Although there are some advancement in the case of CRC treatments, the 5-year survival time of patients with this disease has not yet increased significantly [6]. Therefore, the identification of new molecular biomarkers is essential for CRC diagnosis, prognosis and new therapies.

However, new drug discovery is a tremendously challenging, time-consuming and expensive task. The main challenges are to explore drug target proteins (receptors) responsible for the diseases and drug agents (small molecules) that can reduce the diseases by the interaction with the target proteins. Genomic biomarkers induced proteins are considered as the key receptors. Transcriptomics analysis is a widely used popular approach to explore genomic biomarkers [8–13]. The repurposing of existing drugs for certain diseases could reduce the time and cost compared to de novo drug development. By this time, several authors have

suggested different sets of key genes (KGs) to explore molecular mechanisms and pathogenetic processes of CRC progression [14–45] in which some studies have employed multiple datasets to identify CRC-causing KGs [15–17, 22, 25, 26, 31, 32, 37, 40–43, 46–49]. Few studies also explored their suggested KGs-guided candidate drug molecules for the treatment against CRC [14, 37, 40–42, 50–59]. However, their published data did not display any common KGs as well as common drug molecules (see Additional file 1: Table S1) in all studies. None of those studies investigated the resistance performance of their suggested KGs-guided drug molecules against the CRC-causing independent KGs suggested by others. We found CRC-causing 170 different KGs and associated 64 different drug molecules in those articles. The articles those suggested therapeutics agents applied enrichment approach on Cmap, geneX-pharma or DGIdb databases to select the KGs-guided candidate agents for the treatment against CRC [14, 37, 40–42, 50–59]. They did not provide pairwise drug-target binding affinity scores, since enrichment techniques cannot calculate pairwise binding scores. So, it may be difficult to select most potential drug-target pairs from the existing studies for experimental validation by the wet-lab researchers. On the other hand, though the total number of KGs 170 is much smaller than the whole genome size, it may be yet much laborious, time consuming and costly for the experimental validation of more than $170 \times 64 = 10,880$ drug-target pairs by the wet-lab researchers. Therefore, in this study, our main objectives were to explore (1) more probable CRC-causing KGs from multiple gene expression profile datasets through the verification with different benchmark datasets and independent databases and (2) proposed KGs-guided candidate drug molecules for the treatment of CRC through the verification of their resistance power against the CRC-causing top-ranked independent KGs suggested by others, by molecular docking analysis.

Materials and methods

The overview of this study including materials and methods is summarized in Fig. 1.

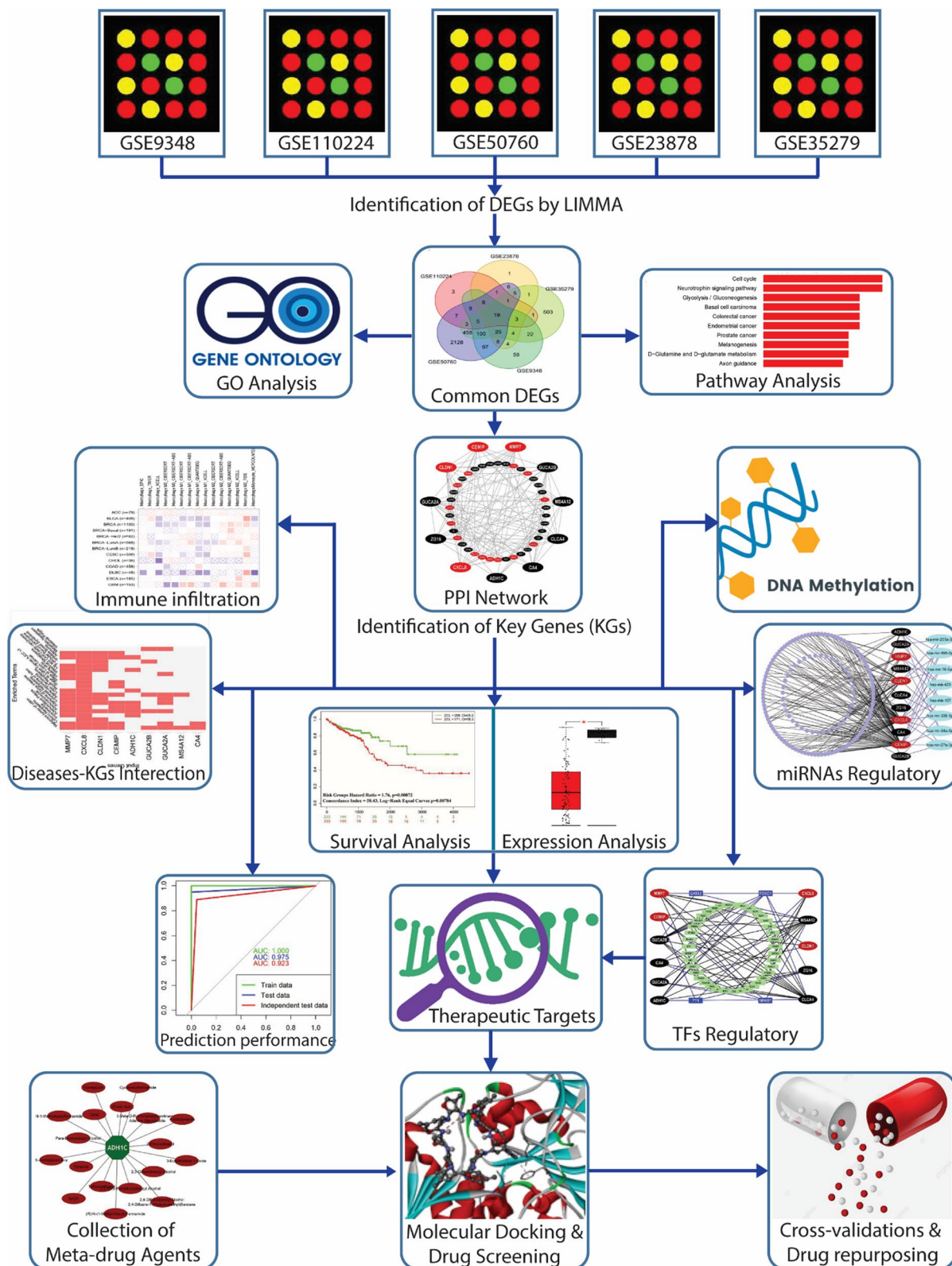


Fig. 1 The pipeline of this study

Data sources and descriptions

We collected gene expression profiles generated from CRC patients for exploring drug targets and small molecules (drug agents) for exploring candidate drugs by molecular docking simulation as described below.

Collection of gene expression profiles for exploring drug-target proteins (receptors)

Four human CRC microarray datasets (GSE9348, GSE110224, GSE23878, and GSE35279) and one RNA-Seq dataset (GSE50760) were downloaded from National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database (<https://www.ncbi.nlm.nih.gov/geo/>). The platform of GSE9348, GSE110224, and GSE23878, were GPL570 [HG-U133_Plus_2] (Affymetrix Human Genome U133 Plus 2.0 Array), GSE35279 was performed by GPL6480 (Agilent-014850 Whole Human Genome Microarray 4 × 44 K G4112F) and GSE50760 was performed by GPL11154 Illumina HiSeq 2000 (Homo sapiens). The summary of this dataset is given in Table 1.

Collection of meta-drug agents for exploring candidate drugs

We collected meta-drug agents from the online database DSigDB [60] with respect to the proposed receptors and FDA approved repurposed drugs for the treatment of CRC patients.

Collection of independent meta-receptors for cross-validation with the proposed drugs

To select the top-ranked receptor proteins (meta-receptors) associated with CRC, we reviewed 33 recently published articles and selected the top-ranked 8 target proteins as the meta-receptors (see Additional file 1: Table S1).

Integrated statistics and bioinformatics approaches

To reach the goal of this study, we applied both statistical and bioinformatics approaches, as discussed below in detail.

Identification of DEGs by using LIMMA

To identify differentially expressed (DEGs) between tumor and normal conditions, we considered the linear models for microarray (LIMMA) data analysis suggested by Smith [61], which can be written as

$$z_g = Y\theta_g + u_g \tag{1}$$

where $z_g = (z_{g1}, z_{g2}, \dots, z_{gn})'$ is the vector of expressions (responses) for gth gene with $n = n_1 + n_2$ samples ($g = 1, 2, \dots, m$), Y is an $n \times 2$ design matrix, $\theta_g = (\theta_{g1}, \theta_{g2})'$ is 2×1 vector ($2 < n$) of effects for two different groups of n samples, and the error vector $u_g \sim N(0, W_g\sigma_g^2)$. Here W_g is a positive definite weight matrix. We want to test the null hypothesis (H_0): $\theta_{g1} = \theta_{g2} \Rightarrow \gamma_g = (\theta_{g1} - \theta_{g2}) = 0$ (that is, gth gene is equally expressed gene (EEG) in both case and control groups) against the alternative hypothesis (H_1): $\theta_{g1} \neq \theta_{g2} \Rightarrow \gamma_g \neq 0$ (that is, the gth gene is a DEG between case and control groups). To test H_0 against H_1 , the moderated t-statistic was formulated by hybridizing the classical and Bayesian approaches in which the posterior variance is substituted into the classical t-statistic in place of the classical sample variance. The moderated t-statistic was defined as

$$\tilde{t}_g = \frac{\hat{z}_g - z_g}{\tilde{s}_g \sqrt{\delta_g}}, \tag{2}$$

which follows t-distribution with $d_g + d_0$ degrees of freedom under H_0 .

Adjusted P values based on the moderated t-statistics and the average of log fold-change (aLog2FC) values of the treatment group with respect to the control group were used to select DEGs as follows

$$DEG_g = \begin{cases} DEG(\text{Upregulated}), & \text{if } adj.p.value < 0.01 \text{ and } aLogFC > 1.0 \\ DEG(\text{Downregulated}), & \text{if } adj.p.value < 0.01 \text{ and } aLogFC < -1.0 \end{cases} \tag{3}$$

Table 1 Details of gene expression profiles that we analyzed

GEO accession	Platform	Year	Country	Normal (n)	Tumor (n)
GSE9348	GPL570	2010	Singapore	12	70
GSE35279	GPL6480	2013	Japan	5	74
GSE23878	GPL570	2010	Saudi Arabia	24	35
GSE110224	GPL570	2018	Greece	17	17
GSE50760	GPL11154	2014	South Korea	18	18

where

$$aLog2FC = \begin{cases} \frac{1}{n_1} \sum_i^{n_1} \log 2(z_{gi}^T) - \frac{1}{n_2} \sum_j^{n_2} \log 2(z_{gj}^C), & \text{if } n_1 \neq n_2 \\ \frac{1}{n} \sum_i^n \log 2\left(\frac{z_{gi}^T}{z_{si}^C}\right), & \text{if } n_1 = n_2 = n \end{cases}$$

Here z_{gi}^T and z_{gj}^C are the expressions for the gth gene with the i th treatment and j th control samples, respectively. We implemented this algorithm using LIMMA R-package to calculate the P values [62] and aLogFC

values to select the DEGs significantly from four gene expression datasets as introduced previously. We separated upregulated and downregulated DEGs for each of four datasets. Then we selected common upregulated and downregulated DEGs for all of four datasets. Then we combine common upregulated DEGs and common downregulated DEGs to construct the common DEGs (cDEGs) set.

Construction of PPI network to identify CRC-causing key genes (KGs)

Protein–protein interaction (PPI) network was constructed to identify common key-genes (KGs). The online STRING-v11 [63] database was used to construct the PPI network of cDEGs. The String database provides critical assessment and integration of protein interactions, including direct (physical) and indirect (functional) associations. To construct a PPI network, the distance ‘D’ between pair of proteins (u, v) is calculated as

$$D(u, v) = \frac{2|N_u \cap N_v|}{|N_u| + |N_v|} \quad (4)$$

where N_u is the neighbor set of u and N_v is the neighbor set of v . Cytoscape plugin cytoHubba was used to rank the nodes of the PPI network, which could be utilized to identify KGs in the network [64, 65]. In the present study, five topological methods, including Degree [66], Bottle-Neck [67], Betweenness [68], and Stress [69] was utilized to identify KGs.

In-silico validation of CRC-causing KGs

An attempt was made to validate the CRC-causing KGs by using different web-tools and independent databases as introduced below.

Expression analysis for KGs by GEPIA web-tool with TCGA RNA-seq data To validate the expression levels of key genes, a gene expression profiling interactive analysis (GEPIA) tool (<http://gepia.cancer-pku.cn/>) was used to explore the related data in TCGA databases, and to analyse the expression levels of key genes in CRC tissues compared with normal tissues [70].

Association of KGs with the immune infiltration levels in different cancers including CRC Tumor Immune Estimation Resource (TIMER) is an integrative resource for investigating the molecular characterization of tumor-immune interactions across various cancer types (<https://cistrome.shinyapps.io/timer/>) [71]. TIMER utilizes a deconvolution statistical method to deduce the abundance of six tumor-infiltrating immune cells, including B cells, CD4⁺ T cells, CD8⁺ T cells, macrophages, neutrophils and DCs from The Cancer Genome Atlas (TCGA).

DNA methylation of KGs MethSurv is used to explore methylation biomarkers associated with the survival of various human cancers [72]. MethSurv is freely available at <https://biit.cs.ut.ee/methsurv>. Through the MethSurv website, we will analyze the DNA methylation analysis of the selected CRC-related genes in the TCGA database.

Association of KGs with different disease The Disease-KGs enrichment analysis was performed using the Enrichr web tool [73] with DisGeNET database [74] to explore other disease risk factors for CRC patients.

Prognostic power analysis of KGs To investigate the prognostic power of KGs, we performed cluster analysis, survival analysis and developed two prediction models using random forest (RF) and AdaBoost classifiers. The survival curve and ROC curve were used to assess the prognosis performance. The online SurvExpress computational tool [75] was used to produce a survival curve. The r-packages ‘gplots’ and ‘ROCR’ were used to produce heatmap and ROC curve, respectively. Exploring drugs by molecular docking simulation.

Exploring GO and KEGG pathway terms that are associated with DEGs including KGs The GO (Gene Ontology) functions [76] and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway enrichment analysis [77] were performed to explore CRC-causing ontology terms (Biological Process (BP), Cellular Component (CC), and Molecular Function (MF)) and pathways that are associated with cDEGs including KGs. To explore the significantly enriched GO terms and KEGG pathways by cDEGs including KGs, let S_i is the annotated gene-set corresponding to the i th type of biological functions or pathways given in the database, and M_i is the number of genes in S_i ($i = 1, 2, \dots, r$); N is the total number of annotated genes those construct the entire combine set $S = \bigcup_{i=1}^r S_i = S_i \cup S_i^c$ such that $N \leq \sum_{i=1}^r M_i$; where S_i^c is the complement set of S_i . Again, let n is the total number of cDEGs of interest and k_i is the number of cDEGs belonging to the annotated gene-set S_i . This problem is summarized by the following contingency table (Table 2):

To find the significantly enriched GO terms and KEGG pathways by our proposed cDEGs, the P value was calculated by the Fisher exact test statistic based on the hypergeometric distribution. We used Enrichr online tool to perform Fisher exact test [78].

Regulatory network analysis of KGs

To identify key transcription factors (TFs) as the transcriptional regulators of KGs, the TFs-KGs interaction

Table 2 Contingency table

Annotated gene-sets (given in the GO terms or KEGG pathway databases)	cDEGs (proposed)	CEEGs (proposed)	Marginal total
i th GO term/KEGG pathway (S_i)	k_i	$M_i - k_i$	M_i
Complement of S_i (S_i^c)	$n - k_i$	$N - M_i - n + k_i$	$N - M_i$
Marginal total	n	$N - n$	N (Grand total)

network was constructed using the publicly available database JASPAR [79]. The interaction network was generated using NetworkAnalyst [80]. To identify key microRNAs (miRNAs) as the post-transcriptional regulators of KGs, the KGs-miRNAs interaction network was constructed by using the publicly available online tool TarBase v8.0 (Release 7.0) [81]. The top degree miRNAs were selected from the networks (miRNAs-KGs) and considered as key miRNAs.

Molecular docking simulation for exploring candidate drug agents

To explore efficient FDA approved repurposed drugs for the treatment of CRC patients, we employed molecular docking simulation between the target receptor proteins and drug agents. We considered our proposed KGs based hub-proteins and associated TFs proteins as the drug target receptor proteins and meta-drug agents collected from online databases and published articles for docking analysis. The molecular docking simulation requires 3-Dimensional (3D) structures of both receptor proteins and candidate drugs. We downloaded the 3D structure of all targeted receptor proteins from Protein Data Bank (PDB) [82] and SWISS-MODEL [66]. The 3D structures of drug agents were downloaded from the PubChem database [83]. The 3D structure of the target proteins was visualized using Discovery Studio Visualizer 2019 [84], and the water molecules, co-crystal ligands which were bound to the protein were removed. Further, the protein was prepared using Swiss-PdbViewer [85] and Auto-Dock Vina [86] in PyRx open-source software by adding charges and minimizing the energy of the protein and subsequently converting it to pdbqt format [86, 87]. The exhaustiveness parameter was set to 8. The Discovery Studio Visualizer 2019 was used to analyze the docked complexes for surface complexes, types and distances of non-covalent bonds. Let A_{ij} denotes the binding affinity between i th target protein ($i = 1, 2, \dots, m$) and j th drug agent ($j = 1, 2, \dots, n$). Then target proteins are ordered according to the descending order of row sums $\sum_{j=1}^n A_{ij}$, $j = 1, 2, \dots, m$, and drug agents are ordered according to the descending order of column sums $\sum_{i=1}^m A_{ij}$, $j = 1, 2, \dots, n$, to select the top ranking few drug agents as the

candidate drugs. Then we validated the proposed repurposed drugs by molecular docking simulation with the top ordered independent receptor proteins associated with CRC published by others.

Results

Identification of cDEGs

We identified 50 cDEGs, including 19 up-regulated (Fig. 2A) and 31 down-regulated (Fig. 2B) genes in CRC tissue, using $\text{adj.}P\text{-Val} < 0.01$ and $\text{logFC} > 1$ as the threshold for down-regulated cDEGs, and $\text{adj.}P\text{-Val} < 0.01$ and $\text{logFC} < -1$ for up-regulated cDEGs. The down and up regulated cDEGs were displayed on the right and left sides respectively in the volcano plot (Fig. 3 and Additional file 2).

Identification of key genes (KGs) from cDEGs

The PPI network of cDEGs was constructed using the STRING database, which includes 49 nodes and 175 edges, with an average node degree of 6.73 and P value $< 1.0e-16$. In the PPI network, Red color indicates up-regulated and black color indicates down-regulated cDEGs, big size and octagon shape indicate common key genes (KGs) (Fig. 4). We used four topological measures (Degree, BottleNeck, Betweenness, and Stress) to select top-ranked 11 KGs (Table 3) that are *CXCL8*, *MMP7*, *CA4*, *ADH1C*, *GUCA2A*, *GUCA2B*, *CEMIP*, *ZG16*, *CLCA4*, *MS4A12* and *CLDN1*, where 4 KGs (*CXCL8*, *CEMIP*, *CLDN1*, and *MMP7*) were up-regulated and the rest 7 KGs were downregulated.

In-silico validation of CRC-causing KGs by using different web-tools and independent databases

Expression analysis for KGs by GEPIA web-tool with TCGA RNA-seq data

In the GEPIA database, differences in transcriptional expression of the hub gene between CRC tissues and normal tissues were again verified. Combining with the box plot results, eleven potential KGs further were screened out. Based on the GEPIA database to test the relative expression of KGs mRNA, it was determined that our proposed KGs (*CXCL8*, *CEMIP*, *MMP7*, *CA4*,

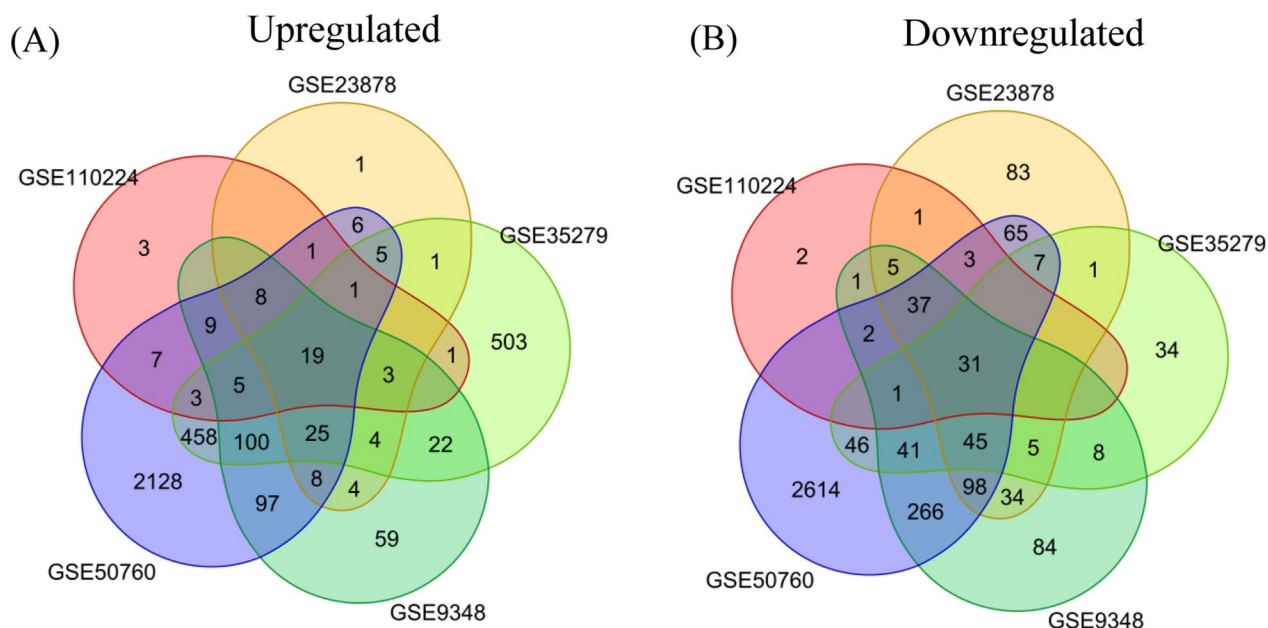


Fig. 2 Common DEGs (cDEGs) among the five GEO datasets for **A** up-regulated and **B** downregulated

ADHIC, GUCA2A, GUCA2B, ZG16, CLCA4, MS4A12 and *CLDN1*) may be closely related to the occurrence and development of CRC (Fig. 5).

Correlation between KGs and immune infiltration levels in different cancers including CRC

We investigated the relationship of different tumors infiltrates immune cell types (B cell, CD8 T cell, CD4 T cell, neutrophil, macrophage and dendritic cell (DC)) with the expressions of KGs (Additional file 3). We observed (Additional file 4) that our proposed KGs are significantly associated with different tumor infiltrates immune cells under different databases of COAD (colon adenocarcinoma) and READ (Rectum adenocarcinoma). Compelling evidence has demonstrated that tumor-infiltrating lymphocytes are significantly associated with survival in cancer. Therefore, we investigated whether KGs expression was related to immune infiltration levels in lung cancer by TIMER. Tumor purity is an important factor affecting the analysis of immune infiltration. Interestingly, our results indicated that KGs expression was correlated with poor prognosis and high immune infiltration in CRC. KGs were highly expressed in monocytes (non-classical and classical) and B cells (naïve). In contrast, KGs expression was not significantly correlated with tumor purity or infiltrating levels of CD8⁺ T cells, CD4⁺ T cells or neutrophils in CRC. CA4 expression levels were positively correlated with infiltrating levels of B cells ($r=0.22, P=2.47E-04$), CD8⁺ T cells ($r=-0.16, P=7.49E-03$), CD4⁺ T cells ($r=0.19, P=1.99E-03$),

macrophages ($r=-0.18, P=2.19E-03$), neutrophils ($r=0.38, P=4.6E-11$) and DCs ($r=0.23, P=9.73E-05$) in COAD (Additional files 3 and 4). CLCA4 expression levels were also positively correlated with infiltrating levels of B cells ($r=0.32, P=1.92E-03$), CD8⁺ T cells ($r=-0.23, P=2.95E-02$), CD4⁺ T cells ($r=0.39, P=1.36E-04$), macrophages ($r=-0.48, P=8.42E-07$), neutrophils ($r=0.5, P=4.50E-07$) and DCs ($r=0.33, P=1.10E-03$) in READ (Additional file 3 and 4). These findings strongly suggest that KGs plays an important role in immune infiltration in CRC, especially infiltration of Macrophage, T cell CD8⁺, T cell CD4⁺, Neutrophil, Myeloid dendritic cell, and B cell.

DNA methylation of KGs

DNA methylation at CpG (CG) sites play the vital role in cancer progression. Therefore, we investigated DNA methylation of KGs (*CXCL8, CEMIP, MMP7, CA4, ADHIC, GUCA2A, GUCA2B, ZG16, CLCA4, MS4A12* and *CLDN1*) at CpG sites by MethSurv web-tool with TCGA database. We observed that seven KGs (*CEMIP, MMP7, CA4, GUCA2B, ZG16, CLCA4, MS4A12*) are significantly methylated at CpG sites (Table 4). The hypermethylation/downregulation gene *CEMIP* has six CpG sites with a P value < 0.05, the hypomethylation/upregulation gene *GUCA2B* has four CpG sites with a P value of < 0.05, the hypomethylation/upregulation gene *MS4A12* has two CpG sites with a P value < 0.05, and the hypomethylation/upregulation gene *MMP7, CLCA4,*

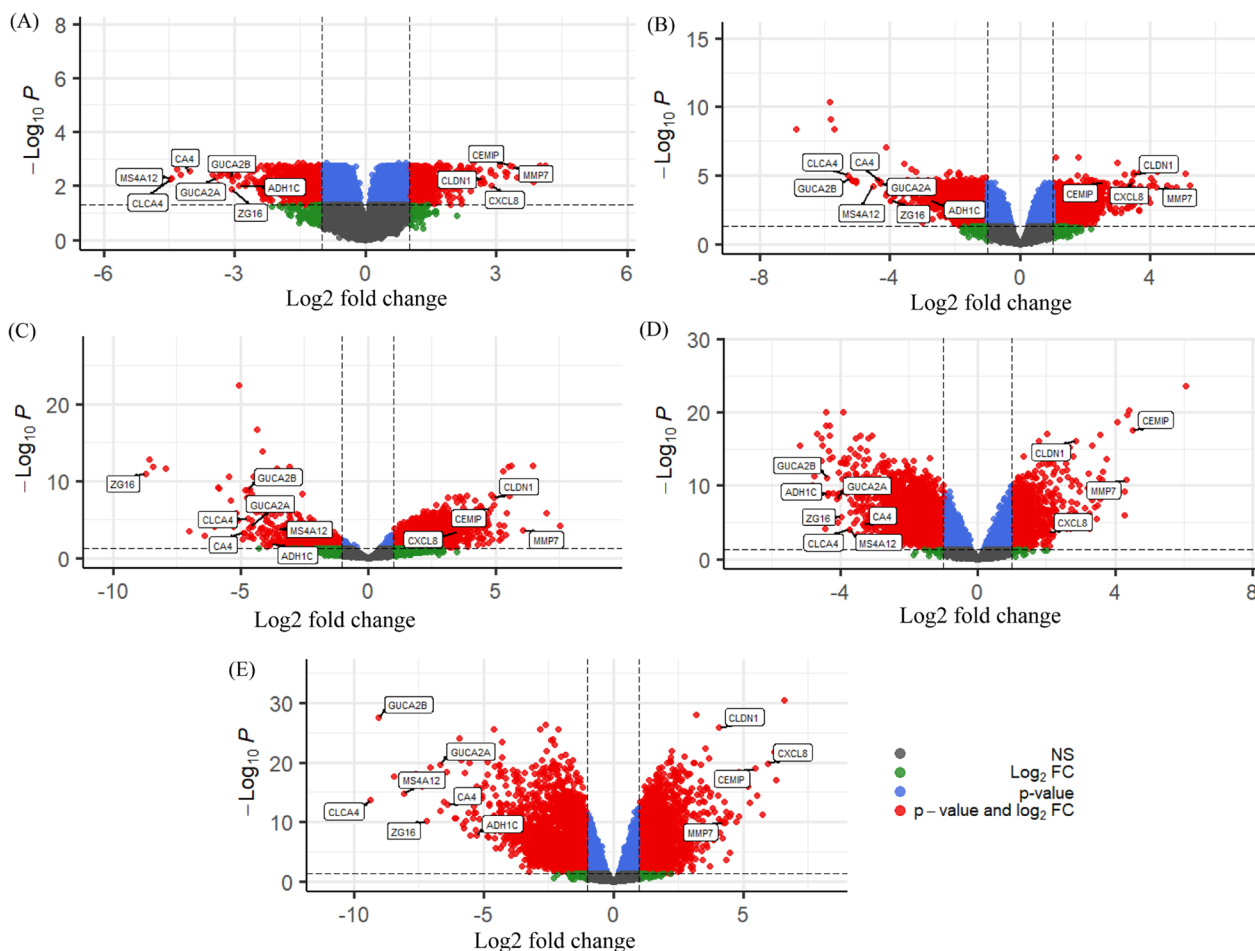


Fig. 3 The five GEO datasets volcano plots of **A** GSE110224, **B** GSE50760, **C** GSE35279, **D** GSE23878 and **E** GSE9348. Ass color point are Not Significant (NS) according to Log_2FC and P value threshold, green color is Log_2FC ($\text{Log}_2\text{FC} < -1$ and $\text{Log}_2\text{FC} > 1$), blue color is P value ≤ 0.05 , and red color points are satisfying the Log_2FC and P value threshold

ZG16 has one CpG site with a P value of < 0.05 , which is statistically significant (Table 4). We found that the difference in DNA methylation between CG12358698 of CEMIP, CG23532119 of MS4A12, CG00656728 of GUCA2B, CG24963041 of MMP7, CG26310643 of CLCA4, CG09229061 of ZG16, CG00200645 of CA4 and CG07510230 of ZNRF2 was most pronounced.

Association of KGs with different diseases including CRC

The disease-KGs interaction analysis showed that KGs are significantly associated with different types of colon or rectal cancers including Malignant tumor of colon, Colonic Neoplasms, Adenomatous Polyps, Adenocarcinoma, Adenoma of large intestine, Colorectal Neoplasms, Adenocarcinoma of colon, Colon Carcinoma, Stage III Colon Cancer AJCC v7, Stage III Colon Cancer, Intestinal Neoplasms, Adenoma and Metastatic Neoplasm (Fig. 6 and Table S2 in Additional file 1).

Prognostic power analysis

We considered both supervised and unsupervised learnings, including multivariate survival analysis, to investigate the prognostic power of KGs. Figure 7A shows that KGs can separate case and control samples accurately by the unsupervised hierarchical clustering (HC). The multivariate survival curves, based on the expressions of 11 KGs, separated the low and high-risk groups significantly (Fig. 7B). In the case of supervised learning, at first, we considered the expression profiles of 11 KGs from three datasets (GSE9348, GSE23878 and GSE110224) that contained 60 tumors and 50 control samples in total. Then we partitioned these datasets in to training (70%) and test (30%) sets. Then we trained one popular classifier known as random forest (RF). To test the prediction performance of the model, we also considered the expressions of 11 KGs from another two dataset GSE35279 and TCGA as the independent test set. Figure 7C showed the

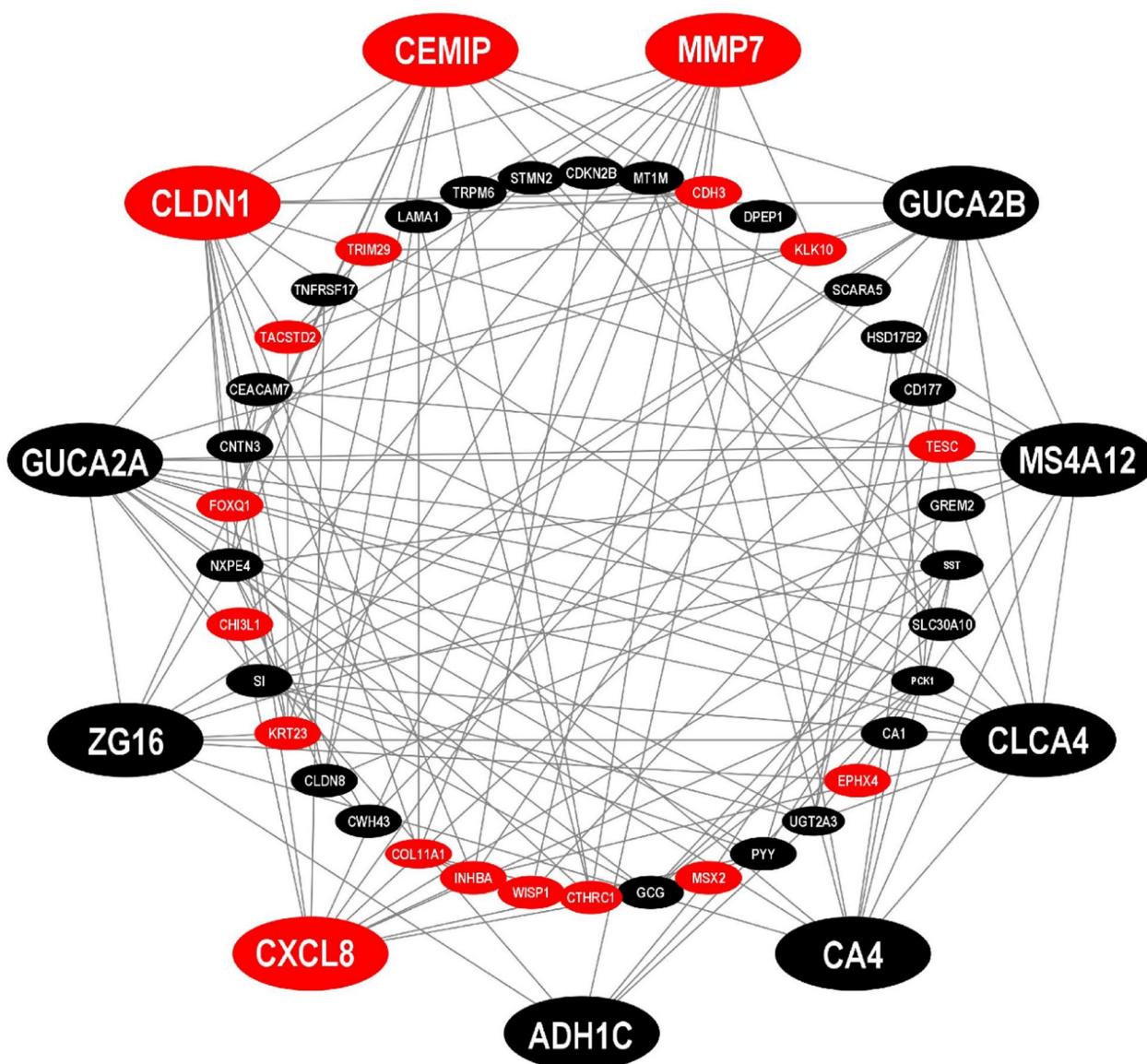


Fig. 4 Network of PPIs for common cDEGs that have been identified. Red color nodes are upregulated and black color nodes are downregulated. The outer circle of the image is common key genes (KGs)

Table 3 Selection of KGs by combining the top ranked genes of five topological measurements with the PPI network

Degree (A)	BottleNeck (B)	Betweenness (C)	Stress (D)	Key genes (KGs) (A ∪ B ∪ C ∪ D)
GUCA2A	CLDN1	CXCL8	CLDN1	GUCA2A, GUCA2B, CLDN1, CLCA4, MS4A12, MMP7, CEMIP, CXCL8, ADH1C, ZG16, CA4
GUCA2B	CXCL8	CLDN1	GUCA2A	
CLDN1	CLCA4	GUCA2A	GUCA2B	
CLCA4	MMP7	MMP7	CXCL8	
MS4A12	ZG16	CA4	CA4	

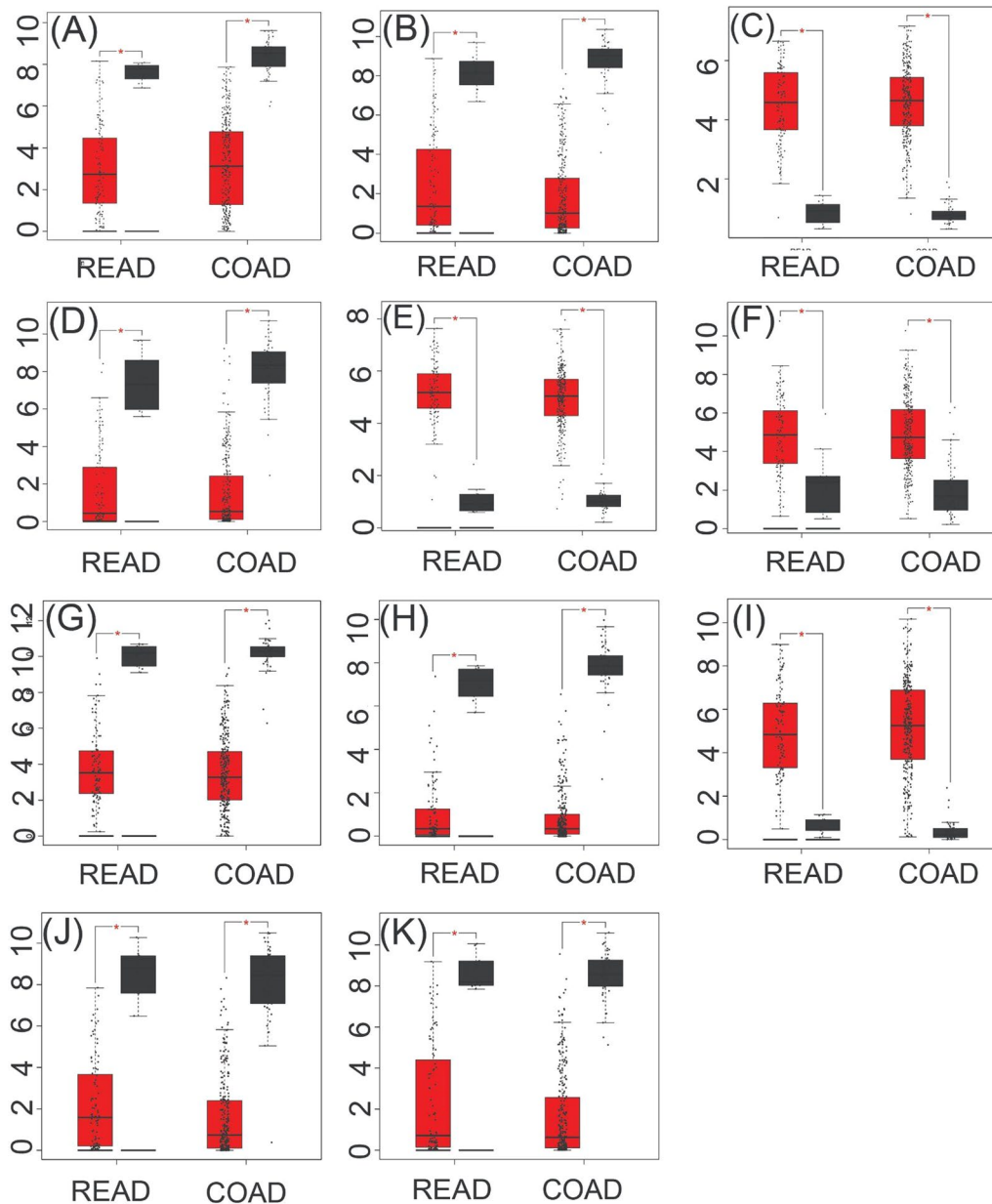


Fig. 5 The expression level of hub genes in CRC. **A** ADH1C; **B** CA4; **C** CEMIP; **D** CLCA4; **E** CLDN1; **F** CXCL8; **G** GUCA2A; **H** GUCA2B; **I** MMP7; **J** MS4A12 and **K** ZG16. The red and gray boxes represent cancer and normal tissues, respectively. Colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ)

ROC curves based on the train, test performance, and independent test dataset of RF prediction model. The AUC values (area under the ROC curve) for RF were 1.00 with train data, 0.988 with test data, 0.943 with independent test data and 0.90 with TCGA dataset. Thus, both prediction models based on RF classifiers showed good performance for each of the dependent and independent test datasets of KGs.

Exploring CRC-causing GO and KEGG pathway terms that are associated with cDEGs including KGs

The GO functional enrichment analysis of showed that 185 GO-BP terms, 9 GO-CC terms and 38 GO-MF terms are enriched by the cDEGs genes, where KGs were involved with 57 BPs, 6 CCs and 21 MFs. Among the enriched GO functions including KGs, 6 GO-BP

Table 4 The significant prognostic value of CpG in three key genes

Gene-CpG	HR	P value
CEMIP-Body-Open_Sea-CG12358698	2.657	0.001
CEMIP-Body-Open_Sea-CG12098156	2.275	0.001
CEMIP-Body-Open_Sea-CG04847610	1.899	0.008
CEMIP-Body-Open_Sea-CG17820039	3.085	0.027
CEMIP-Body-Open_Sea-CG21838329	2.665	0.045
CEMIP-5'UTR-Open_Sea-CG09579081	3.836	0.019
MMP7-TSS1500-Open_Sea-cg24963041	1.822	0.016
MS4A12_5'UTR-Open_Sea-cg09257456	0.196	0.003
MS4A12_TSS200-Open_Sea-cg23532119	4.164	0.009
CLCA4-Body-Island-cg26310643	0.259	0.018
ZG16-TSS1500-Open_Sea-cg09229061	6.022	0.001
CA4-Body-Open_Sea-cg00200645	3.173	0.022
GUCA2B-TSS1500-Open_Sea-cg00656728	11.395	0.001
GUCA2B-TSS200-Open_Sea-cg10179693	3.585	0.009
GUCA2B-TSS200-Open_Sea-cg14848143	3.185	0.023
GUCA2B-1stExon-Open_Sea-cg19728577	7.457	0.001

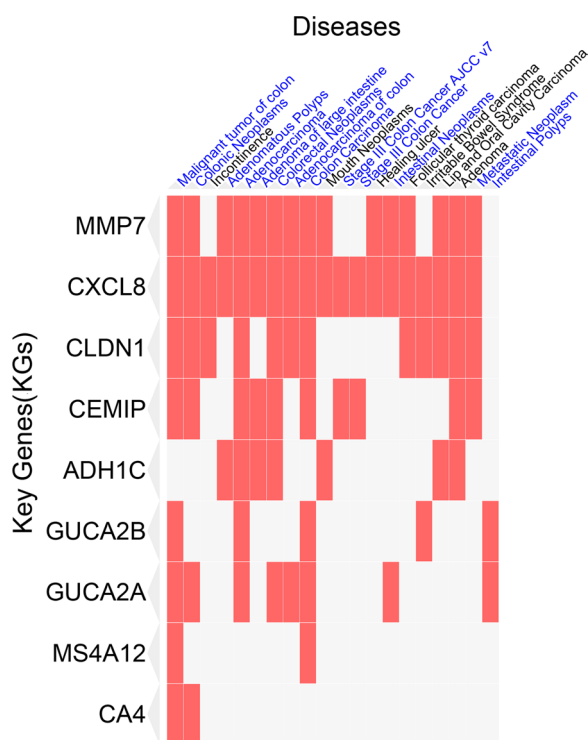


Fig. 6 KGs-Diseases interaction, where blue color highlighted risk factors are CRC related

terms (GO:0034,31~cell junction maintenance, GO:0098742~cell-cell adhesion via plasma-membrane adhesion molecules, GO:0045216~cell-cell junction organization, GO:0008285~negative regulation of cell

population proliferation, GO:0030334~regulation of cell migration, and GO:0048565~digestive tract development), 5 GO-CC terms (GO:0046658~anchored component of plasma membrane, GO:0062023~collagen-containing extracellular matrix, GO:0005923~bicellular tight junction, GO:0043296~apical junction complex, and GO:0005911~cell-cell junction) and 6 GO-MF terms (GO:0005179~hormone activity, GO:0030250~guanylate cyclase activator activity, GO:0048018~receptor ligand activity, GO:0005254~chloride channel activity, GO:0008237~metallopeptidase activity, and GO:0045236~CXCR chemokine receptor binding) were reported by other researchers as the BPs of CRC (see Table 3 and discussion section for more details). The KEGG pathway enrichment analysis of cDEGs showed that 8 pathways are enriched by the KGs. Among them, KGs involving Nitrogen metabolism, Proximal tubule bicarbonate reclamation, Cell adhesion molecules, Pathogenic *Escherichia coli* infection, Human T-cell leukemia virus 1 infection, Amoebiasis, Leukocyte transendothelial migration, and Cytokine-cytokine receptor interaction was also reported by other researchers as the pathways of CRC development (see Table 5 and discussion section for more details as before).

Regulatory network analysis of KGs

We constructed KGs versus transcription factors (KGs-TFs) interaction network to identify top ranking few TFs as the key transcriptional regulators of KGs. We selected the top 4 key TFs (FOXC1, YY1, GATA2 and NFKB1) as the vital transcriptional regulators of KGs with degree ≥ 4, where the green color rectangle indicates top degree key TFs and, red and black color ellipse indicates KGs (Fig. 8A). To identify top ranking few micro-RNA (miRNA) as the key post-transcriptional regulators of KGs, we constructed a KGs-miRNAs interaction network. We selected the top 8 key miRNAs (hsa-mir-16-5p, hsa-mir-195-5p, hsa-mir-203a-3p, hsa-mir-34a-5p, hsa-mir-107, hsa-mir-27a-3p, hsa-mir-429, and hsa-mir-335-5p) as the vital regulators of KGs with degree ≥ 4, where green color rectangle indicates top degree key miRNAs and, red and black color ellipse indicates KGs (Fig. 8B).

Exploring candidate drug agents by molecular docking simulation

To explore candidate drugs for CRC, we considered 11 KGs based proteins (CXCL8, MMP7, CA4, ADH1C, GUCA2A, GUCA2B, CEMIP, ZG16, CLCA4, MS4A12 and CLDN1) and its regulatory key 4 TFs proteins (FOXC1, YY1, GATA2 and NFKB1) as the m=15 drug target receptors. The 3-Dimension (3D) structure of CXCL8, MMP7, ZG16, CA4, YY1 and NFKB1 were

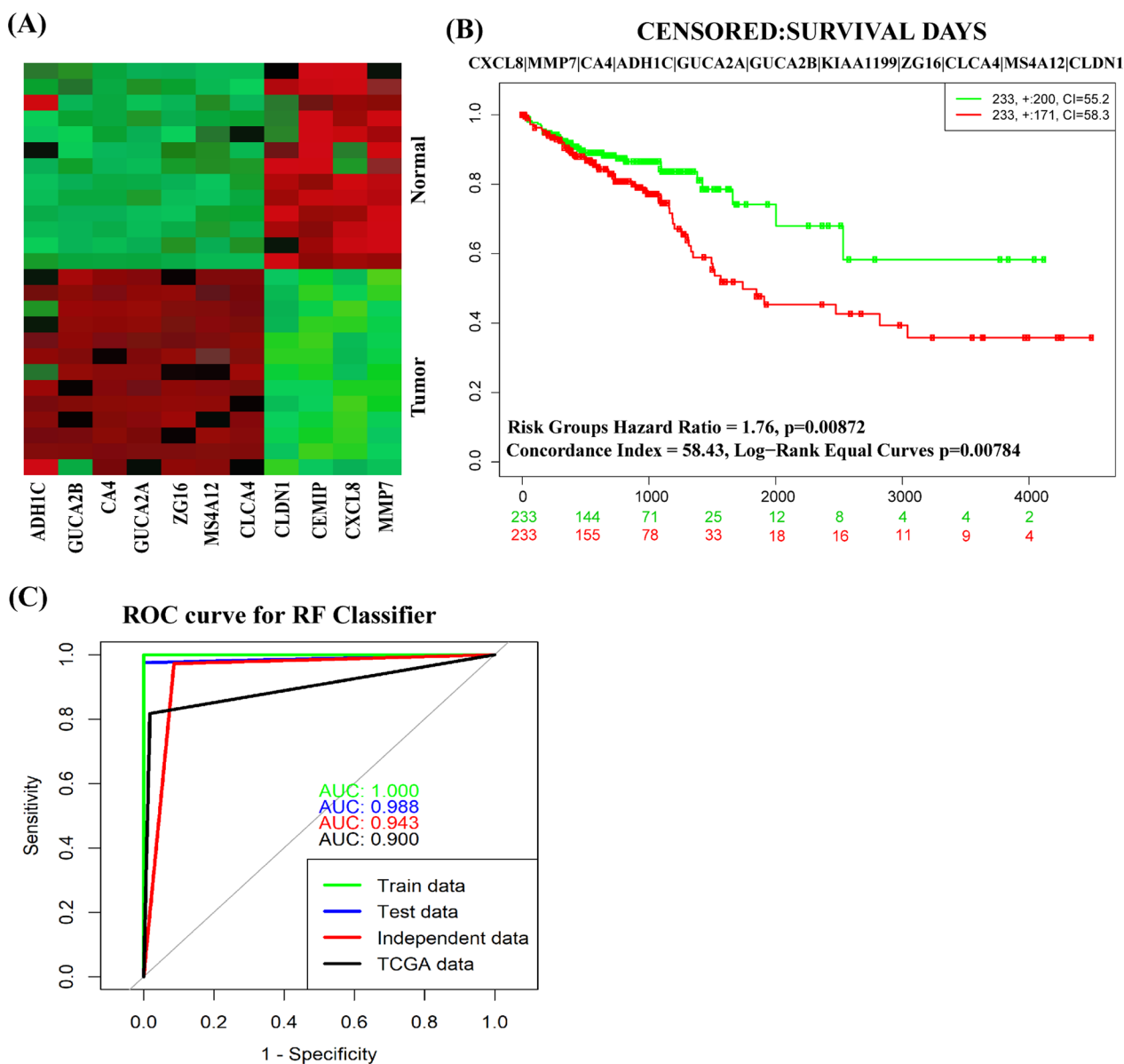


Fig. 7 The prognostic powers of KGs were displayed by **A** a Heatmap of hierarchical clustering, **B** multivariate survival curves with KGs, and **C** ROC curves of prediction models with KGs

downloaded from Protein Data Bank (PDB) with the PDB codes 6N2U, 1MMQ, 3APA, 5KU6, 4C5I and 1NFI and the rest of them, such as GUCA2A, GUCA2B, CLDN1, CLCA4, MS4A12, FOXC1 and GATA2 targets were downloaded from SWISS-MODEL using UniProt with IDs Q02747, Q16661, O95832, Q14CN2, Q9NXJ0, Q12948, and P23769 respectively. Then we considered 92 meta-drug molecules from the DSigDB database and 64 meta-drugs from the published articles and the Food and Drug Administration (FDA) as drug agents. The 3D structures of drug agents were downloaded from

the PubChem database. Then we performed a molecular docking simulation between our proposed receptors and meta-drug agents. The binding affinity score matrix between the ordered receptors and ordered drug agents were displayed in Fig. 9A. We observed that Cyclosporin A produces highly significant binding affinity scores with all $m=15$ target proteins, and their average binding affinity scores across all targets were -9.46 (kcal/mol). The 2th and 3th top ordered drugs (Manzamine A and Cardidigin) produced highly significant binding affinity scores with 14 target proteins, and their average binding

Table 5 Top Enriched gene ontology (GO) terms and KEGG pathways by the proposed cDEGs highlighting cKGs

Term	Overlap	P value	cKGs
<i>Biological process</i>			
Cell junction maintenance (GO:0034331) [88]	2/14	6.14E−04	CLDN1
Cell–cell adhesion via plasma-membrane adhesion molecules (GO:0098742) [89]	4/170	0.001067	CLDN1
Calcium-independent cell–cell adhesion via plasma membrane cell-adhesion molecules (GO:0016338) [89]	2/20	0.00127	CLDN1
Cell–cell junction organization (GO:0045216) [89]	3/82	0.001342	CLDN1
Negative regulation of cell population proliferation (GO:0008285) [90]	5/379	0.003239	CXCL8
Regulation of cell migration (GO:0030334) [91]	5/408	0.00443	MMP7;CLDN1
<i>Molecular function</i>			
Hormone activity (GO:0005179) [92]	5/78	1.96E−06	GUCA2A
Guanylate cyclase activator activity (GO:0030250) [93]	2/5	6.85E−05	GUCA2B;GUCA2A
Receptor ligand activity (GO:0048018) [94]	6/307	1.56E−04	GUCA2A
Chloride channel activity (GO:0005254) [95]	2/64	0.012507	CLCA4
Metallopeptidase activity (GO:0008237) [96]	2/121	0.040942	MMP7
CXCR chemokine receptor binding (GO:0045236) [97]	1/17	0.044125	CXCL8
<i>Cellular component</i>			
Anchored component of plasma membrane (GO:0046658) [98]	2/46	0.006619	CA4
Collagen-containing extracellular matrix (GO:0062023) [99]	4/380	0.018081	ZG16
Bicellular tight junction (GO:0005923) [100]	2/78	0.018197	CLDN1
Apical junction complex (GO:0043296) [101]	2/98	0.027852	CLDN1
Cell–cell junction (GO:0005911) [102]	3/271	0.035073	CLDN1
<i>KEGG</i>			
Nitrogen metabolism [22]	1/7	9.13E−04	CA4
Proximal tubule bicarbonate reclamation [103]	1/23	0.001682	CA4
Cell adhesion molecules [104]	3/148	0.007098	CLDN1
Pathogenic Escherichia coli infection [105]	3/197	0.015369	CXCL8;CLDN1
Amoebiasis [44]	2/102	0.029984	CXCL8
Leukocyte transendothelial migration [106]	2/114	0.036749	CLDN1
Cytokine-cytokine receptor interaction [107]	3/295	0.043339	CXCL8

affinity scores across all $m=15$ targets were -8.22 and -8.19 , respectively. The 4th to 10th top ordered drug Staurosporine, Benzo[A] Pyrene, Sitosterol, Nocardiosis Sp, Troglitazone, K-252a, and Riccardin D produced significant binding affinity scores with 14 target proteins, and the average binding affinity score was -7.76 , -7.71 , -7.69 , -7.68 , -7.66 , -7.64 , and -7.62 respectively. The other drugs (lead compounds) produced significant binding affinity scores with less than 13 target proteins out of 15, and their average binding affinity scores were negatively smaller than -7.5 . Therefore, we considered the top ordered nine drugs (Cyclosporin A, Manzamine A, Cardidigin, Staurosporine, Benzo[A]Pyrene, Sitosterol, Nocardiosis Sp, Troglitazone and Riccardin D) as the candidate drugs in our study. We also examined their complete interaction profile, including hydrogen bonds, hydrophobic, halogen/ salt Bridge and electrostatic interactions in Table 6.

Performance investigation of proposed drugs by cross-validation with the top-ranked independent receptors

To investigate the resistance performance of our proposed 9 candidate drugs against the state-of-the-art alternative receptors for CRC by molecular docking, we considered the top-ranked 8 independent receptors (MYC, CDK1, CXCL1, CXCL8, CXCL12, TIMP1, AURKA, and TOP2A) published by others in different 36 articles for CRC (see Additional file 1: Table S1), where the receptor CXCL8 was common with our proposed receptor. The 3D structure of MYC, CDK1, CXCL12, TIMP1, AURKA, and TOP2A was downloaded from the PDB database with the PDB codes 6G6K, 6GU2, 6SHR, 2J0T, 6VPM, AND 1ZXM, respectively and for another one CXCL1, downloaded from SWISS-MODEL using UniProt with ID P09341. The we

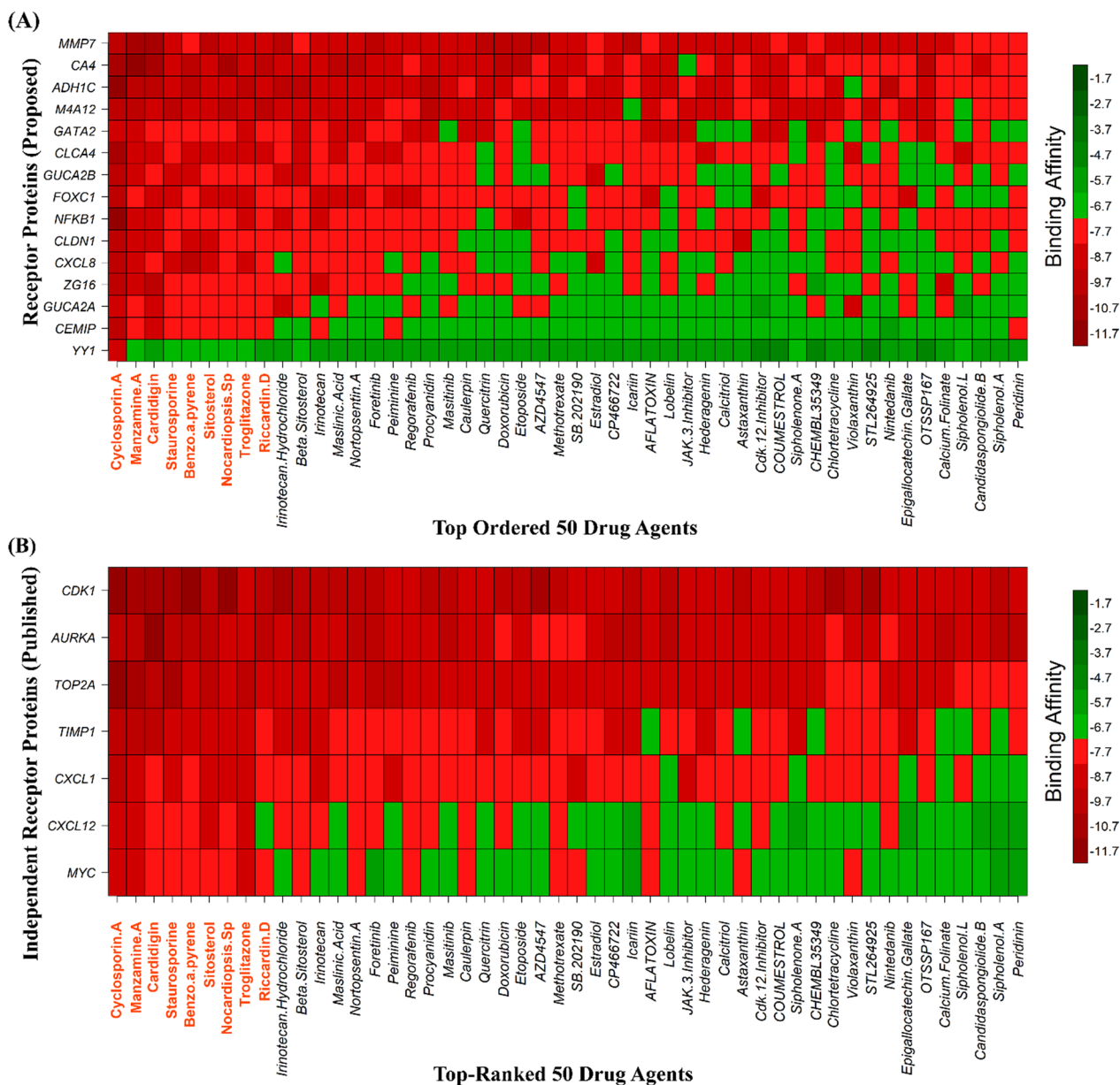


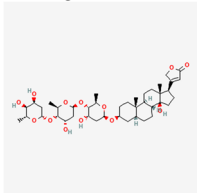
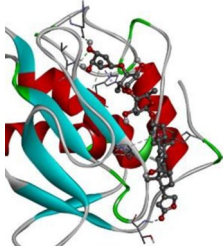
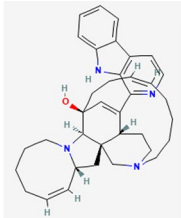
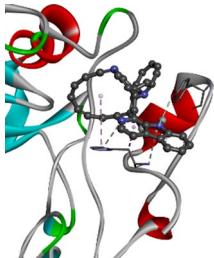
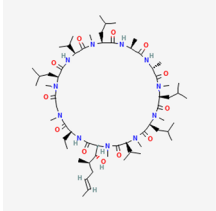
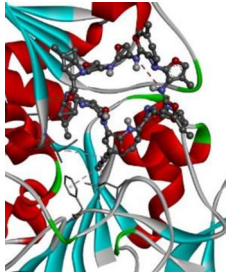
Fig. 9 Molecular docking simulation results for exploring candidate drugs against CRC. **A** Image of binding affinity scores of proposed ordered receptor proteins with the top 50 ordered. **B** Image of binding affinity scores of the top-ranked independent receptors published by others with the top 50 ordered

performed molecular docking analysis of 7 independent receptors with all of 160 drug agents. Figure 9B showed that our proposed 9 candidate drugs are also detected as the independent receptor-guided top-ranked 9 drugs. Therefore, we can strongly recommend that the proposed drugs might be more effective candidates than the other drugs for the treatment against CRC.

Connectivity map (CMap) analysis to discover the mechanism of action of drug agents

In an effort to elucidate its mechanism of action, we defined a signature for Troglitazone, Cardidigin and Staurosporine. High connectivity scores were found for multiple instances of five heat shock protein inhibitors: Angiotensin receptor antagonist, Topoisomerase

Table 6 The 3-dimension view of strong binding interactions between targets and drugs is shown in the 4th column

Potential targets	Structure of lead compounds	Binding affinity (kCal/mol)	The 3d view and interactions of complex	Interacting amino acids		
				Hydrogen bond	Hydrophobic interactions	Electrostatic
MMP7	Cardidigin 	- 10.4		LEU181, THR189, ASN179, GLU219	HIS218, PHE103, PHE185	-
CA4	Manzamine A 	- 10.8		HIS4, TYR11, HIS4	HIS4, HIS4	-
ADH1C	Cyclosporin A 	- 11.7		ALA317	LEU116, ILE318, PHE93	-

Key interactions amino acids and their binding types with potential targets were shown in the last column

inhibitor, Glycogen synthase kinase inhibitor, DNA dependent protein kinase inhibitor, and MTOR inhibitor. Despite the differences in the cells used to generate the query signature and reference profiles, the three highest-scoring compounds in the Con nectivity Map were Troglitazone, Cardidigin (Digitoxin use this name to use in Cmap) and Staurosporine (Fig. 10A). More important, the Connectivity Map also revealed strong connectivity with ten structurally distinct compounds, mocetinostat, ryuidine, cyclopamine, dorsomorphin, JNJ-7706621, quinoclamine, SU-11652, bisacodyl, alvocidib, and rottlerin respective inhibitor are show in Fig. 10B. Cyclopamine and alvocidib compounds are not connected with Troglitazone.

AK3, *OAZ2*, *NDUFAF4*, *EGFR*, *GNE*, *MAPK14*, *CSNK1G2*, *HSP90AB1*, *ZNF449*, and *GATAD1* genes depicts high positive connectivity with each of the drugs Troglitazone, Cardidigin and Staurosporine and their median connectivity score belongs to 97.96–97.53 (out

of ± 100) which display in the Fig. 10C with corresponding enriched pathways of the connected gene. Moreover, the drug staurosporine was positively connected with 6 other genes namely *OTUD3*, *TCF7L1*, *C2*, *TAAR1*, *PRDM1*, and *BMP2* with corresponding enriched pathways.

Discussion

The molecular mechanism of colorectal cancer (CRC) is not yet completely clear to the researchers. So potential molecular signatures are required to disclose molecular mechanisms of CRC and its therapeutic agents. The integrated statistics and bioinformatics analyses are now widely using to explore potential molecular signatures of malignant tumors [108]. Transcriptomics analysis is a popular way of identifying DEGs between normal and tumor tissue samples [109]. Therefore, in this study, we considered the integrated bioinformatics analyses for exploring common genomic biomarkers

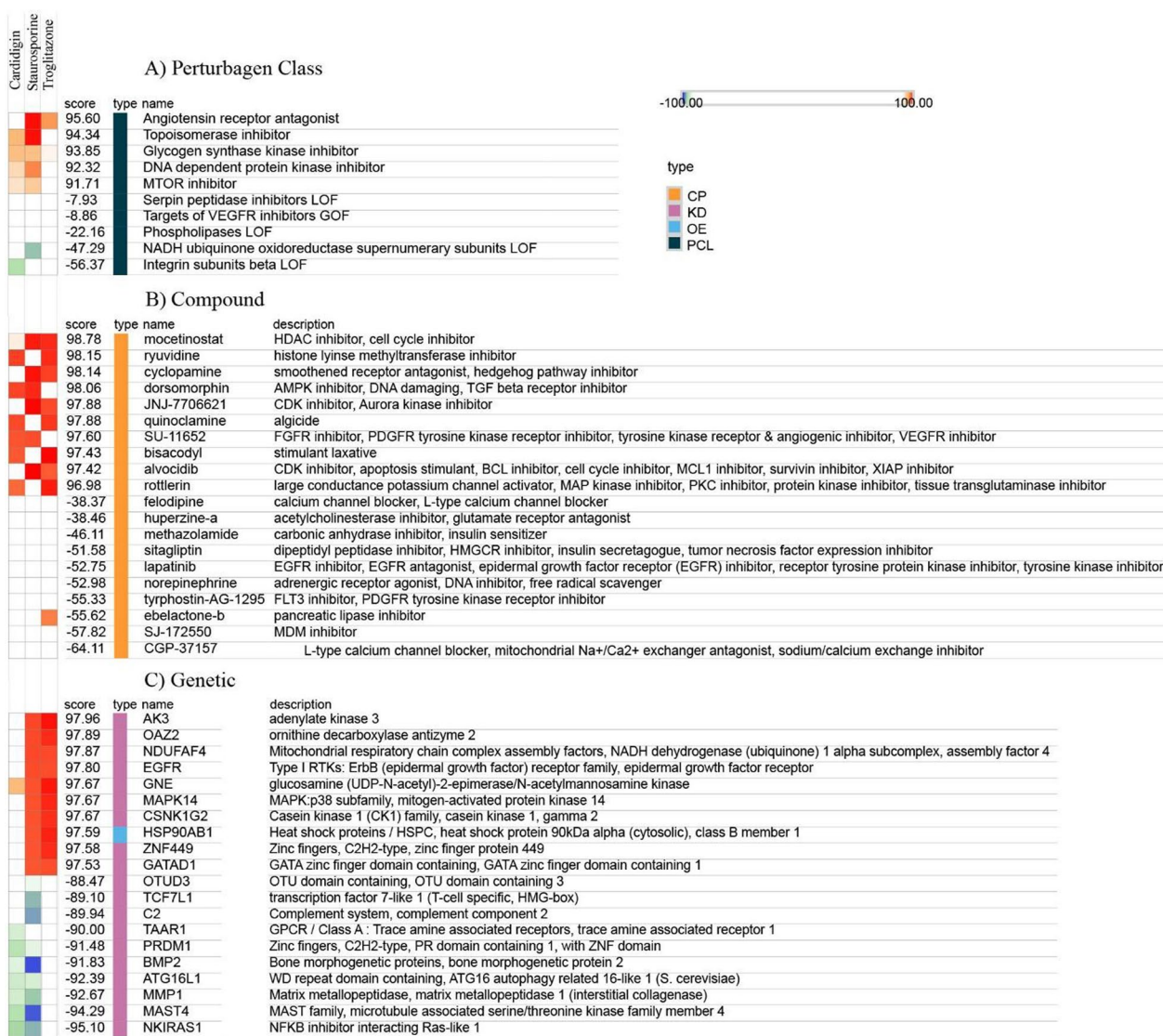


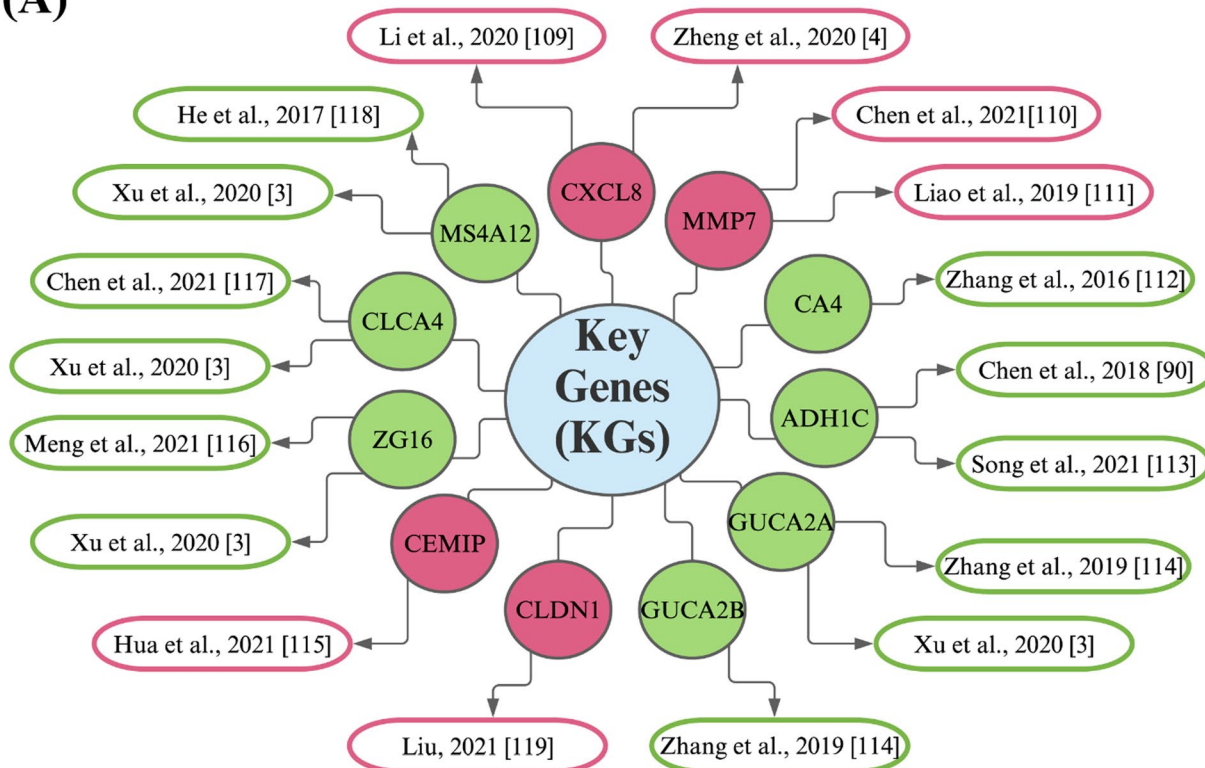
Fig. 10 The Connectivity Map for three small molecules of Troglitazone, Cardidigin and Staurosporine

from five transcriptomics profiles (GSE9348, GSE35279, GSE23878, GSE110224 and GSE50760) for diagnosis, prognosis, and therapies of CRC. At first, we identified 11 KGs (*CXCL8*, *MMP7*, *CA4*, *ADHIC*, *GUCA2A*, *GUCA2B*, *CEMIP*, *ZG16*, *CLCA4*, *MS4A12*, and *CLDN1*) by PPI analysis of 50 common DEGs (cDEGs). Some literature reviews also agreed with our results that these KGs are associated with CRC [14–45, 54] (see Fig. 11A). For example, Li et al. [110] reported that the gene *CXCL8* plays a vital role in CRC progression by mediating the differentiation, proliferation, and apoptosis within a regulatory network. So, they suggested this gene as a drug target for CRC also [110]. Chen and Ke [111] detected the gene *MMP7* as a potential biomarker

of CRC by bioinformatics analysis. Another study found that it regulates cancer progression and mediate the differentiation, proliferation, invasion and metastasis of various cancer cell types by different mechanisms [112]. A study reported that *CA4* is a newly identified tumor suppressor gene in CRC by targeting the WTAP–WT1–TBL1 axis through the inhibition of the *Wnt* signaling pathway [113]. The gene *ADHIC* might lead the increasing production of proinflammatory mediators by decreasing its expressions in the ulcerative colitis colon through the activation of the STAT1/NF-κB signaling pathway [114].

The abnormally expressed peptide hormones *GUCA2A* and *GUCA2B* play as paracrine endogenous ligands for

(A)



(B)

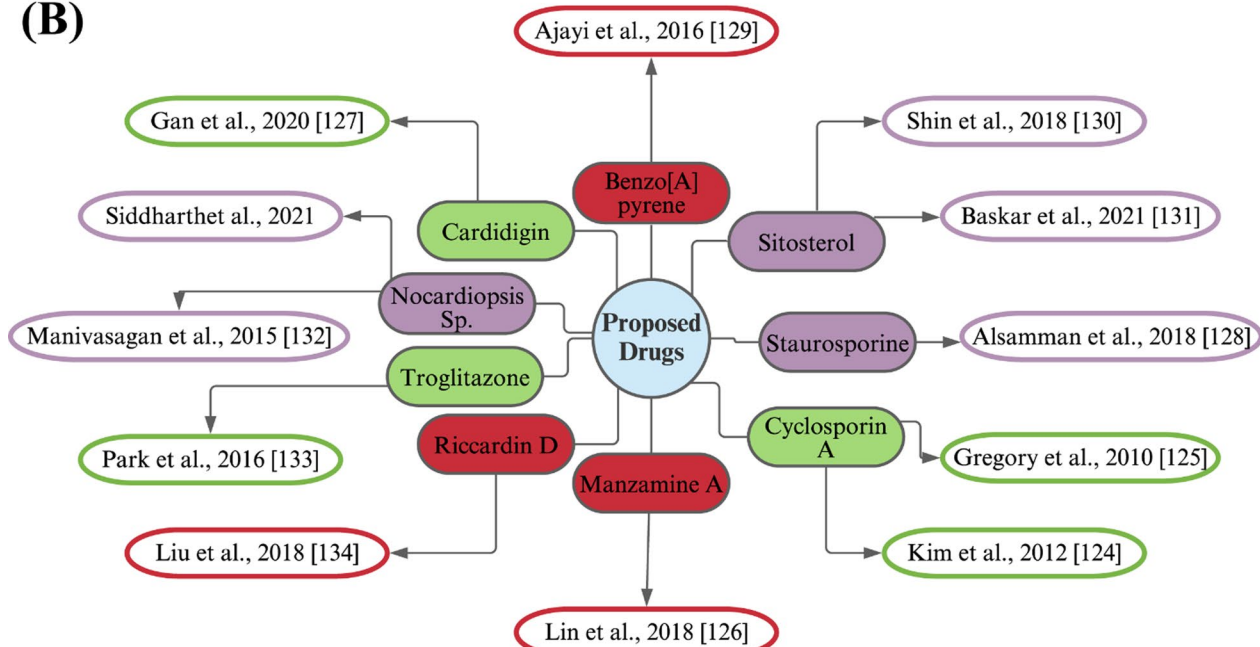


Fig. 11 Validation of the proposed KGs (receptors) and candidate drugs in favor of CRC by the literature review. **A** Validation of the proposed KGs: circles with green color indicate down-regulated KGs, and pink color indicates up-regulated KGs, and each connected network with a circle indicates the reference in which the cKG is associated with CRC, **B** validation of the proposed candidate drugs: circles with green color indicate FDA approved and investigational drugs, purple color indicate investigational drugs and red color indicate unapproved drugs and each connected network with a circle indicates the references in which our suggested drugs might be effective against CRC treatment

the guanylate cyclase-C (GUCY2C) receptor and help for the development of tumors in CRC by the association comparatively in lower levels with the disrupted intestinal homeostasis [115]. CEMIP is an adaptor protein of the O-GlcNAc transferase that can be reprogramming the glutamine metabolism through the reciprocal regulation of β -catenin and thereby promotes CRC metastasis. So, CRC metastasis can be prevented by the combinational inhibition of CEMIP and glutamine metabolism that would be a useful therapeutic strategy [116]. ZG16 can modulate the immune response in CRC by blocking the PD-L1 expression and the strong correlation designate ZG16 as a biomarker for the stratification of patient of immunotherapy [117]. Cancer-associated fibroblasts exosomes decreased the sensitivity of CRC towards the radiation and over-expressed miR-590-3p that promote CLCA4-dependent PI3K/Akt signaling pathway as well as cancer cell survival [118]. MS4A12 gene belonging to the MS4A family encodes a protein found in the apical membrane of colonocytes that plays an important role in the differentiation, proliferation, and cell cycle regulation and is believed to be a risk classification marker for early-stage colon cancer [119]. Upregulation of CLDN1 expression was observed in patients with colorectal cancer, which could be a possible biomarker for colorectal cancer treatment [120].

Moreover, different pertinent bioinformatic analyses based on independent databases significantly supported the relationship of KGs with the CRC progression as discussed below. The expression analysis with box-plots by GEPIA web-tool with TCGA RNA-seq data showed that KGs significantly separated CRC groups from the control groups (Fig. 5). We investigated the relationship of tumor infiltrates immune cells with the KGs and observed that KGs are significantly associated with different tumor infiltrates immune cells under different databases of CRC (Additional files 3 and 4). We investigated the DNA methylation of KGs at CpG sites by MethSurv web-tool with TCGA database and observed that seven KGs (*CEMIP*, *MMP7*, *CA4*, *GUCA2B*, *ZG16*, *CLCA4*, *MS4A12*) are significantly methylated at CpG sites (Table 5) that may play the vital role in CRC progression. To investigate the prognostic power of KGs, we performed multivariate survival analysis and developed a prediction model through RF classifiers in Fig. 7C. Our developed prediction models showed good performance with both training and test datasets generated from the main data collected from NCBI with accession numbers GSE9348, GSE23878, and GSE110224. The AUC values were 1.000 for the training dataset and 0.988 for the test dataset for RF model. To investigate their performance unbiasedly, we also considered independent test datasets from other NCBI sources with accession numbers GSE35279, respectively. We

observed that predictor show good performance with the independent test data and TCGA dataset. The values of AUC were 0.943 and 0.90 for independent test data and TCGA dataset based on RF model. These results indicate the good prediction performance for the identified KGs, so we suggested the prognostic model for the classifier (RF). The GO functional and KEGG pathway enrichment analyses of cDEGs significantly revealed some GO terms of BPs, MFs and CCs, and KEGG pathways by involving KGs that are highly linked with CRC patients (see Table 2). Our literature review also supported their link with CRC. As for examples with the enriched BPs, six GO terms *cell junction maintenance* [88], *cell-cell adhesion* [89], *calcium-independent cell-cell adhesion* [89], and *cell-cell junction organization* [89] these is associated with one KG (*CLDN1*). GO terms *negative regulation of cell population proliferation* [90] (associated with *CXCL8*), *regulation of cell migration* [91] (associated with *MMP7* and *CLDN1*) and (associated with *CA4*) were reported as important BPs for CRC progression. Among the enriched MFs, two GO terms *hormone activity* [92] and *receptor ligand activity* [94] were associated with *GUCA2A*. The *guanylate cyclase activator activity* [93] were associated with *GUCA2A*, *GUCA2B*. The MFs terms *chloride channel activity* [95] was associated with *CLCA4*. The *metallopeptidase activity* [96] was associated with *MMP7*. Among the enriched CCs, *anchored component of plasma membrane* [98] was associated with *CA4*. *collagen-containing extracellular matrix* [99] was associated with *ZG16*. Four CCs term *bicellular tight junction* [100] *apical junction complex* [101] and *cell-cell junction* [102] were associated with *CLCA1*. Among the enriched KEGG pathways, two KEGG terms *Nitrogen metabolism* [22] and *Proximal tubule bicarbonate reclamation* [103] were associated with *CA4*. Three pathways *Cell adhesion molecules* [104], *Pathogenic Escherichia coli infection* [105] and *Leukocyte transendothelial migration* [106] were associated with *CLDN1*. Two pathways *Pathogenic Escherichia coli infection* [105], *Amoebiasis* [44], and *Cytokine-cytokine receptor interaction* [107] were associated with *CXCL8*.

The KGs-TFs interaction network analysis indicated that 4 TFs proteins (*FOXC1*, *YY1*, *GATA2* and *NFKB1*) are the key transcriptional regulatory factors of KGs (see Fig. 4A). Among them, *FOXC1* (a regulator of *CA4*, *ADH1C*, *GUCA2A*, *GUCA2B*, *ZG16*, and *CLCA4*) is connected with lymphatic vessel formation, arterial cell specification, and cardiovascular development [121]. The expression of TF-protein *YY1* (a regulator of *CXCL8*, *ADH1C*, *CEMIP*, *ZG16*, and *CLCA4*) contributes to tumor growth differs in different cancers [122]. The TF-protein *GATA2* (a regulator of *GUCA2B*, *MMP7*, *CXCL8* and *MS4A12*) is connected with Hematopoietic and

immune defects [123]. The TF-protein NFKB1 (a regulator of *GUCA2A*, *CA4*, *CEMIP* and *MS4A12*) is a suppressor of inflammation, ageing and cancer [124]. We also constructed the proteins-disease interaction network to detect other diseases connected with the proposed target proteins. Total 9 target proteins out of 11 were associated with other 547 diseases that can be considered as the risk factors of CRC. Especially, two diseases, "Malignant tumor of colon" and "Colonic Neoplasms", were mostly related to our target proteins.

To explore our proposed KGs-guided new and repurposable candidate drugs for the treatment against CRC, we considered the proposed KGs based 11 key proteins (*CXCL8*, *MMP7*, *CA4*, *ADH1C*, *GUCA2A*, *GUCA2B*, *CEMIP*, *ZG16*, *CLCA4*, *MS4A12* and *CLDN1*) and their regulatory 4 TFs proteins (FOXC1, YY1, GATA2 and NFKB1) as the drug target receptors and performed their docking simulation with 167 drug molecules collected from the DSigDB database and published articles (Fig. 9A). Then we selected top-ranked 10 drugs (Cyclosporin A, Manzamine A, Cardidigin, Staurosporine, Benzo[A]Pyrene, Sitosterol, Nocardiosis Sp, Troglitazone, K-252a and Riccardin D) as the most probable repurposable candidate drugs for CRC patients based on their strong binding affinity scores (kCal/mol) with all the target proteins (Fig. 9A, B). Then we investigated the resistance performance of both the proposed and already published candidate drugs against the state-of-the-art alternatives of already published top-ranked 7 independent receptors for CRC and observed that our proposed candidate drugs are more effective compared to the already published drugs against the independent receptors also (Fig. 11B). We also tried to validate our proposed drugs in favor of CRC by the literature review (Fig. 11B).

Among the identified candidate drugs Cyclosporin A, a calcineurin inhibitor, traditionally used for its immunosuppressive effects, inhibits the activity of the non-canonical Wnt/Ca⁺⁺/NFAT signaling pathway [125, 126]. It has been reported that Manzamine A exhibits an anti-proliferative effect on human colorectal carcinoma cells and displays broad effects on gene expression to down-regulate fundamental maintenances of cell survival and induce apoptotic cell death and EMT inactivation [127]. This study demonstrates the efficacy of Cardidigin (digitoxin) against cervical cancer in vivo and elucidates its molecular mechanisms, including DSBs, cell cycle arrest and mitochondrial apoptosis. These results will contribute to the development of Cardidigin as a chemotherapeutic agent in the treatment of cervical cancer [128]. Staurosporine alleviates cisplatin chemoresistance in human cancer (colon) cell models by suppressing the induction of SQSTM1/p62 [129]. Ajayi et al. [130]

showed that Benzo[A] Pyrene induces oxidative stress, pro-inflammatory cytokines, expression of nuclear factor-kappa B, and deregulation of Wnt/ β -catenin signaling in colons of exposed mice. Sitosterol (Beta-sitosterol) suppresses tumor growth without toxicity in AGS xenograft mouse models and induces apoptosis in human gastric adenocarcinoma cells [131]. Sitosterol prevents lipid peroxidation and improves antioxidant status and histoarchitecture in rats with 1,2-dimethylhydrazine-induced colon cancer [132]. The marine actinobacterium *Nocardiosis* sp. MBRC-48 is an excellent microbial resource for the biosynthesis of gold nanoparticles with various biomedical applications such as antimicrobial, antioxidant, and anticancer activities [133]. The anti-proliferative and apoptotic activities of PDT in combination with the PPAR γ ligand troglitazone and provide a strong rationale for testing the therapeutic potential of combination treatment in colon cancer [134]. Liu et al. showed that Riccardin D might inhibit cell proliferation and induce apoptosis in HT-29 cells, which may be associated with the blocking of the NF- κ B signaling pathway [135]. Among the proposed nine candidate drugs, Cyclosporin A, Cardidigin, and Troglitazone are approved by the FDA for a different disease, the three other drugs (Staurosporine, Sitosterol and *Nocardiosis* Sp.) are still investigational, and the rest of the three drugs (Manzamine A, Benzo[A]Pyrene, and Riccardin D) are not yet approved. The approved drugs for different diseases and unapproved drugs should be further assessed at the molecular level by the wet-lab experiments prior to clinical investigation in the treatment of CRC.

Conclusion

The main purpose of this study was to identify key genomic biomarkers from multiple gene expression profiles for diagnosis, prognosis and therapies of CRC by using integrated bioinformatics and statistical approaches. We identified 11 common key genes (KGs) from multiple transcriptomics datasets, where 4 KGs (*CXCL8*, *CEMIP*, *MMP7*, and *CLDN1*) were up-regulated and the rest 7 KGs (*CA4*, *ADH1C*, *GUCA2A*, *GUCA2B*, *ZG16*, *CLCA4*, and *MS4A12*) were down-regulated. Different pertinent bioinformatic analyses including box plots of KGs-expressions with CRC and control groups, multivariate survival probability curves based on KGs-expressions, DNA methylation of KGs, correlation of KGs with immune infiltration levels in CRC, (different diseases)-KGs interaction, CRC-causing GO and KEGG pathways based on independent databases significantly supported the relationship of KGs with the CRC progression. Their association was also supported by several other independent studies directly or indirectly that we mentioned in the discussion

section. We detected four TFs proteins (FOXC1, YY1, GATA2 and NFKB) and eight microRNAs (hsa-mir-16-5p, hsa-mir-195-5p, hsa-mir-203a-3p, hsa-mir-34a-5p, hsa-mir-107, hsa-mir-27a-3p, hsa-mir-429, and hsa-mir-335-5p) as the key transcriptional and post-transcriptional regulators that may play a vital role in the regulation of KGs. Then we considered the proposed 11 key proteins and their regulatory 4 TFs-proteins as the drug target receptors to explore effective drug agents for CRC by molecular docking simulation with the 156 meta-drug agents. We detected nine small molecules (Cyclosporin A, Manzamine A, Cardidigin, Staurosporine, Benzo[A]Pyrene, Sitosterol, Nocardiosis Sp, Troglitazone, and Riccardin D) as the top-ranked candidate drugs for the treatment against CRC. Then we investigated the resistance performance of the proposed drugs against the state-of-the-art already published top-ranked 11 independent receptors for CRC and observed that our proposed repurposable candidate drugs are more effective compared to the already published drugs against the independent receptors also. Therefore, the proposed candidate drugs might be played a vital role in the treatment of CRC.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12920-023-01488-w>.

Additional file 1. Table S1. Different lists of hub-genes [HubGs] for colorectal cancer [CRC] published in different articles. **Table S2.** Association of cKGs with other disease risks.

Additional file 2. The details of common differentially expressed genes (cDEGs) for the five GEO datasets (GSE110224, GSE50760, GSE35279, and GSE23878) including Gene.symbol, ID, adj.P.Val, P.Value, t, B, logFC, and Gene.title.

Additional file 3. Correlation plots illustrating the relationship between gene expression and immune infiltration levels across multiple types of cancers, using the bioinformatics tool TIMER 2.0.

Additional file 4. Correlation between key genes (KGs) and immune infiltration levels in colorectal cancers (CRC).

Acknowledgements

We would like to acknowledge both editors and reviewers for their valuable comments and suggestions that help us to improve the quality of the manuscript.

Author contributions

MAH and MNHM conceived the idea of the study. MAH, MAI and MKK jointly analyzed gene expression profiles using different statistical and bioinformatics tools. MAH performed molecular docking and drafted the manuscript. MJA, SRK and MNHM edited the manuscript and provided important suggestions and MNHM supervised the project. All authors read and approved the final manuscript.

Funding

This work was completed without funding.

Availability of data and materials

The necessary R-codes and the datasets analyzed in this study is available at the web link: <https://github.com/Horaira29/crc.git>. (We collected these

datasets from the National Center of Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database with the accession numbers GSE9348, GSE110224, GSE23878, GSE35279 and GSE50760 using the web links. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE9348>, <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE110224>, <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE23878>, <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE35279>, and <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50760> respectively. We collected meta-drug agents from the online database DSigDB [56] with respect to the proposed receptors and FDA approved repurposed drugs for the treatment of NSCLC patients <https://www.cancer.gov/about-cancer/treatment/drugs/colorectal>.

Declarations

Ethics approval and consent to participate

All methods in the study were performed in accordance with the relevant guidelines.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Bioinformatics Lab, Department of Statistics, University of Rajshahi, Rajshahi 6205, Bangladesh. ²Department of Biochemistry and Molecular Biology, University of Rajshahi, Rajshahi 6205, Bangladesh.

Received: 12 April 2022 Accepted: 14 March 2023

Published online: 29 March 2023

References

1. WHO. Cancer, fact sheet, World Health Organization. Newsroom. 2018. <https://www.who.int/news-room/fact-sheets/detail/cancer>. Accessed 18 Mar 2021.
2. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. *CA Cancer J Clin*. 2015;65:87–108.
3. Arnold M, Sierra MS, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global patterns and trends in colorectal cancer incidence and mortality. *Gut*. 2017;66:683–91.
4. Kijima S, Sasaki T, Nagata K, Utano K, Lefor AT, Sugimoto H. Preoperative evaluation of colorectal cancer using CT colonography, MRI, and PET/CT. *World J Gastroenterol*. 2014;20:16964–75.
5. Brenner H, Kloor M, Pox CP. Colorectal cancer. In: *The Lancet*. Elsevier B.V.; 2014. p. 1490–502.
6. Ohhara Y, Fukuda N, Takeuchi S, Honma R, Shimizu Y, Kinoshita I, et al. Role of targeted therapy in metastatic colorectal cancer. *World J Gastrointest Oncol*. 2016;8:642.
7. Watanabe T, Muro K, Ajioka Y, Hashiguchi Y, Ito Y, Saito Y, et al. Japanese Society for Cancer of the Colon and Rectum (JSCCR) guidelines 2016 for the treatment of colorectal cancer. *Int J Clin Oncol*. 2018;23:1–34.
8. Santiago JA, Bottero V, Potashkin JA. Dissecting the molecular mechanisms of neurodegenerative diseases through network biology. *Front Aging Neurosci*. 2017;9:166.
9. Rahman MR, Islam T, Turanli B, Zaman T, Faruquee HM, Rahman MM, et al. Network-based approach to identify molecular signatures and therapeutic agents in Alzheimer's disease. *Comput Biol Chem*. 2019;78:431–9.
10. Islam T, Rahman R, Gov E, Turanli B, Gulfidan G, Haque A, et al. Drug targeting and biomarkers in head and neck cancers: insights from systems biology analyses. *Omi A J Integr Biol*. 2018;22:422–36.
11. Shahjaman M, Rezanur Rahman M, Shahinul Islam SM, Mollah MNH. A robust approach for identification of cancer biomarkers and candidate drugs. *Medicina*. 2019;55:269.
12. Moni MA, Islam MB, Rahman MR, Rashed-Al-Mahfuz M, Awal MA, Islam SMS, et al. Network-based computational approach to identify

- delineating common cell pathways influencing type 2 diabetes and diseases of bone and joints. *IEEE Access*. 2020;8:1486–97.
13. Mosharaf MP, Reza MS, Kibria MK, Ahmed FF, Kabir MH, Hasan S, et al. Computational identification of host genomic biomarkers highlighting their functions, pathways and regulators that influence SARS-CoV-2 infections and drug repurposing. *Sci Rep*. 2022;12:4279.
 14. Dong S, Ding Z, Zhang H, Chen Q. Identification of prognostic biomarkers and drugs targeting them in colon adenocarcinoma: a bioinformatic analysis. *Integr Cancer Ther*. 2019;18:1534735419864434.
 15. Rahman F, Mahmud P, Karim R, Hossain T, Islam F. Determination of novel biomarkers and pathways shared by colorectal cancer and endometrial cancer via comprehensive bioinformatics analysis. *Inform Med Unlocked*. 2020;20:100376.
 16. Dai GP, Wang LP, Wen YQ, Ren XQ, Zuo SG. Identification of key genes for predicting colorectal cancer prognosis by integrated bioinformatics analysis. *Oncol Lett*. 2020;19:388–98.
 17. Guo Y, Bao Y, Ma M, Yang W. Identification of key candidate genes and pathways in colorectal cancer by integrated bioinformatical analysis. *Int J Mol Sci*. 2017;18:722.
 18. Ding X, Duan H, Luo H. Identification of core gene expression signature and key pathways in colorectal cancer. *Front Genet*. 2020;11:45.
 19. Yuan Y, Chen J, Wang J, Xu M, Zhang Y, Sun P, et al. Identification hub genes in colorectal cancer by integrating weighted gene co-expression network analysis and clinical validation in vivo and vitro. *Front Oncol*. 2020;10:638.
 20. Liang B, Li C, Zhao J. Identification of key pathways and genes in colorectal cancer using bioinformatics analysis. *Med Oncol*. 2016;33:1–8.
 21. Tang L, Lei YY, Liu YJ, Tang B, Yang SM. The expression of seven key genes can predict distant metastasis of colorectal cancer to the liver or lung. *J Dig Dis*. 2020;21:639–49.
 22. Hozhabri H, Lashkari A, Razavi SM, Mohammadian A. Integration of gene expression data identifies key genes and pathways in colorectal cancer. *Med Oncol*. 2021;38:1–14.
 23. Berg KCG, Eide PW, Eilertsen IA, Johannessen B, Bruun J, Danielsen SA, et al. Multi-omics of 34 colorectal cancer cell lines—a resource for biomedical studies. *Mol Cancer*. 2017;16:1–16.
 24. Qi Y, Qi H, Liu Z, He P, Li B. Bioinformatics Analysis of key genes and pathways in colorectal cancer. *J Comput Biol*. 2019;26:364–75.
 25. Chen Z, Lin Y, Gao J, Lin S, Zheng Y, Liu Y, et al. Identification of key candidate genes for colorectal cancer by bioinformatics analysis. *Oncol Lett*. 2019;18:6583–93.
 26. Xu H, Ma Y, Zhang J, Gu J, Jing X, Lu S, et al. Identification and verification of core genes in colorectal cancer. *Biomed Res Int*. 2020;2020.
 27. Mastrogamvraki N, Zaravinos A. Signatures of co-deregulated genes and their transcriptional regulators in colorectal cancer. *Npj Syst Biol Appl*. 2020;6:23.
 28. Liu X, Bing Z, Wu J, Zhang J, Zhou W, Ni M, et al. Integrative gene expression profiling analysis to investigate potential prognostic biomarkers for colorectal cancer. *Med Sci Monit*. 2020;26:e918906–11.
 29. Pirim D. Integrative analyses of molecular pathways and key candidate biomarkers associated with colorectal cancer. *Cancer Biomark*. 2020;27:555–68.
 30. Cui X, Shen K, Xie Z, Liu T, Zhang H. Identification of key genes in colorectal cancer using random walk with restart. *Mol Med Rep*. 2017;15:867–72.
 31. Zhu H, Ji Y, Li W, Wu M. Identification of key pathways and genes in colorectal cancer to predict the prognosis based on mRNA interaction network. *Oncol Lett*. 2019;18:3778–86.
 32. Peng WF, Bai F, Shao K, Shen LS, Li HH, Huang S. The key genes underlying pathophysiology association between the type 2-diabetic and colorectal cancer. *J Cell Physiol*. 2018;233:8551–7.
 33. Liu S, Zeng F, Fan G, Dong Q. Identification of hub genes and construction of a transcriptional regulatory network associated with tumor recurrence in colorectal cancer by weighted gene co-expression network analysis. *Front Genet*. 2021;12:649752.
 34. Asghari M, Abazari MF, Bokharai H, Aleagha MN, Poortahmasebi V, Askari H, et al. Key genes and regulatory networks involved in the initiation, progression and invasion of colorectal cancer. *Futur Sci OA*. 2018;4:FSO278.
 35. Lin T, Liang C, Peng W, Qiu Y, Peng L. Mechanisms of core Chinese herbs against colorectal cancer: A study based on data mining and network pharmacology. *Evidence-based Complement Altern Med*. 2020;2020.
 36. Kasap E, Gerceker E, Boyacioglu SÖ, Yuceyar H, Yildirim H, Ayhan S, et al. The potential role of the NEK6, AURKA, AURKB, and PAK1 genes in adenomatous colorectal polyps and colorectal adenocarcinoma. *Tumor Biol*. 2016;37:3071–80.
 37. Zheng Z, Xie J, Xiong L, Gao M, Qin L, Dai C, et al. Identification of candidate biomarkers and therapeutic drugs of colorectal cancer by integrated bioinformatics analysis. *Med Oncol*. 2020;37:1–11.
 38. Hameed Y, Usman M, Liang S, Ejaz S. Novel diagnostic and prognostic biomarkers of colorectal cancer: capable to overcome the heterogeneity-specific barrier and valid for global applications. *PLoS ONE*. 2021;16:e0256020.
 39. Yang X, Wei W, Tan S, Guo L, Qiao S, Yao B, et al. Identification and verification of HCAR3 and INSL5 as new potential therapeutic targets of colorectal cancer. *World J Surg Oncol*. 2021;19:248.
 40. Zhang J, Zhang H, Li F, Song Z, Li Y, Zhao T. Identification of intestinal flora-related key genes and therapeutic drugs in colorectal cancer. *BMC Med Genom*. 2020;13:1–8.
 41. Rahman MR, Islam T, Gov E, Turanli B, Gulfidan G, Shahjaman M, et al. Identification of prognostic biomarker signatures and candidate drugs in colorectal cancer: insights from systems biology analysis. *Medicina*. 2019;55:20.
 42. Chen J, Wang Z, Shen X, Cui X, Guo Y. Identification of novel biomarkers and small molecule drugs in human colorectal cancer by microarray and bioinformatics analysis. *Mol Genet Genomic Med*. 2019;7:e00713.
 43. Yu C, Chen F, Jiang J, Zhang H, Zhou M. Screening key genes and signaling pathways in colorectal cancer by integrated bioinformatics analysis. *Mol Med Rep*. 2019;20:1259–69.
 44. Yang J, Gao S, Qiu M, Kan S. Integrated analysis of gene expression and metabolite data reveals candidate molecular markers in colorectal carcinoma. *Cancer Biother Radiopharm*. 2020. <https://doi.org/10.1089/cbr.2020.3980>.
 45. Zhao Z, Fan X, Yang L, Song J, Fang S, Tu J, et al. The identification of a common different gene expression signature in patients with colorectal cancer. *Math Biosci Eng*. 2019;16:2942–58.
 46. Ma Y, Wen J, Wang J, Wang C, Zhang Y, Zhao L, et al. Asiaticoside antagonizes proliferation and chemotherapeutic drug resistance in hepatocellular carcinoma (HCC) cells. *Med Sci Monit*. 2020;26:e924435-1.
 47. Leng X, Yang J, Liu T, Zhao C, Cao Z, Li C, et al. A bioinformatics framework to identify the biomarkers and potential drugs for the treatment of colorectal cancer. *Front Genet*. 2022;13:2739.
 48. Wang Q, Huang X, Zhou S, Ding Y, Wang H, Jiang W, et al. IL1RN and PRRX1 as a prognostic biomarker correlated with immune infiltrates in colorectal cancer: evidence from bioinformatic analysis. *Int J Genom*. 2022;2022.
 49. Sharma A, Yadav D, Rao P, Sinha S, Goswami D, Rawal RM, et al. Identification of potential therapeutic targets associated with diagnosis and prognosis of colorectal cancer patients based on integrated bioinformatics analysis. *Comput Biol Med*. 2022;146:105688.
 50. Ekanem TI, Tsai WL, Lin YH, Tan WQ, Chang HY, Huang TC, et al. Identification of the effects of aspirin and sulindac sulfide on the inhibition of HMG2-mediated oncogenic capacities in colorectal cancer. *Molecules*. 2020;25:3826.
 51. Li B, Flaveny CA, Giambelli C, Fei DL, Han L, Hang BI, Bai F, Pei XH, Nose V, Burlingame O, Capobianco AJ. Repurposing the FDA-approved pinworm drug pyvium as a novel chemotherapeutic agent for intestinal polyposis. *PLoS ONE*. 2014;9:e101969.
 52. Sun Z, Liu C, Cheng SY. Identification of four novel prognosis biomarkers and potential therapeutic drugs for human colorectal cancer by bioinformatics analysis. *J Biomed Res*. 2020;1:1–15.
 53. Zheng Y, Zhou J, Tong Y. Gene signatures of drug resistance predict patient survival in colorectal cancer. *Pharmacogenomics J*. 2015;15:135–43.
 54. Que W, Chen M, Yang L, Zhang B, Zhao Z, Liu M, et al. A network pharmacology-based investigation on the bioactive ingredients and

- molecular mechanisms of *Gelsemium elegans Benth* against colorectal cancer. *BMC Complement Med Ther.* 2021;21:1–18.
55. Zhang M, Wang D, Lu F, Zhao R, Ye X, He L, et al. Identification of the active substances and mechanisms of ginger for the treatment of colon cancer based on network pharmacology and molecular docking. *BioData Min.* 2020. <https://doi.org/10.21203/rs.3.rs-38111/v1>.
 56. Huang S, Zhang Z, Li W, Kong F, Yi P, Huang J, et al. Network pharmacology-based prediction and verification of the active ingredients and potential targets of Zuojinwan for treating colorectal cancer. *Drug Des Dev Ther.* 2020;14:2725–40.
 57. Yang F, Cai S, Ling L, Zhang H, Tao L, Wang Q. Identification of a five-gene prognostic model and its potential drug repurposing in colorectal cancer based on TCGA, GTEx and GEO databases. *Front Genet.* 2021;11:622659.
 58. L L, Y C, X L, M W, J L, Q X, et al. Upregulation of SNTB1 correlates with poor prognosis and promotes cell growth in colorectal cancer. 2021. <https://doi.org/10.21203/RS.3.RS-549083/V1>.
 59. Mokgautsi N, Wang Y-C, Lawal B, Khedkar H, Sumitra MR, Wu ATH, et al. Network pharmacological analysis through a bioinformatics approach of novel NSC765600 and NSC765691 compounds as potential inhibitors of CCND1/CDK4/PLK1/CD44 in cancer types. *Cancers (Basel).* 2021;13:2523.
 60. Yoo M, Shin J, Kim J, Ryall KA, Lee K, Lee S, et al. DSigDB: drug signatures database for gene set analysis. *Bioinformatics.* 2015;31:3069–71.
 61. Smith GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol.* 2004;3:1–25.
 62. Ritchie ME, Phipson B, Wu DI, Hu Y, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015;43(7):e47.
 63. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 2019;47:D607–13.
 64. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003;13:2498–504.
 65. Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY. cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst Biol.* 2014;8:1–7.
 66. Jeong H, Mason SP, Barabási AL, Oltvai ZN. Lethality and centrality in protein networks. *Nature.* 2001;411:41–2.
 67. Pržulj N, Wagle DA, Jurisica I. Functional topology in a network of protein interactions. *Bioinformatics.* 2004;20:340–8.
 68. Freeman LC. A set of measures of centrality based on betweenness. *Sociometry.* 1977;40:35–41.
 69. Shimmel A. Structural parameters of communication networks. *Bull Math Biophys.* 1953;15:501–7.
 70. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* 2017;45:W98–102.
 71. Li T, Fu J, Zeng Z, Cohen D, Li J, Chen Q, et al. TIMER2.0 for analysis of tumor-infiltrating immune cells. *Nucleic Acids Res.* 2020;48:W509–14.
 72. Modhukur V, Iljasenko T, Metsalu T, Lökk K, Laisk-Podar T, Vilo J. MethSurv: a web tool to perform multivariable survival analysis using DNA methylation data. *Epigenomics.* 2018;10:277–88.
 73. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 2016;44:W90–7.
 74. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* 2020;48:D845–55.
 75. Aguirre-Gamboa R, Martí 'ñez-Ledesma E, Martí 'ñez-Torteya A, Chacolla-Huaringa R, et al. G-RH. SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis. 2018.
 76. Doms A, Schroeder M. GoPubMed: exploring PubMed with the gene ontology. *Nucleic Acids Res.* 2005;33(SUPPL):2.
 77. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000;28:27–30.
 78. Xie Z, Bailey A, Kuleshov MV, Clarke DJB, Evangelista JE, Jenkins SL, et al. Gene set knowledge discovery with Enrichr. *Curr Protoc.* 2021;1:e90.
 79. Fornes O, Castro-Mondragon JA, Khan A, Van Der Lee R, Zhang X, Richmond PA, JASPAR, et al. Update of the open-Access database of transcription factor binding profiles. *Nucleic Acids Res.* 2020;2020:48.
 80. Zhou G, Soufan O, Ewald J, Hancock REW, Basu N, Xia J. NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. *Nucleic Acids Res.* 2019;47:W234–41.
 81. Chou CH, Shrestha S, Yang CD, Chang NW, Lin YL, Liao KW, MiRTarBase update, et al. A resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.* 2018;2018:46.
 82. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, et al. The protein data bank. *Acta Crystallogr Sect D Biol Crystallogr.* 2002;58(6):899–907.
 83. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* 2019;47:D1102–9.
 84. Systèmes D. BIOVIA, discovery studio visualizer, release 2019. San Diego Dassault Systèmes. 2020.
 85. Gruber A, Durham AM, Huynh C, del Portillo HA. Defining and searching for structural motifs using DeepView/Swiss-PdbViewer. *BMC Bioinform.* 2008;13:1–11.
 86. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multi-threading. *J Comput Chem.* 2009. <https://doi.org/10.1002/jcc.21334>.
 87. Dallakyan S, Olson AJ. Small-molecule library screening by docking with PyRx. *Methods Mol Biol.* 2015;1263:243–50.
 88. Ibáñez Gaspar V, Catozzi S, Ternet C, Luthert PJ, Kiel C. Analysis of Ras-effector interaction competition in large intestine and colorectal cancer context. *Small GTPases.* 2021;12:209–25.
 89. Faux MC, King LE, Kane SR, Love C, Sieber OM, Burgess AW. APC regulation of ESRP1 and p120-catenin isoforms in colorectal cancer cells. *Mol Biol Cell.* 2021;32:120–30.
 90. Fan Q, Liu B. Identification of the anticancer effects of a novel proteasome inhibitor, ixazomib, on colorectal cancer using a combined method of microarray and bioinformatics analysis. *Oncotargets Ther.* 2017;10:3591–606.
 91. Gao Y, Zhang S, Zhang Y, Qian J. Identification of microRNA-target gene-transcription factor regulatory networks in colorectal adenoma using microarray expression data. *Front Genet.* 2020;11:463.
 92. Lv J, Li L, Duan B. Hub genes and key pathway identification in colorectal cancer based on bioinformatic analysis. *Biomed Res Int.* 2019;2019.
 93. Ai D, Wang Y, Li X, Pan H. Colorectal cancer prediction based on weighted gene co-expression network analysis and variational auto-encoder. *Biomolecules.* 2020;10:1–11.
 94. Xu K, He J, Zhang J, Liu T, Yang F, Ren T. A novel prognostic risk score model based on immune-related genes in patients with stage IV colorectal cancer. *Biosci Rep.* 2020;40.
 95. Chen TJ, He HL, Shiu YL, Yang CC, Lin LC, Tian YF, et al. High chloride channel accessory 1 expression predicts poor prognoses in patients with rectal cancer receiving chemoradiotherapy. *Int J Med Sci.* 2018;15:1171–8.
 96. Wang W, Sun JF, Wang XZ, Ying HQ, You XH, Sun F. A novel prognostic score based on zg16 for predicting crc survival. *Pharmgenomics Pers Med.* 2020;13:735–47.
 97. Zheng W, Yang C, Qiu L, Feng X, Sun K, Deng H. Transcriptional information underlying the generation of CSCs and the construction of a nine-mRNA signature to improve prognosis prediction in colorectal cancer. *Cancer Biol Ther.* 2020;21:688–97.
 98. Sugiyama Y, Wakazaki M, Toyooka K, Fukuda H, Oda Y. A Novel plasma membrane-anchored protein regulates xylem cell-wall deposition through microtubule-dependent lateral inhibition of rho GTPase domains. *Curr Biol.* 2017;27:2522–2528.e4.
 99. Chu XD, Zhang YR, Lin ZB, Zhao Z, Huangfu SC, Qiu SH, et al. A network pharmacology approach for investigating the multitarget mechanisms of Huangqi in the treatment of colorectal cancer. *Transl Cancer Res.* 2021;10:681–93.
 100. Angius A, Uva P, Pira G, Muroli MR, Sotgiu G, Saderi L, et al. Integrated analysis of miRNA and mRNA endorses a twenty miRNAs signature for colorectal carcinoma. *Int J Mol Sci.* 2019;20:4067.

101. Gehren AS, Rocha MR, de Souza WF, Morgado-Díaz JA. Alterations of the apical junctional complex and actin cytoskeleton and their role in colorectal cancer progression. *Tissue Barriers*. 2015;3:1–12.
102. Kim WK, Kwon Y, Jang M, Park M, Kim J, Cho S, et al. **B**-catenin activation down-regulates cell–cell junction-related genes and induces epithelial-to-mesenchymal transition in colorectal cancers. *Sci Rep*. 2019;9:1–15.
103. Patil AR, Leung MY, Roy S. Identification of hub genes in different stages of colorectal cancer through an integrated bioinformatics approach. *Int J Environ Res Public Health*. 2021;18:5564.
104. Codrich M, Dalla E, Mio C, Antoniali G, Malfatti MC, Marzinotto S, et al. Integrated multi-omics analyses on patient-derived CRC organoids highlight altered molecular pathways in colorectal cancer progression involving PTEN. *J Exp Clin Cancer Res*. 2021;40:1–17.
105. Veziat J, Gagnière J, Jouberton E, Bonnin V, Sauvanet P, Pezet D, et al. Association of colorectal cancer with pathogenic *Escherichia coli*: focus on mechanisms using optical imaging. *World J Clin Oncol*. 2016;7:293–301.
106. Huang C, Ou R, Chen X, Zhang Y, Li J, Liang Y, et al. Tumor cell-derived SPON2 promotes M2-polarized tumor-associated macrophage infiltration and cancer progression by activating PYK2 in CRC. *J Exp Clin Cancer Res*. 2021;40:1–17.
107. Singh MP, Rai S, Singh NK, Srivastava S. Transcriptomic landscape of early age onset of colorectal cancer identifies novel genes and pathways in Indian CRC patients. *Sci Rep*. 2021;11:1–11.
108. Wu D, Wang X. Application of clinical bioinformatics in lung cancer-specific biomarkers. *Cancer Metastasis Rev*. 2015;34:209–16.
109. Liang Y, Diehn M, Watson N, Bollen AW, Aldape KD, Nicholas MK, et al. Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme. *Proc Natl Acad Sci U S A*. 2005;102:5814–9.
110. Li J, Liu Q, Huang X, Cai Y, Song L, Xie Q, et al. Transcriptional profiling reveals the regulatory role of CXCL8 in promoting colorectal cancer. *Front Genet*. 2020;10:1360.
111. Chen L, Ke X. MMP7 as a potential biomarker of colon cancer and its prognostic value by bioinformatics analysis. *Medicine (Baltimore)*. 2021;100:e24953.
112. Liao HY, Da CM, Liao B, Zhang HH. Roles of matrix metalloproteinase-7 (MMP-7) in cancer. *Clin Biochem*. 2021;92:9–18.
113. Zhang J, Tsoi H, Li X, Wang H, Gao J, Wang K, et al. Carbonic anhydrase IV inhibits colon cancer development by inhibiting the Wnt signalling pathway through targeting the WTAP-WT1-TBL1 axis. *Gut*. 2016;65:1482–93.
114. Song F, Zhang Y, Pan Z, Hu X, Zhang Q, Huang F, et al. The role of alcohol dehydrogenase 1C in regulating inflammatory responses in ulcerative colitis. *Biochem Pharmacol*. 2021;192:114691.
115. Zhang H, Du Y, Wang Z, Lou R, Wu J, Feng J. Integrated analysis of oncogenic networks in colorectal cancer identifies GUCA2A as a Molecular marker. *Biochem Res Int*. 2019;2019.
116. Hua Q, Zhang B, Xu G, Wang L, Wang H, Lin Z, et al. CEMIP, a novel adaptor protein of OGT, promotes colorectal cancer metastasis through glutamine metabolic reprogramming via reciprocal regulation of β -catenin. *Oncogene*. 2021;2021:1–13.
117. Meng H, Ding Y, Liu E, Li W, Wang L. ZG16 regulates PD-L1 expression and promotes local immunity in colon cancer. *Transl Oncol*. 2021;14:101003.
118. Chen X, Liu Y, Zhang Q, Liu B, Cheng Y, Zhang Y, et al. Exosomal miR-590-3p derived from cancer-associated fibroblasts confers radioresistance in colorectal cancer. *Mol Ther Nucleic Acids*. 2021;24:113–26.
119. He L, Deng HY, Wang XC. Decreased expression of MS4A12 inhibits differentiation and predicts early stage survival in colon cancer. *Neoplasma*. 2017;64:65–73.
120. Liu W. Long non-coding RNA VPS9D1-AS1 promotes growth of colon adenocarcinoma by sponging miR-1301-3p and CLDN1. *Hum Cell*. 2021;34:1775–87.
121. Kume T. The cooperative roles of Foxcl and Foxc2 in cardiovascular development. *Adv Exp Med Biol*. 2009;665:63–77.
122. Sarvagalla S, Kolapalli SP, Vallabhapurapu S. The two sides of YY1 in cancer: a friend and a foe. *Front Oncol*. 2019;9:1230.
123. Collin M, Dickinson R, Bigley V. Haematopoietic and immune defects associated with GATA2 mutation. *Br J Haematol*. 2015;169:173–87.
124. Cartwright T, Perkins ND, L. Wilson C. NFKB1: a suppressor of inflammation, ageing and cancer. *FEBS J*. 2016;283:1812–22.
125. Kim J, Kim DW, Chang W, Choe J, Kim J, Park C-S, et al. Wnt5a is secreted by follicular dendritic cells to protect germinal center B cells via Wnt/Ca²⁺/NFAT/NF- κ B Cell Lymphoma 6 signaling. *J Immunol*. 2012;188:182–9.
126. Gregory MA, Phang TL, Neviani P, Alvarez-Calderon F, Eide CA, O'Hare T, et al. Wnt/Ca²⁺/NFAT signaling maintains survival of Ph⁺ leukemia cells upon inhibition of Bcr-Abl. *Cancer Cell*. 2010;18:74–87.
127. Lin LC, Kuo TT, Chang HY, Liu WS, Hsia SM, Huang TC. Manzamine a exerts anticancer activity against human colorectal cancer cells. *Mar Drugs*. 2018;16:252.
128. Gan H, Qi M, Chan C, Leung P, Ye G, Lei Y, et al. Digitoxin inhibits HeLa cell growth through the induction of G2/M cell cycle arrest and apoptosis in vitro and in vivo. *Int J Oncol*. 2020;57:562–73.
129. Alsamman K, El-Masry OS. Staurosporine alleviates cisplatin chemoresistance in human cancer cell models by suppressing the induction of SQSTM1/p62. *Oncol Rep*. 2018;40:2157–62.
130. Ajayi BO, Adedara IA, Farombi EO. Benzo(a)pyrene induces oxidative stress, pro-inflammatory cytokines, expression of nuclear factor-kappa B and deregulation of wnt/beta-catenin signaling in colons of BALB/c mice. *Food Chem Toxicol*. 2016;95:42–51.
131. Shin EJ, Choi HK, Sung MJ, Park JH, Chung MY, Chung S, et al. Antitumour effects of beta-sitosterol are mediated by AMPK/PTEN/HSP90 axis in AGS human gastric adenocarcinoma cells and xenograft mouse models. *Biochem Pharmacol*. 2018;152:60–70.
132. Baskar AA, AlNumair KS, GabrielPaulraj M, Alsaif MA, Muamar MA, Ignacimuthu S. **B**-sitosterol prevents lipid peroxidation and improves antioxidant status and histoarchitecture in rats with 1,2-dimethylhydrazine-induced colon cancer. *J Med Food*. 2012;15:335–43.
133. Manivasagan P, Alam MS, Kang KH, Kwak M, Kim SK. Extracellular synthesis of gold bionanoparticles by *Nocardiaopsis* sp. and evaluation of its antimicrobial, antioxidant and cytotoxic activities. *Bioprocess Biosyst Eng*. 2015;38:1167–77.
134. Park H, Ko SH, Lee JM, Park JH, Choi YH. Troglitazone enhances the apoptotic response of DLD-1 colon cancer cells to photodynamic therapy. *Yonsei Med J*. 2016;57:1494–9.
135. Liu H, Li G, Zhang B, Sun D, Wu J, Chen F, et al. Suppression of the NF- κ B signaling pathway in colon cancer cells by the natural compound Riccardin D from *Dumortierahirsute*. *Mol Med Rep*. 2018;17:5837–43.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

