Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

CRCDB: A comprehensive database for integrating and analyzing multi-omics data of early-onset and late-onset colorectal cancer

Danyi Zou [a,d,e,1], Wanshan Ning [a,d,e,f,1], Luming Xu [b,d,e,1], Shijun Lei [a,d,e], Lin Wang [a,d,e,*], Zheng Wang [b,c,d,e,*]

[a] *Department of Clinical Laboratory, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China*
[b] *Research Center for Tissue Engineering and Regenerative Medicine, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China*
[c] *Department of Gastrointestinal Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China*
[d] *Hubei Key Laboratory of Regenerative Medicine and Multi-disciplinary Translational Research, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China*
[e] *Hubei Provincial Engineering Research Center of Clinical Laboratory and Active Health Smart Equipment, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China*
[f] *Institute for Clinical Medical Research, the First Affiliated Hospital of Xiamen University, School of Medicine, Xiamen University, Xiamen, Fujian 361003, China*

ARTICLE INFO

Keywords:
Colorectal cancer
Early-onset colorectal cancer
Data mining platform
Multi-omics data

ABSTRACT

The incidence of early-onset colorectal cancer (EOCRC) has increased significantly worldwide. Uncovering biomarkers that are unique to EOCRC is of great importance to facilitate the prevention and detection of this growing cancer subtype. Although efforts have been made in the data curation about CRC, there is no integrated platform that gives access to data specifically related to young CRC patients. Here, we constructed a user-friendly open integrated resource called CRCDB (URL: http://crcdb-hust.com) which contains multi-omics data of 785 EOCRC, 4898 late-onset CRCs (LOCRC), and 1110 normal control samples from tissue, whole blood, platelets, and serum exosomes. CRCDB manages the differential analysis, survival analysis, co-expression analysis, and immune cell infiltration comparison analysis results in different CRC groups. Meta-analysis results were also provided for users for further data interpretation. Using the resource in CRCDB, we identified that genes associated with the metabolic process were less expressed in EOCRC patients, while up regulated genes most associated with the mitosis process might play an important role in the molecular pathogenesis of LOCRC. Survival-related genes were most enriched in oxidoreduction pathways in EOCRC while in immune-related pathways in LOCRC. With all the data gathered and processed, we anticipate that CRCDB could be a practical data mining platform to help explore potential applications of omics data and develop effective prevention and therapeutic strategies for the specific group of CRC patients.

1. Introduction

Colorectal cancer (CRC) is the third most common malignancy and the second leading cause of cancer-related death worldwide [1]. Although CRC is previously considered a disease primarily affecting older individuals, there has been a global increase in early-onset CRC (EOCRC) patients, defined as patients under 50 years old, over the past few decades [2]. Given that more than 86 % of EOCRC are diagnosed at later stages with worse outcomes which causes a large burden of disease among young adults [3], uncovering biomarkers unique to EOCRC is of great importance to facilitate the prevention and detection of this growing cancer subtype.

With more researches focusing on EOCRC, it has been revealed that EOCRC exhibits distinct molecular characteristics compared to late-onset CRC (LOCRC; age >= 50 years old). For instance, EOCRC patients were found to have higher rates of mutations in genes related to cancer-predisposing syndromes, such as MSH2 and MSH6; the prevalence of MSI-high tumors was found to be twice higher in EOCRC

* Corresponding authors at: Hubei Key Laboratory of Regenerative Medicine and Multi-disciplinary Translational Research, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China.

E-mail addresses: lin_wang@hust.edu.cn (L. Wang), zhengwang@hust.edu.cn (Z. Wang).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.csbj.2024.05.051
Received 18 March 2024; Received in revised form 30 May 2024; Accepted 31 May 2024
Available online 2 June 2024
2001-0370/© 2024 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

compared to LOCRC [4,5]. Besides, EOCRC patients may have stronger immune reactions and a lower prevalence of activation of WNT and MYC signaling pathways and epithelial-mesenchymal transition compared to CRC patients aged 50–69 years old [6,7]. Thus, the current identified biomarkers or used disease prevention strategies for overall CRC might not apply to EOCRC. As for disease treatment, since long-term treatment-related complications should be considered, we need to find more appropriate ways for these younger patients than traditional therapies [8]. Therefore, further elucidating their characteristics, especially at the molecular level, may provide important insights into the underlying disease processes, which might ultimately help to develop effective prevention, early detection, and therapeutic strategies for the specific group of patients.

Until now, some efforts have been made in the data curation about CRC [9,10] and several databases have been constructed. Colorectal Cancer Atlas was established as a resource for the genomic and proteomic annotations identified in CRC tissues and cell lines [11]. CBD collected 870 CRC biomarkers from articles in PubMed published from 1986 to 2017 and provided users with data for further statistics and bioinformatics analyses [12]. These databases provide valuable information for researchers; however, no integrated platform gives access to data specifically related to EOCRC.

In the current study, we carefully curated the public datasets from four kinds of data resources (tissue, whole blood, platelets, and serum exosomes) that cover 6793 clinical samples in EOCRC, LOCRC, and normal samples to depict the differential omics landscapes of specific groups of CRC patients. We visualized all the data and analysis results in an open integrated platform CRCDB (http://crcdb-hust.com). Users can compare the gene expression, methylation level, copy number variation (CNV), and immune infiltration status in different groups of CRC patients and explore the survival outcome-related genes in four kinds of data resources. We believe that CRCDB will serve as a useful platform for the understanding of tumor progression and may have further practical value in CRC clinical treatment.

2. Materials and methods

2.1. Data extraction and preprocessing

TCGA CRC data was downloaded from UCSC Xena (https://xena.ucsc.edu/) and other datasets were downloaded from the GEO database (https://www.ncbi.nlm.nih.gov/geo/)(Table S1). For the GEO database, we only obtained the datasets with age information and with sample size > 20 (EOCRC > 3). All the samples were divided into

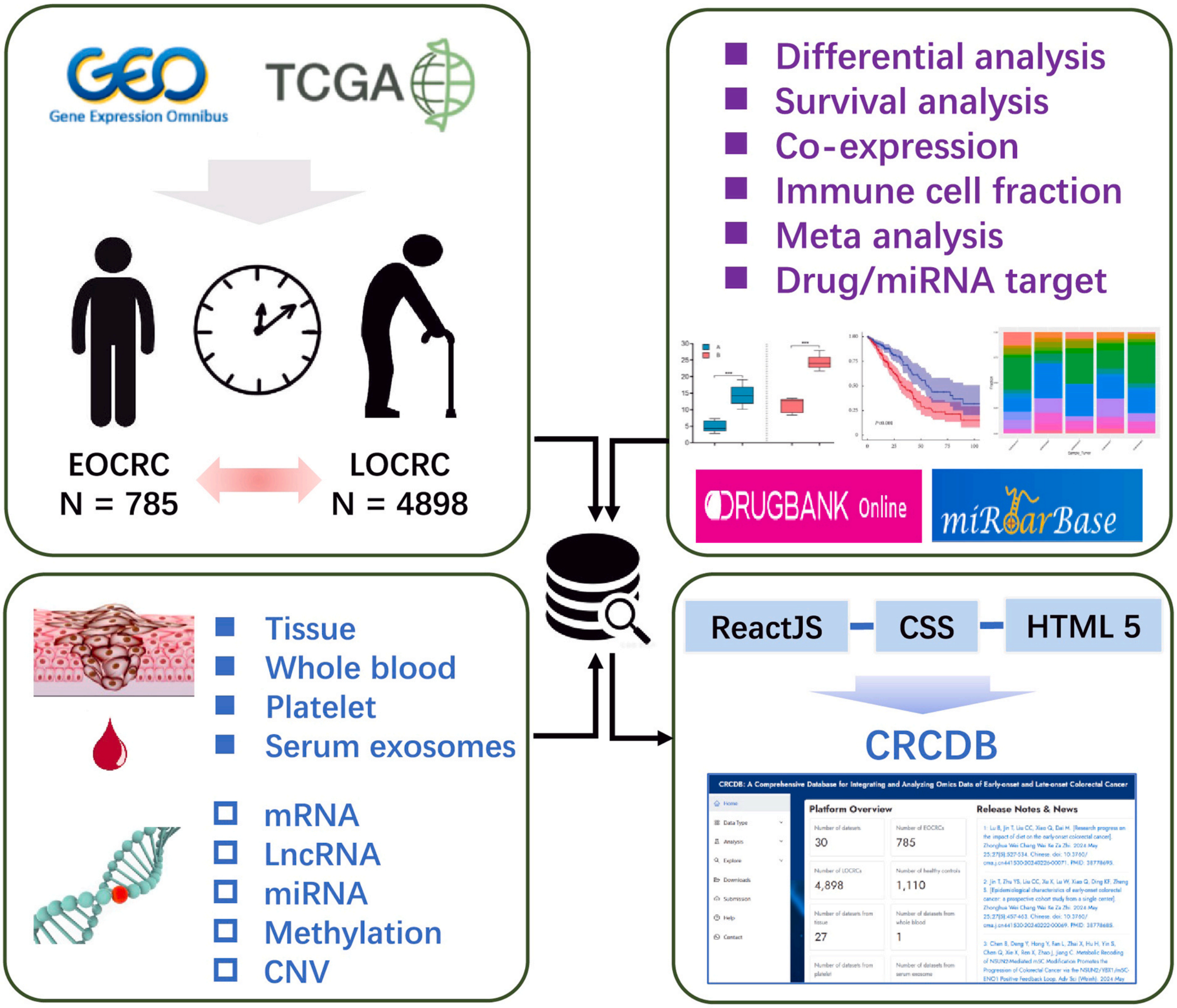

Gene Expression Omnibus
TCGA
EOCRC
N = 785
LOCRC
N = 4898
Differential analysis
Survival analysis
Co-expression
Immune cell fraction
Meta analysis
Drug/miRNA target
DRUGBANK Online
miRBase
Tissue
Whole blood
Platelet
Serum exosomes
mRNA
LncRNA
miRNA
Methylation
CNV
ReactJS
CSS
HTML 5
CRCDB
CRCDB: A Comprehensive Database for Integrating and Analyzing Omics Data of Early-onset and Late-onset Colorectal Cancer
Home
Data Type
Analysis
Explore
Downloads
Submission
Help
Contact
Platform Overview
Number of datasets 30
Number of EOCRCs 785
Number of LOCRCs 4,898
Number of healthy controls 1,110
Number of datasets from tissue 27
Number of datasets from whole blood 1
Number of datasets from platelet
Number of datasets from serum exosome
Release Notes & News
1: Lu B, Jin T, Liu CC, Xiao Q, Dai M. [Research progress on the impact of diet on the early-onset colorectal cancer]. Zhonghua Wei Chang Wai Ke Za Zhi. 2024 May 25;27(5):527-534. Chinese. doi: 10.3760/cma.j.cn441530-20240226-00071. PMID: 38778695.
2: Jin T, Zhu YS, Liu CC, Xu X, Lu W, Xiao Q, Ding KF, Zheng S. [Epidemiological characteristics of early-onset colorectal cancer: a prospective cohort study from a single center]. Zhonghua Wei Chang Wai Ke Za Zhi. 2024 May 25;27(5):457-463. Chinese. doi: 10.3760/cma.j.cn441530-20240222-00069. PMID: 38778685.
3: Chen B, Deng Y, Hong Y, Fan L, Zhai X, Hu H, Yin S, Chen Q, Xie X, Ren X, Zhao J, Jiang C. Metabolic Recoding of NSUN2-Mediated m5C Modification Promotes the Progression of Colorectal Cancer via the NSUN2/YBX1/m5C-ENO1 Positive Feedback Loop. Adv Sci (Weinh). 2024 May

Fig. 1. The work flowchart of the data collection and process of the platform. All the datasets were downloaded from the GEO or TCGA database. Samples are divided into three different age groups: EOCRC group, LOCRC group, and CA group composed of both EOCRC and LOCRC. The datasets include mRNA, lncRNA, and miRNA expression data, methylation data, and copy number variation data. Data were extracted from CRC tissue, whole blood, platelets, or serum exosomes. CRCCP presents the differential analysis, survival analysis, co-expression analysis, immune cell fraction comparison analysis, and meta-analysis results.

different age groups: EOCRC group, LOCRC group, or groups with both EOCRC and LOCRC patients which called CA group in this study. We downloaded the series matrix files from the GEO database. The datasets include mRNA, lncRNA, miRNA, methylation data, or CNV data. Data was extracted from tissue, whole blood, platelets, or serum exosomes from CRC patients and the healthy control samples. Gene expression was log-transformed when needed. The flowchart of the data processing procedure and overall design of CRCDB is shown in Fig. 1. Detailed information on all the datasets included in this platform can be found on the 'Help' page online.

2.2. Identification of differentially expressed genes and methylated probes

We regrouped the CRC patients as well as the normal control samples according to age and identified the differentially expressed genes between CRC patients and the normal controls in different groups, respectively. Genes with very low expression (average expression < 0.1) were removed; Student's *t*-test and "limma" package in R studio were used for the differential analysis and paired *t*-test was conducted when there was matched normal tissue. For methylation data, we used DMP in the R package "ChAMP" to identify the differentially methylated probes.

2.3. Survival analysis

For datasets with survival information, we calculated the association between gene expression and cancer outcome. Genes were divided into two groups according to the median value and log-rank analysis was used for the prognostic analysis. The cancer outcomes include OS (overall survival), DFS (disease free survival), DSS (disease specific survival), DFI (disease free interval), PFS (progression free survival), PFI (progression free interval), and RFS (recover free interval).

2.4. Correlation and network analysis

We used Pearson analysis to obtain the correlation value r and *P*-value between the selected two genes of a certain group of CRC patients. R packages "Hmisc" and "igraph" were used for the network calculation and graph presentation, respectively. The network chart was based on the related genes with *P*-value < 0.05.

2.5. Integration analysis

The robust rank aggregation algorithm (RRA) was used to integrate the results of differential analysis in an unbiased manner [13]. The aggregation rank score represents the integrated rank from the meta-analysis of the fold-changes in different studies. Meta-analysis was also calculated for the prognostic analysis by using "metagen" in the R package "meta".

2.6. Immune cell fraction analysis

For the tissue-originated gene expression data, we used CIBERSORT [14] coupled with LM22, which contains 547 genes that distinguish 22 human hematopoietic cell phenotypes to obtain the abundance of the 22 leukocyte subsets in each patient. Then we calculated the differences of the proportion of 22 immune cell types between CRC and normal controls in different age groups using Student's *t*-test, respectively.

2.7. Identification of the differences between EOCRC and LOCRC using partial data of CRCDB

We used Chi-square test or Fisher's exact test to compare the clinical or histological characteristics such as gender, disease stage, gene mutation, or relapse status in EOCRC and LOCRC groups. The R package "clusterProfiler" was then used for the GO enrichment analysis using the RRA and meta-analysis results.

2.8. Implementation

CRCDB was built using the NodeJS 8.10.0 (https://nodejs.org/en/) framework. MongoDB 3.6.5 (https://www.mongodb.com/) was used as the platform engine. The web interfaces were implemented in the JavaScript library of ReactJS (https://reactjs.org/). The platform runs on a Linux-based Nginx Web server and the CRCDB website is available online (http://crcdb-hust.com) with no registration requirement.

3. Results

3.1. Summary of the platform

We filtered out 30 datasets from 472 datasets with age information and enough samples for the database construction, which included 29 GEO datasets and the TCGA dataset. The data came from tissue, whole blood, platelets, or serum exosomes with 14 datasets had normal control samples, 9 datasets had paired normal tissue samples and 7 had the prognosis information. There were in total 5683 cancer samples with 785 EOCRCs, 4898 LOCRCs, and 1110 normal controls. Three RNA types (4916 lncRNAs, 19177 mRNAs, and 2402 miRNAs), methylation probes (486710 cpg probes), and CNVs (29590 genes) were included in the platform (Table S1).

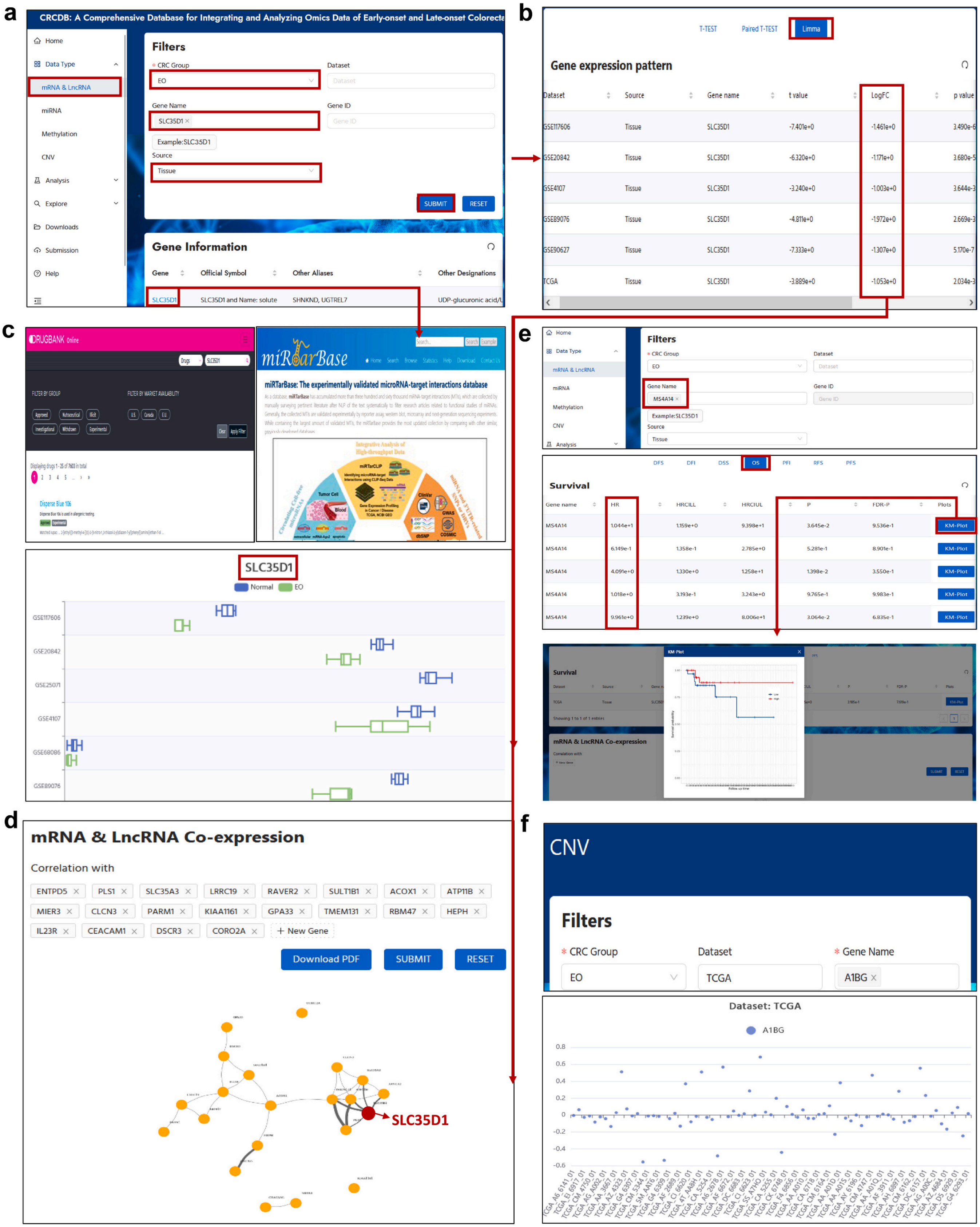
3.2. Clinical characteristics of EOCRC and LOCRC

By comparing the clinical and histological features between the LOCRC and EOCRC patients, we identified that there were more females ($P = 0.008$) and advanced stage patients ($P = 0.001$) in the EOCRC group. The frequency of *BRAF* mutation was lower ($P < 0.001$) and the frequency of *PIK3CA* mutation was higher ($P = 0.005$) in EOCRC. The results also showed that the EOCRC patients more frequently had lymph node involvement ($P = 0.007$) and liver and lung metastasis ($P = 0.020$) than LOCRC patients did. We did not find any statistically significant difference in BMI, microsatellite status, *KRAS* mutation, *TP53* mutation, lymphatic invasion, and relapse in EOCRC and LOCRC patients (Table S2).

3.3. Module usage in CRCDB

On the Home page of CRCDB, when users click on the examples presented at the bottom of the page, they will be led to the corresponding data type module. Research paper related to LOCRC and EOCRC were presented on the top right of the Home page.

In the "Data Type" module, we provide the basic gene information, differential analysis, survival analysis, or co-expression analysis results of each dataset of omics data. For example, in the mRNA&LncRNA submodule, after choosing the "EO" option of CRC group, filling in the gene name of "*SLC35D1*", and choosing the data source of "tissue", the platform presented the gene information, the differential analysis results, and survival analysis results of the gene. When choosing "limma" analysis, the results showed that this gene was down-regulated in all the EOCRC datasets (Fig. 2a-c). In addition, when clicking the gene name presented in the gene information table, users can be linked to: 1) the DrugBank database, which shows the target drugs of *SLC35D1*; 2) the miRTarBase database, which provides the miRNAs related to *SLC35D1* and users could search for the miRNAs' expression in the miRNA module; 3) the methylation probes related to *SLC35D1* in CRCDB, users could identify whether there were differentially methylated probes in the gene; or 4) the copy number variation of *SLC35D1* in each sample of the selected CRC group in CRCDB (Fig. 2c). Also, users can explore the co-expression and network analysis results between *SLC35D1* and a set of genes in the Co-expression part at the bottom of the page (Fig. 2d). *SLC35D1* was reported to be one of the key genes involved in UC-associated CRC [15] and its function in EOCRC is worth exploring. Here, we give another example for the survival analysis in the



CRCDB: A Comprehensive Database for Integrating and Analyzing Omics Data of Early-onset and Late-onset Colorecta

Filters

CRC Group: EO
Dataset: Dataset
Gene Name: SLC35D1
Gene ID: Gene ID
Example:SLC35D1
Source: Tissue
SUBMIT RESET

Gene Information

Gene	Official Symbol	Other Aliases	Other Designations
SLC35D1	SLC35D1 and Name: solute	SHNKND, UGTREL7	UDP-glucuronic acid/U

Gene expression pattern

Dataset	Source	Gene name	t value	LogFC	p value
GSE117606	Tissue	SLC35D1	-7.401e+0	-1.461e+0	3.490e-6
GSE20842	Tissue	SLC35D1	-6.320e+0	-1.171e+0	3.680e-5
GSE4107	Tissue	SLC35D1	-3.240e+0	-1.003e+0	3.644e-3
GSE89076	Tissue	SLC35D1	-4.811e+0	-1.972e+0	2.669e-3
GSE90627	Tissue	SLC35D1	-7.333e+0	-1.307e+0	5.170e-7
TCGA	Tissue	SLC35D1	-3.889e+0	-1.053e+0	2.034e-3

mRNA & LncRNA Co-expression

Correlation with: ENTPD5, PLS1, SLC35A3, LRRC19, RAVER2, SULT1B1, ACOX1, ATP11B, MIER3, CLCN3, PARM1, KIAA1161, GPA33, TMEM131, RBM47, HEPH, IL23R, CEACAM1, DSCR3, CORO2A
+ New Gene
Download PDF SUBMIT RESET

Filters

CRC Group: EO
Dataset: Dataset
Gene Name: MS4A14
Gene ID: Gene ID
Example:SLC35D1
Source: Tissue

Survival

Gene name	HR	HRCILL	HRCIUL	P	FDR-P	Plots
MS4A14	1.044e+1	1.159e+0	9.398e+1	3.645e-2	9.536e-1	KM-Plot
MS4A14	6.149e-1	1.358e-1	2.785e+0	5.281e-1	8.901e-1	KM-Plot
MS4A14	4.091e+0	1.330e+0	1.258e+1	1.398e-2	3.550e-1	KM-Plot
MS4A14	1.018e+0	3.193e-1	3.243e+0	9.765e-1	9.983e-1	KM-Plot
MS4A14	9.961e+0	1.239e+0	8.006e+1	3.064e-2	6.835e-1	KM-Plot

CNV

Filters

CRC Group: EO
Dataset: TCGA
Gene Name: A1BG

Fig. 2. Data type module of CRCDB. a. The basic information of the searched gene *SLC35D1*. b. Differential analysis results of tissue samples in EOCRC samples. c. Genes can be linked to the DrugBank database, the miRTarBase database, the methylation and CNV modeule. The box plot shows the expression of *SLC35D1* in each dataset. d. The co-expression analysis results of *SLC35D1*. e. *MS4A14* gene expression is associated with the overall survival outcome of EOCRC patients. d. Users can take a view of the copy number variation of their interested gene(s) in each sample of the input dataset.

mRNA&LncRNA sub-module. When we put "MS4A14″ in the "Gene name" box and chose the CRC groups and data source, the data showed that *MS4A14* was significantly associated with OS in EOCRC tissue in 3 of 5 datasets with *P* value < 0.05. By clicking on the KM plot, the survival curves could be shown and downloaded (Fig. 2e). *MS4A14* was identified to be associated with advanced stage and worse prognosis in both clear-cell renal cell carcinoma and gastric cancer [16,17], the function of *MS4A14* in EOCRC was unknown. For the CNV module, after choosing the specific CRC group and dataset, and putting in gene name (s), the platform provides scatter plots to show the variation of the gene (s) in each sample of the selected dataset (Fig. 2f).

In the "Analysis" module, the RRA or meta-analysis results are presented. Users can compare the gene expression profiles in three CRC groups in the "differential analysis" sub-module and the "survival analysis" sub-module. The color levels of the heatmap show the meta-FC or meta-HR values and the *P* values are presented in the boxes (Fig. 3a, b). Besides, users can have an overview of the expression of 22 immune cell fractions in the EOCRC, LOCRC, and overall CRC group of patients in each dataset in the TME sub-module (Fig. 3c). Then CRCDB would provide the differential analysis results of the cell infiltration between tumor and normal tissue, and the results would be shown in both table and bar chart form (Fig. 3d).

For the "Explore" module, in the Explore-Meta analysis part, users could choose the datasets and genes of their interest and get the new meta-analysis results in the heatmap format; and on the Explore-differential analysis page, users could upload the transcriptome data or the methylation data and do the differential analysis in combination with the data in the CRCDB (Figs. 3e-3f).

3.4. Query on the "Download" and "Submission" page

We provide users with the analysis results on the download page. In addition, we provided a submission interface to invite the community to upload novel expression profiles derived from CRC samples with clinical information, especially age information. Notably, users would be required to offer the other basic information including names, email addresses, and institution names, and please provide the information for us to access the data in the "Message box". More information about the dataset is also welcomed.

3.5. Bioinformatic analysis using the database

After we did the differential analysis in the individual dataset, RRA was used to integrate the results. There are some interesting observations. When considering the differentially expressed genes identified using limma analysis followed by RRA in tissue samples, the most differentially expressed genes ($|\text{LogFC}| > 2$, *P* value < 0.05) in EOCRC patients including up regulated genes of *MMP7*, *CDH3*, *KRT23*, *CLDN1* [18–21], and down regulate genes of *CD177*, *PCK1,* and *GUCA2A* [22,23], which were associated with immunity or proliferation, migration, and invasion of the tumor. The GO enrichment analysis identified that the most enriched pathways were the small molecule catabolic process, lipid catabolic process, fatty acid metabolic process, and leukocyte migration in EOCRC, which mainly consisted of down regulated genes (Fig. 4a**,** Fig. S1a-S1b), suggesting inefficient metabolic prosses in EOCRC. While in LOCRC, differentially expressed genes are mainly enriched in pathways associated with the mitosis process. Those genes were up regulated in LOCRC, such as *BOP1*, *DKC1*, and *WDR43*, illustrating a higher cell proliferative activity in LOCRC patients (Fig. 4b**,** Fig. S1c-S1d). When we pooled the data together, which formed the overall CRC group, the differentially expressed genes were similar to LOCRC (Fig. 4c). These results informed us that general analysis of the CRC patients might conceal the specific characteristics of EOCRC.

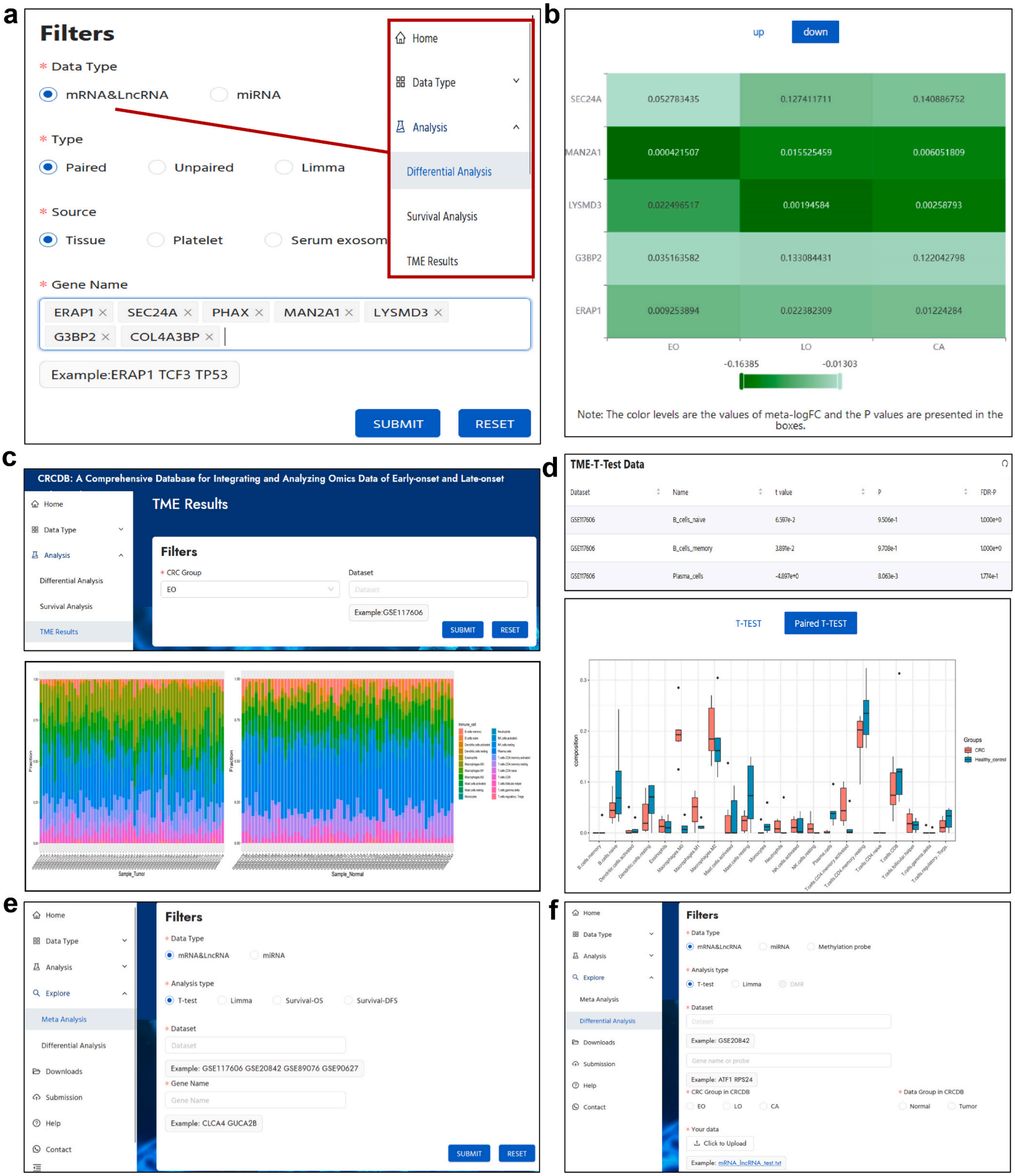
We did meta-analysis in different groups of patients integrating the survival analysis results from individual dataset, which could also lead to some inspiring results. Taking the OS related genes as an example, the GO analysis results showed that genes related to mitochondrial and oxidoreductase function were identified to play an important role in EOCRC. While in LOCRC, genes were most enriched in immune-related pathways including immunoglobulin production, phagocytosis, antigen receptor-mediated signaling pathway, regulation of B cell activation, and the pathway of plasma membrane invagination as well. As for overall CRC, genes were also enriched in the mitochondrial translation pathway but mainly in the immune-related pathways (Fig. 5).

4. Discussion

CRCDB is the first platform containing EOCRC patients and is a valuable resource for studying the molecular mechanisms of CRC. With the increasing number of high-through sequencing data, consistent patterns across multiple datasets would be beneficial to identify reliable biomarkers of CRC. Integrating data from GEO and TCGA, CRCDB provides a one-stop online resource for exploring molecular information of CRC with user-friendly interfaces.

In CRCDB, we presented the analysis results of differential analysis, prognosis analysis, and co-expression analysis of individual dataset and we provided meta-analysis results for the integration of all the datasets in different CRC groups. Thus, CRCDB might help to overcome technical problems when analyzing data from diverse studies. In addition, there is also an overview of the 22 immune cell fractions in the EOCRC, LOCRC, and overall CRC group of patients in each dataset. The mRNA, lncRNA, miRNA expression data, methylation level data, and CNV data come from not only tissue, but also whole blood, platelets, and serum exosomes which broaden the option of biomarkers for both early detection and therapy target identification. By using meta results of CRCDB combined with GO enrichment analysis, we confirmed that EOCRC has specific features and the previous research which pooled all the patients of EOCRC and LOCRC may hide the unique characteristics of EOCRC. Of note, we identified some important genes in the key pathways such as the metabolic process, mitosis process, oxidoreduction pathways, or in immune-related pathways that might be associated with the tumor progression and outcomes in the specific subtypes of CRC. These results might help in the understanding of the molecular mechanisms of the different CRC types, suggesting the effectiveness and utility of CRCDB to filter EOCRC or LOCRC biomarkers for follow-up experiments. From a treatment point of view, more over-expression of proliferation-associated genes among LOCRC patients may suggest a better response to proliferation-targeted therapy in that group. However, in the current study, we presented the bioinformatic analysis results using genes with *P* value < 0.05 considering taking into consideration of all the potentially cancer related genes. The analysis results might need validation in other datasets. In addition, considering long-term impact on quality of life such as fertility preservation, the economic and financial ramifications, and the psychosocial stressors, different treatment method may be used for the younger patients, which might result in the differences in omics [3,24], especially those omics data collected from patients who have been treated before the tumor sample collected. Thus, there might be some omics observations that were consequences of treatment rather than characteristic of the age of onset - early or late.

There are also some limitations of the study. Since the research related to CRC is accumulating, the datasets included in our platform are limited. We will improve CRCDB by providing more RNA expression profiles in the future version of this platform, such as circle RNA [25] and cell-free DNA [26]. And more data resources might be included in CRCDB especially data on liquid biopsy which has promising future applications in clinical needs in tumor patients [27,28]. Furthermore, datasets that include multi-omics information are limited and the majority multi-omics data comes from the different individuals. In CRCDB, we only obtained the datasets with diagnosed age information and sample size > 20 (EOCRC > 3), which might limit the datasets that include all the datatype. We will try to collect the datasets included multi-omics data in the future and add them to our database. We would



Filters

* Data Type: ● mRNA&LncRNA ○ miRNA

* Type: ● Paired ○ Unpaired ○ Limma

* Source: ● Tissue ○ Platelet ○ Serum exosom

* Gene Name: ERAP1 × SEC24A × PHAX × MAN2A1 × LYSMD3 × G3BP2 × COL4A3BP ×

Example:ERAP1 TCF3 TP53

	EO	LO	CA
SEC24A	0.052783435	0.127411711	0.140886752
MAN2A1	0.000421507	0.015525459	0.006051809
LYSMD3	0.022496517	0.00194584	0.00258793
G3BP2	0.035163582	0.133084431	0.122042798
ERAP1	0.009253894	0.022382309	0.01224284

Note: The color levels are the values of meta-logFC and the P values are presented in the boxes.

CRCDB: A Comprehensive Database for Integrating and Analyzing Omics Data of Early-onset and Late-onset

TME Results

Filters

* CRC Group: EO

Dataset: Example:GSE117606

TME-T-Test Data

Dataset	Name	t value	P	FDR-P
GSE117606	B_cells_naive	6.597e-2	9.506e-1	1.000e+0
GSE117606	B_cells_memory	3.891e-2	9.708e-1	1.000e+0
GSE117606	Plasma_cells	-4.897e+0	8.063e-3	1.774e-1

Filters

* Data Type: ● mRNA&LncRNA ○ miRNA

* Analysis type: ● T-test ○ Limma ○ Survival-OS ○ Survival-DFS

* Dataset: Example: GSE117606 GSE20842 GSE89076 GSE90627

* Gene Name: Example: CLCA4 GUCA2B

Filters

* Data Type: ● mRNA&LncRNA ○ miRNA ○ Methylation probe

* Analysis type: ● T-test ○ Limma ○ DMR

* Dataset: Example: GSE20842

Gene name or probe: Example: ATF1 RPS24

* CRC Group in CRCDB: ○ EO ○ LO ○ CA

* Data Group in CRCDB: ○ Normal ○ Tumor

* Your data: Click to Upload

Example: mRNA_lncRNA_test.txt

Fig. 3. Analysis module of CRCDB. a,b. The meta-analysis results of differential analysis using RRA. The color levels of the heatmap are in accordance with the meta-FC values and the *P* values are presented in the boxes. c. The 22 immune cells fraction in each sample of the selected dataset. d. The differential analysis results of the immune cell fraction between tumor and normal samples in both table and bar chart form. e. In the Explore-Meta analysis part, users could choose the datasets of their interest and get the new meta-analysis results in the heatmap format. f. On the Explore-differential analysis page, users could upload the transcriptome data or the methylation data and do the differential analysis in combination with the data in the CRCDB.

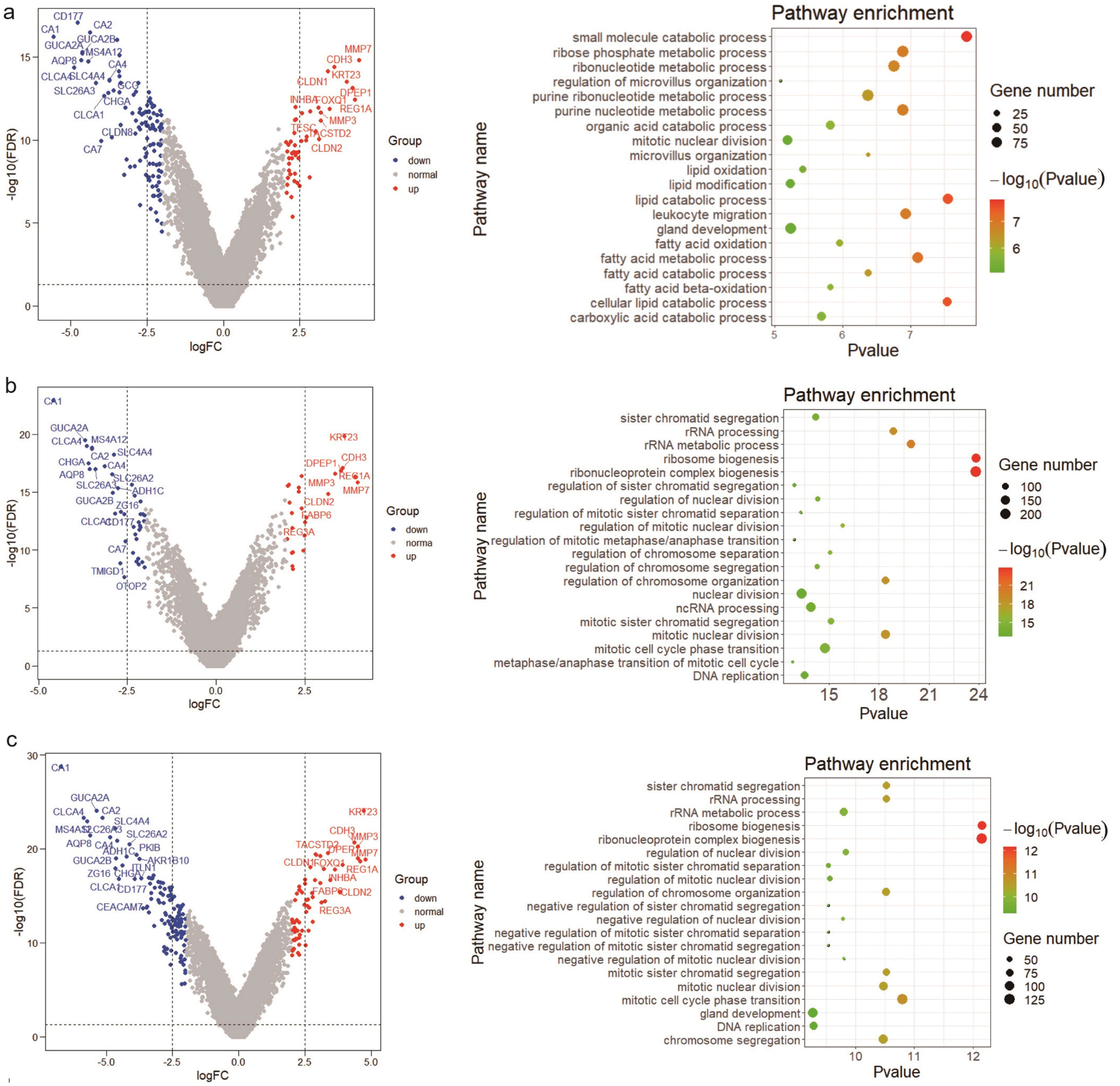


Fig. 4. The integrated differential analysis results using RRA. The volcano plots showed the up and down-regulated genes in EOCRC (a), LOCRC (b), and overall CRC (c) patients, respectively. GO enrichment analysis was conducted for the differentially expressed genes identified using limma analysis followed by RRA in tissue samples to explore the top 20 pathways in EOCRC, LOCRC, and overall CRC patients, respectively (adjust P value < 0.1).

obtain more public multi-omics data on different groups of CRC patients, including proteomic [29], metagenomic, and metabolomic data [30,31]. With more data collected, we will also upgrade the analysis results of our platform.

In conclusion, we presented the genomic landscape in tissue, platelets, whole blood, and serum exosomes in both EOCRC, LOCRC and overall CRC patients and presented the profiles in an open platform CRCDB. We believe CRCDB could serve as a very useful public resource for researchers and contribute to clinical studies.

Funding

This work was supported by the National Key Research and Development Program of China (2022YFC2408100), the National Natural Science Foundation of China (82173315 and 82202617), the Major Scientific and Technological Innovation Projects in Hubei Province (2022BCA013), and the China Postdoctoral Science Foundation (2021M701337 and 2022T150242).

Author statement

The authors declare no competing interests.

The work has not been published previously and it is not under consideration for publication elsewhere. All authors approved the final report for publication.

All the data used could be downloaded from UCSC Xena (https://xena.ucsc.edu/) and the GEO database (https://www.ncbi.nlm.nih.gov/geo/).

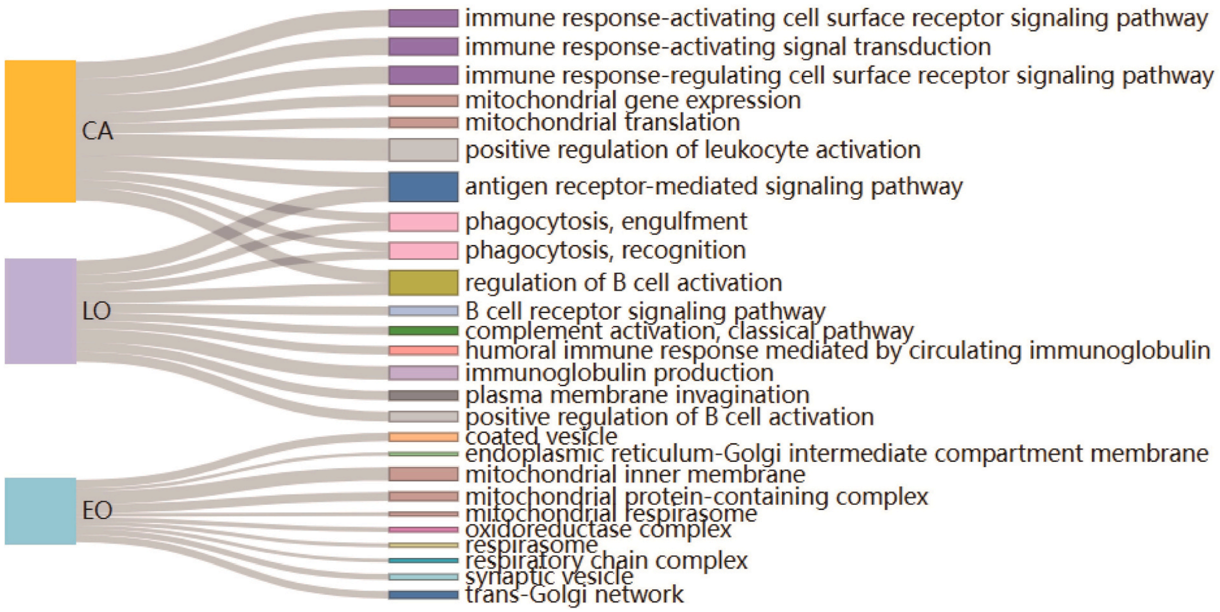


Fig. 5. The GO enrichment analysis of the survival-related genes**.** Meta-analyses were performed to integrate the OS analysis results in EOCRC, LOCRC, and overall CRC patients. The biological pathways were presented in the Sankey diagram after using GO enrichment analysis (adjust P value < 0.1).

The CRCDB website is available online (http://crcdb-hust.com) with no registration requirement.

CRediT authorship contribution statement

Luming Xu: Validation, Data curation. **Wanshan Ning:** Writing – original draft, Software, Resources, Methodology, Formal analysis, Data curation. **Danyi Zou:** Writing – original draft, Software, Resources, Methodology, Formal analysis, Data curation. **Zheng Wang:** Supervision, Project administration, Investigation, Funding acquisition, Conceptualization. **Lin Wang:** Supervision, Project administration, Investigation, Funding acquisition, Conceptualization. **Shijun Lei:** Writing – review & editing.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgments

We thank Weilin Nie for their help with web design and building.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.05.051.

References

[1] Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020, GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71:209–49.
[2] Siegel RL, Wagle NS, Cercek A, et al. Colorectal cancer statistics, 2023. CA Cancer J Clin 2023;73:233–54.
[3] Spaander MCW, Zauber AG, Syngal S, et al. Young-onset colorectal cancer. Nat Rev Dis Prim 2023;9:21.
[4] Stoffel EM, Koeppe E, Everett J, et al. Germline genetic features of young individuals with colorectal cancer. Gastroenterology 2018;154:897.
[5] Pearlman R, Frankel WL, Swanson B, et al. Prevalence and spectrum of germline cancer susceptibility gene mutations among patients with early-onset colorectal cancer. Jama Oncol 2017;3:464–71.
[6] Burnett-Hartman AN, Lee JK, Demb J. An update on the epidemiology, molecular characterization, diagnosis, and screening strategies for early-onset colorectal cancer. Gastroenterology 2021;160:1041–9.
[7] Guinney J, Dienstmann R, Wang X, et al. The consensus molecular subtypes of colorectal cancer. Nat Med 2015;21:1350–6.
[8] Spaander MCW, Zauber AG, Syngal S, et al. Young-onset colorectal cancer. Nat Rev Dis Prim 2023;9:21.
[9] Wang X, Liu J, Wang D, Feng M, et al. Epigenetically regulated gene expression profiles reveal four molecular subtypes with prognostic and therapeutic implications in colorectal cancer. Brief Bioinf 2021;22:bbaa309.
[10] Cui L, Li H, Bian J, Wang G, et al. Unsupervised construction of gene regulatory network based on single-cell multi-omics data of colorectal cancer. Brief Bioinf 2023;24:bbad011.
[11] Chisanga D, Keerthikumar S, Pathan M, et al. Colorectal cancer atlas, an integrative resource for genomic and proteomic annotations from colorectal cancer cell lines and tissues. Nucleic Acids Res 2016;44:D969–74.
[12] Zhang XL, Sun XF, Cao Y, et al. CBD, a biomarker database for colorectal cancer. Database-Oxf 2018:bay046.
[13] Kolde R, Laur S, Adler P. Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics 2012;28:573–80.
[14] Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 2015;12:453.
[15] Zhang D, Yan PG, Han TT, et al. Identification of key genes and biological processes contributing to colitis associated dysplasia in ulcerative colitis. Peerj 2021;9:e11321.
[16] Sun L, Zhang Y, Zhang C. Distinct expression and prognostic value of MS4A in gastric cancer. Open Med 2018;13:178–88.
[17] Li K, Li Y, Lyu Y, et al. Development of a phagocytosis-dependent gene signature to predict prognosis and response to checkpoint inhibition in clear-cell renal cell carcinoma. Front Immunol 2022;13:853088.
[18] Zhang Q, Liu S, Parajuli KR, et al. Interleukin-17 promotes prostate cancer via MMP7-induced epithelial-to-mesenchymal transition. Oncogene 2017;36:687–99.
[19] Han S, Wang Y, Ma J, et al. Sulforaphene inhibits esophageal cancer progression via suppressing SCD and CDH3 expression, and activating the GADD45B-MAP2K3-p38-p53 feedback loop. Cell Death Dis 2020;11:713.
[20] Kim D, Brocker CN, Takahashi S, et al. Keratin 23 is a peroxisome proliferator-activated receptor alpha-dependent, MYC-amplified oncogene that promotes hepatocyte proliferation. Hepatology 2019;70:154–67.
[21] Pope JL, Bhat AA, Sharma A, et al. Claudin-1 regulates intestinal epithelial homeostasis through the modulation of Notch-signalling. Gut 2014;63:622–34.
[22] Kim MC, Borcherding N, Ahmed KK, et al. CD177 modulates the function and homeostasis of tumor-infiltrating regulatory T cells. Nat Commun 2021;12:5764.
[23] Liu Y, Jin M, Wang Y, et al. MCU-induced mitochondrial calcium uptake promotes mitochondrial biogenesis and colorectal cancer growth. Signal Transduct Target Ther 2020;5:59.
[24] Lang D, Ciombor KK. Diagnosis and management of rectal cancer in patients younger than 50 years: rising global incidence and unique challenges. J Natl Compr Canc Netw 2022;20:1169–75.
[25] Xiong L, Liu HS, Zhou C, et al. A novel protein encoded by circINSIG1 reprograms cholesterol metabolism by promoting the ubiquitin-dependent degradation of INSIG1 in colorectal cancer. Mol Cancer 2023;22:72.

[26] Kotani D, Oki E, Nakamura Y, et al. Molecular residual disease and efficacy of adjuvant chemotherapy in patients with colorectal cancer. Nat Med 2023;29:127–34.

[27] Singh S, Gupta S. Promise and perils of blood-based signatures for detecting early-onset colorectal cancer. Gastroenterology 2022;163:1155–7.

[28] Nakamura K, Hernández G, Sharma GG, et al. A liquid biopsy signature for the detection of patients with early-onset colorectal cancer. Gastroenterology 2022;163. , 1242-1251.e2.

[29] Bech JM, Terkelsen T, Bartels AS, et al. Proteomic profiling of colorectal adenomas identifies a predictive risk signature for development of metachronous advanced colorectal Neoplasia. Gastroenterology 2023;165. , 121-132.e5.

[30] Kong C, Liang L, Liu G, et al. Integrated metagenomic and metabolomic analysis reveals distinct gut-microbiome-derived phenotypes in early-onset colorectal cancer. Gut 2023;72:1129–42.

[31] Du Z, Su J, Lin S, et al. Hydroxyphenylpyruvate Dioxygenase is a metabolic immune checkpoint for UTX-deficient colorectal cancer. Gastroenterology 2023;164. , 1165-1179.e13.