Discover Oncology

RESEARCH Open Access

Screening and validation of long non-coding RNAs associated with colorectal cancer based on random forest and LASSO regression algorithm

Yujia Zhao[1†], Qian Li[1,2†], Xintong Cui[1], Zhiyu Zhang[1], Yong You[3], Xiaowen Hou[1], Yan Wang[1,4] and Xu Feng[1*]

[†]Yujia Zhao and Qian Li have contributed equally to this work.

*Correspondence:
Xu Feng
fxsymc@163.com
[1]Department of Health Statistics, School of Public Health, Shenyang Medical College, 146 Huanghe North Street, Shenyang 110034, China
[2]Department of Childhood and Maternal and Child Health Care, School of Public Health, Jinzhou Medical University, Jinzhou 121001, China
[3]The Fourth People's Hospital of Shenyang, Shenyang 110034, China
[4]Department of Occupational Health and Occupational Medicine, School of Public Health, Shenyang Medical College, Shenyang 110034, China

Abstract

Objective Colorectal cancer (CRC) ranks as the third most prevalent contributor to global disease burden and represents the second highest mortality rate among all malignancies worldwide. Long non-coding RNAs (lncRNAs) are a new class of regulatory RNAs, which play a crucial role in the occurrence and development of colorectal cancer. Therefore, it is potentially important to use bioinformatics and machine learning methods to study novel biomarkers for CRC.

Methods The RNA-seq data of colorectal cancer and normal colorectal tissue were downloaded from the GEO database. Random forest (RF) and LASSO (Least Absolute Shrinkage and Selection Operator (LASSO) regression algorithms were constructed to screen lncRNAs closely related to CRC, and their screening efficiency was verified. Predict the regulatory genes of lncRNA and construct the ceRNA regulatory network of lncRNA-miRNA-mRNA. Quantitative real-time PCR (qRT-PCR) was used to verify its expression in colorectal cancer tissues and adjacent tissues, as well as its relationship with clinical features of CRC patients.

Result A total of 3028 CRC-related lncRNAs were initially screened from the GEO database, and 55 differentially expressed lncRNAs (DE lncRNAs) were finally selected through difference analysis. The key lncRNAs were further screened using RF and LASSO. The same gene in the screening results of the above two methods was selected as the key lncRNA of CRC. Finally, five key lncRNAs (NCAL1, CRNDE, HMGA1P4, EPIST and MT1JP) were selected, among them, the expressions of NCAL1, CRNDE and HMGA1P4 were upregulated compared with normal CRC tissues, while the expressions of EPIST and MT1JP were downregulated compared with normal colorectal tissues. The expression of 5 key CRC lncRNAs was verified, and each AUC is greater than 0.7, indicating a good screening effect. Since CRNDE has been studied by members of this research group before, it will not be further studied. It was predicted that 4 lncRNAs would interact with 16 miRNAs and 57 mRNAs. Four key lncRNAs, namely NCAL1, HMGA1P4, EPIST and MT1JP, were experimentally verified. qRT-PCR results showed that the expression of four key lncRNAs in CRC tissues and adjacent tissues had statistical significance ($p < 0.001$).

Conclusion In summary, we obtained 5 lncRNAs that may be closely related to colorectal cancer, including NCAL1, CRNDE, HMGA1P4, EPIST and MT1JP. This study found that NCAL1, HMGA1P4, EPIST and MT1JP may be candidate biomarkers for colorectal cancer.

Keywords Colorectal cancer, Long non-coding RNA, Random forest (RF), Least absolute shrinkage and selection operator (LASSO)

1 Introduction

Colorectal cancer (CRC) is one of the most common malignant tumors of the digestive system worldwide, ranking third in incidence and second in mortality [1]. At present, the treatment of colorectal cancer is mainly surgical resection and chemotherapy [2]. Despite the continuous progress in the diagnosis and treatment of CRC in recent years, the early diagnosis and treatment of CRC is still the focus and difficulty of current cancer research [3]. Therefore, it is the focus of clinical research to fully understand the mechanism of occurrence, development and metastasis of colorectal cancer, search for effective biomarkers, and improve diagnosis and treatment strategies. Long non-coding RNAs (lncRNAs) are RNA transcripts larger than 200 nucleotides long that do not encode proteins or peptides [4]. With the continuous deepening of research, it has been found that lncRNA can participate in many important biological processes [5, 6], including chromatin modification, chromosome silencing, transcriptional interference and transcriptional activation. lncRNA, as a biomarker, has demonstrated significant value in the diagnosis, prognosis assessment and treatment guidance of various diseases, especially in the field of precision medicine for cancer [7]. Aberrant expression patterns of lncRNA have been implicated in tumorigenic processes across multiple cancer type [8, 9]. Although the specific pathogenesis of CRC has not been fully studied, the important role of lncRNA in CRC has been proposed [10]. Random Forest (RF) is an algorithm that integrates multiple tree models based on Bagging in ensemble learning [11]. LASSO (Least absolute shrinkage and selection operator) algorithm is a method for feature selection and regularization at the same time. The basic idea is to minimize the sum of absolute values of regression coefficients by reducing them to less than a threshold value. The coefficients of low correlation features are compressed to 0 and deleted to achieve the purpose of dimensionality reduction [12]. In previous studies, RF and LASSO have been used to screen for factors that influence disease occurrence and prognosis [13, 14], as well as to screen tumor-associated differentially expressed genes to more accurately identify key hub genes associated with specific diseases [15, 16]. In this study, RF and LASSO were used to search for lncRNAs that may be closely related to CRC in GEO database, and their expression in CRC tissues and neighboring tissues was further verified by experiments. The screened lncRNAs can be used as potential diagnostic biomarkers and therapeutic targets in clinical studies, providing references for further studies.

2 Methods

2.1 Data collection

From a GEO database (https://www.ncbi.nlm.nih.gov/geo/) to download multiple data sets of colorectal cancer and normal colorectal tissues lncRNA expression spectrum data, After the batch effect is removed from the data sets of different GPL platforms, the data sets are then merged and used as the training set data. Then, the R 4.2.2 "limma"

software package is used to filter, normalize, estimate missing values, logarithm conversion and other processing of the data to obtain a standardized representation matrix.

2.2 Differential expression analysis of CRC-associated LncRNA

DEseq2 was used to analyze differentially expressed genes (DEGs) of lncRNA. DE lncRNAs (Differentially Expressed Long noncoding RNAs) data with the screening condition of |log2FoldChange| > 1 and $p < 0.05$ were selected for follow-up analysis.

2.3 Construct RF and LASSO regression algorithms to screen key CRC LncRNAs

Taking the DE lncRNAs obtained in the previous step as the basic data, the random forest model is constructed by using the "random forest" package in R 4.2.2. Two important parameters are set, ntree: the number of decision trees contained in the random forest, which is 500 by default; mtry: specifies the number of variables in the node used in the binary tree, and the quadratic root of the number of variables in the data set by default. After artificial successive selection, the optimal parameters of the model are screened. The model randomly generates lncRNA classification trees and scores the classification results, which are sorted according to the importance of genes by Gini index, and several key lncRNAs are obtained. Also based on the DE lncRNAs obtained in the previous step, LASSO regression algorithm is constructed using the R 4.2.2"glmnet" package, and penalty parameter adjustment (X) is performed based on the 10x cross-validation of the minimum standard. Lasso regression modeling is divided into two steps: finding the best K value and building model. In the trajectory diagram, if it tends to be stable after a certain point, the K value corresponding to that point is the best K value, and the smaller the K value, the better, so as to build the Lasso regression model; Then, the point with the smallest error is found, which is the target number of Lasso regression, and then the key lncRNA selected by Lasso regression is obtained. The common gene screened by the above two methods is the key lncRNA associated with CRC.

2.4 To evaluate the diagnostic efficacy of screened key LncRNAs for CRC

By downloading other chip data sets that meet the inclusion criteria, the batch effect of different GPL platforms is removed, and multiple data sets are merged to use this as a validation set. The ROC curve of each key lncRNA is drawn based on the lncRNA chip data in the verification set, so as to further evaluate their diagnostic efficiency. Quantitative evaluation of diagnostic performance is achieved by calculating the Area Under Curve (AUC) for each ROC curve. Generally speaking, the value of AUC ranges from 0 to 1, and AUC > 0.7 indicates that the model has good diagnostic performance.

2.5 Predict the genes interacting with LncRNA and construct the CeRNA network

The miRNAs that regulate the key lncRNAs were predicted using three online databases of lncRNA-miRNA, namely miRcode, starbase, and lncRNABase V.2. Then, the intersections of the genes in the three databases were taken to obtain the miRNAs that interact with lncRNAs. The target mRNA of the miRNA obtained in the previous step was predicted using the three databases of Targetscan, miRDB and miTarBase. Then, the intersection of the genes in the three databases was taken to obtain the target mRNA. The above interaction relationship was imported into the Cytoscape_v3.10.2 software, Construct the CRC-related lncRNA-miRNA-mRNA ceRNA network.

2.6 qRT-PCR verification of CRC key LncRNAs

A total of 73 human colorectal cancer tissue samples and adjacent tissue samples were collected from Shenyang Fourth People's Hospital. All patients signed the informed consent forms before the specimens were obtained. All patients were newly diagnosed and underwent surgical resection of colorectal cancer between October 2021 and August 2023 and did not receive chemotherapy or radiotherapy before surgery. Inclusion criteria were: Patients with pathological diagnosis of colon cancer. Patients were excluded if they had other malignancies or had received radiation or chemotherapy in the 3 months prior to enrollment. The collection of specimens was approved by the Ethics Committee of Shenyang Fourth People's Hospital (Approval Comment Number: 2021-kt-010). All methods were carried out in accordance with the institutional guidelines and regulations.

Total RNA was extracted from CRC and adjacent tissues using Trizol reagent (Thermo Fisher Scientific). QRT-PCR was conducted with the TB Green Premix Ex Taq kit (Takara) using Applied Biosystems 7500 Fast Real Time PCR System (Thermo Fisher Scientific) following the manufacturer's instruction. In all assays, GAPDH served as normalization control. Gene expression was calculated using the 2-△△CT method with each test performed in triplicate. The primers used were listed in Table 1.

2.7 Statistical analysis

R4.2.2 software was used to construct RF and LASSO regression algorithms for lncRNA screening, and SPSS 26.0 was used for statistical analysis. Median and quartile ranges were used to describe lncRNAs expression in colorectal cancer tissue and adjacent tissues. For comparison of expressions between groups, wilcoxon signed rank test and wilcoxon rank sum test were used, and $p < 0.05$ was considered statistically significant.

3 Results

3.1 Screening of key LncRNAs in CRC

3.1.1 Data collection

Three CRC-related datasets that met the inclusion criteria: GSE70880, GSE115856, and GSE102340 were downloaded from the GEO database as training sets. The original data of gene chip expression were sorted into gene-sample expression matrix. The three datasets are from the GPL19748, GPL16956, and GPL13825 platform. Overall, 41 CRC and 41 normal samples were included and their expression profiles were identified. After

Table 1 The primer sequences for qRT-PCR

lncRNA	Primer sequence (F = Forward; *R* = Reverse)
HMGA1P4	F: CCTAGCACACCCTCCTCCACTG
	R: TCAAACTCCTCCTGCTTTGTTTCCTG
MT1JP	F: CTTACCGCGGCTCGAAATGG
	R: GAGCTGTTCCCACATCAGGC
EPIST	F: GTTCGTTCGTTCGTTCGTTCGTTC
	R: GCGGGAATGTCTTTATTGGACGTTAC
NCAL1	F: TTGTCTGAAGGGCGAAGGAATGC
	R: AATCCTACATTAGTCATCCAGCCAACC
GAPDH	F: CAGGAGGCATTGCTGATGAT
	R: GAAGGCTGGGGCTCATTT

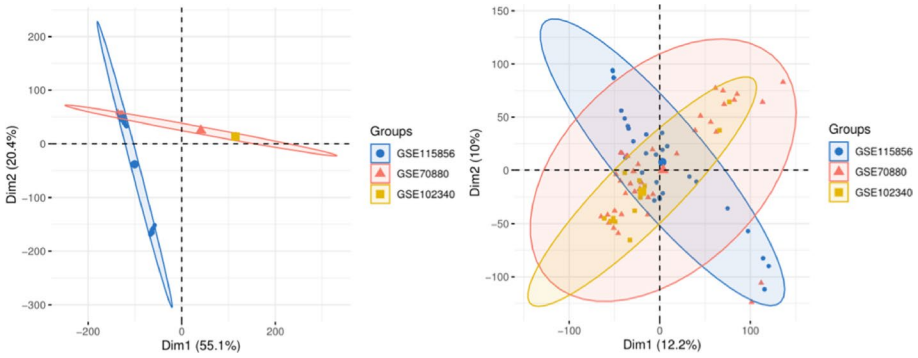

Fig. 1 Remove the batch effect of GSE70880, GSE115856 and GSE102340 (The left picture is before the batch is removed, and the right picture is after the batch is removed)

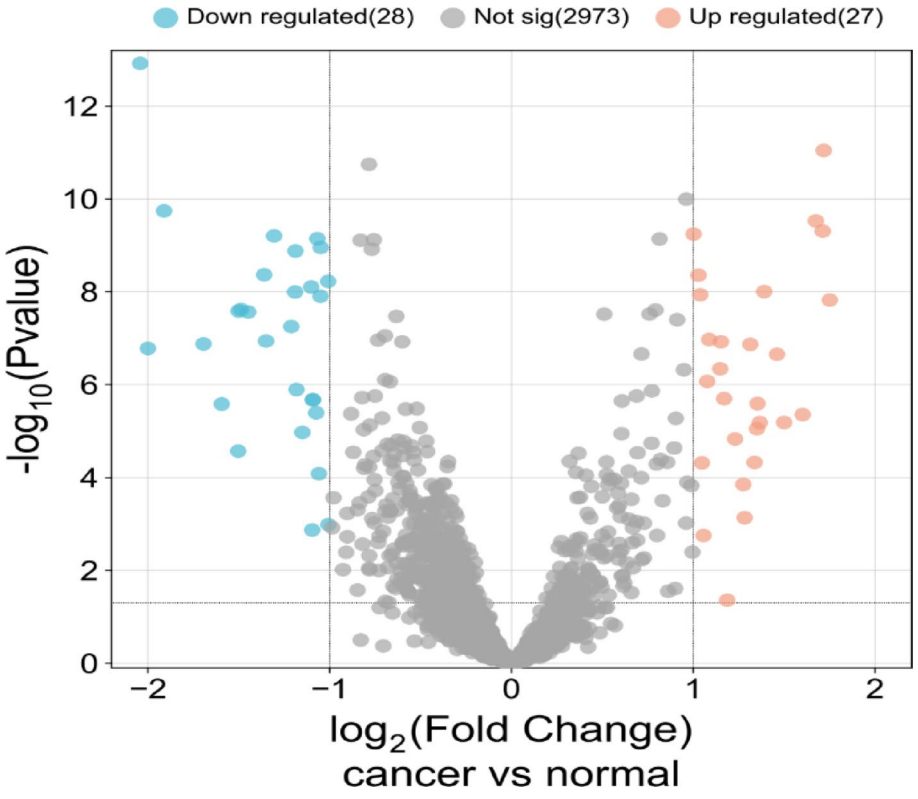

Fig. 2 lncRNA volcano map (In the volcano map, red represents genes with high expression, blue represents genes with low expression, and black represents genes with no difference in expression)

removing the batch effect of different GPL platforms, combine the data as a training set. The result is shown in Fig. 1.

3.2 Differential expression analysis of CRC-associated LncRNA

Finally, a total of 55 DE lncRNAs (Differentially Expressed long noncoding RNAs) were screened, of which 27 lncRNAs were up-regulated and 28 lncRNAs were down-regulated, which was used for subsequent analysis. The result is shown in Fig. 2.

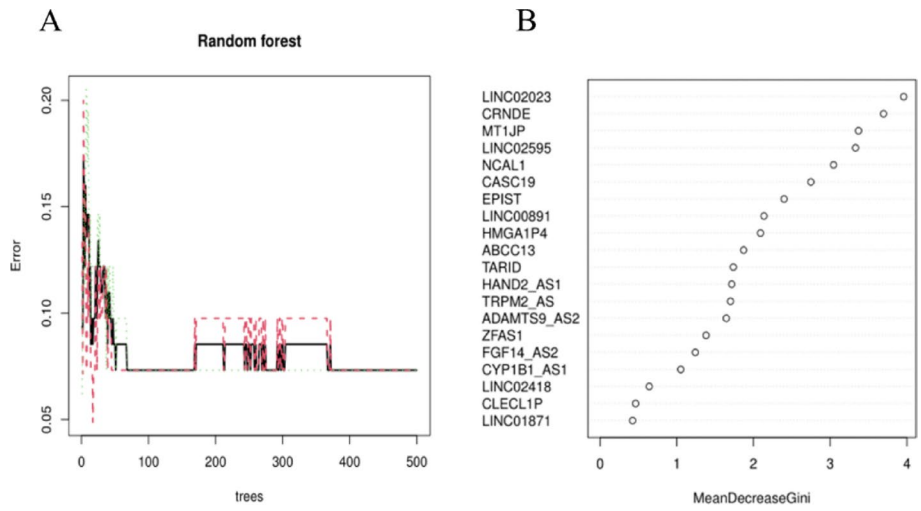

Fig. 3 Screening of CRC key lncRNAs by random forest

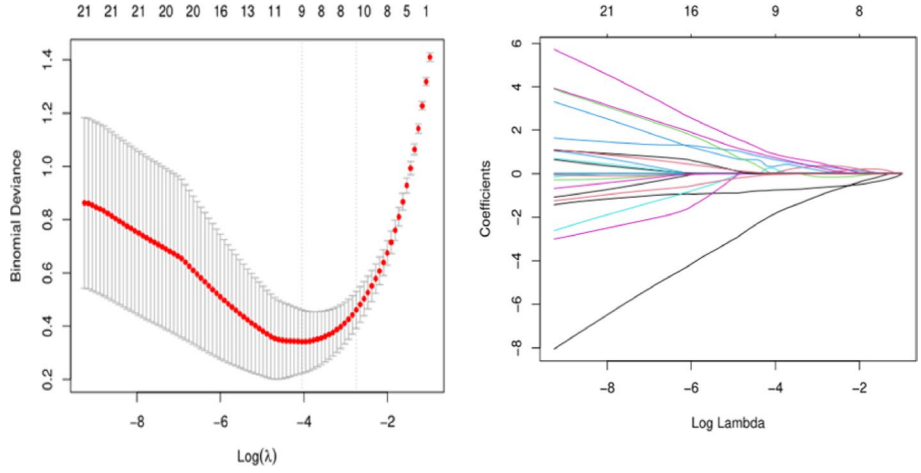

Fig. 4 LASSO regression algorithm screening key CRC lncRNAs

3.3 Construct RF and LASSO regression algorithms to screen key CRC LncRNAs

Take the DE lncRNAs obtained in the previous step as basic datas, sequence genes according to adj.P.Val in increasing order, select 20 genes successively from smallest to largest, and then use R package "random forest" for random forest classification, ntree is 500 by default. According to the obtained results (Fig. 3A), it can be determined that the error is minimal when ntree = 51, so the ntree is set to 51 for another random forest classification, and the top ten lncRNAs are sequenced according to the importance of genes by MeanDecrease Gini. The top ten lncRNAs are as follows: LINC02023, CRNDE, MT1JP, LINC02595, NCAL1, CASC19, EPIST, LINC00891, HMGA1P4 and ABCC13 (Fig. 3B).

Also based on the DE lncRNAs obtained in the previous step, LASSO regression algorithm is constructed using the R language "glmnet" package, family = "binomial" is used to fit the binary classification model, and 10 folds cross-validation is performed (nfolds = 10). lambda (lambda.min), which minimizes the mean square error, is selected as the optimal parameter. 9 lncRNAs selected by LASSO: MT1JP, CRNDE, EPIST, NCAL1, HMGA1P4, PCAT18, DUXAP10, DUXAP8, GATA2-AS1 (Fig. 4).

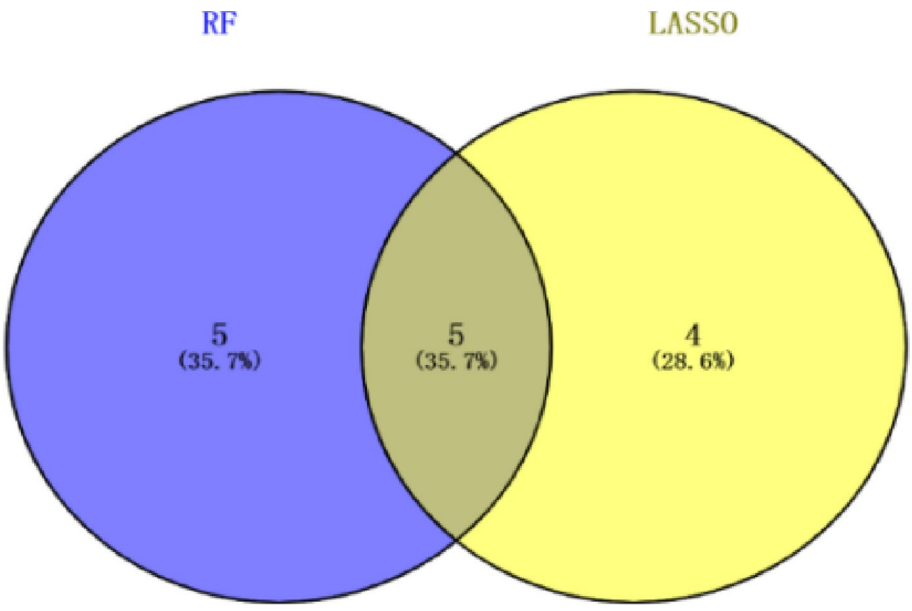

Fig. 5 Veen plots identical lncRNAs screened by random forest and LASSO regression

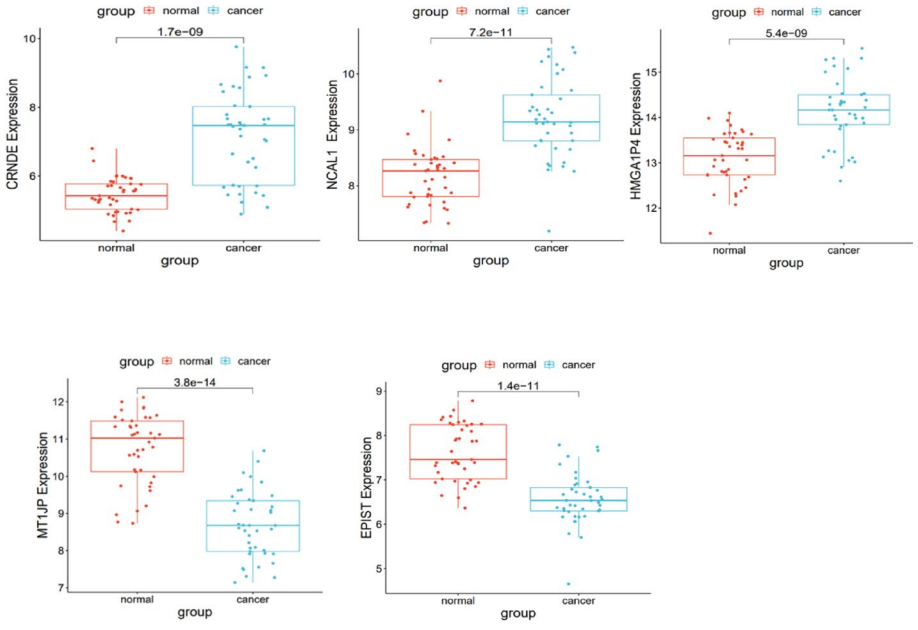

Fig. 6 The expression of five lncRNAs

The common lncRNAs of RF and LASSO screening were obtained by Venny 2.1.0 online software. The result is shown in Fig. 5.

Five key lncRNAs (NCAL1, CRNDE, HMGA1P4, EPIST and MT1JP) were selected, among them, the expressions of NCAL1, CRNDE and HMGA1P4 were upregulated compared with normal CRC tissues, while the expressions of EPIST and MT1JP were downregulated compared with normal colorectal tissues. The result is shown in Fig. 6.

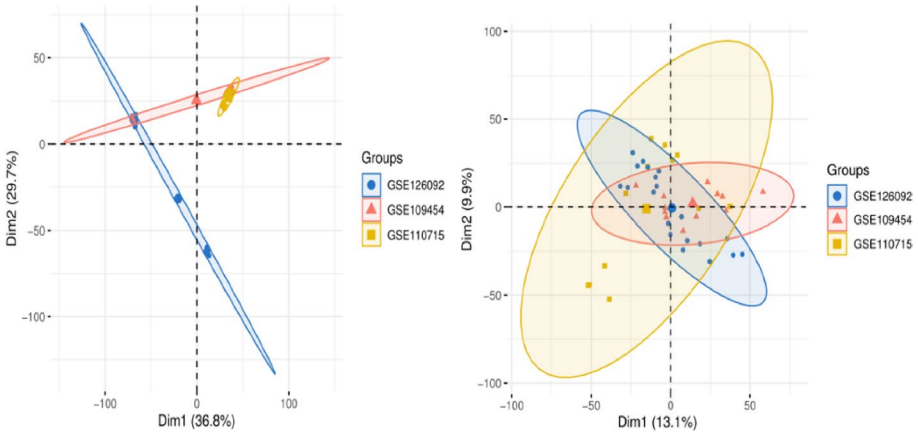

Fig. 7 Remove the batch effect of GSE126092, GSE109454 and GSE110715(The left picture is before the batch is removed, and the right picture is after the batch is removed)

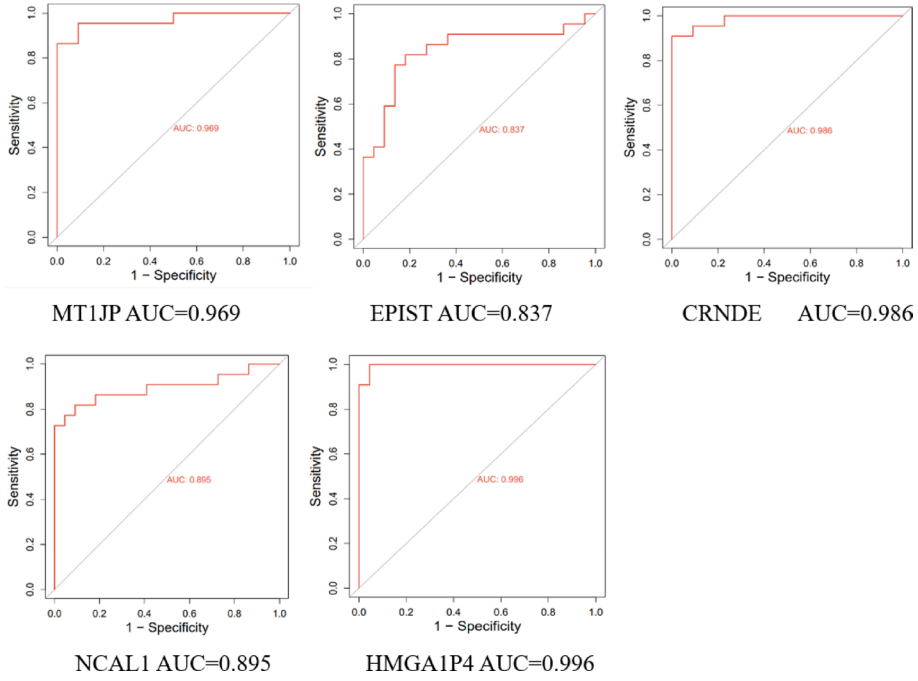

Fig. 8 ROC curves of five key lncRNAs were plotted based on verification set chip data

3.4 To evaluate the diagnostic efficacy of screened key LncRNAs for CRC

The CRC-related GSE126092, GSE109454 and GSE110715 were downloaded from the GEO database as validation sets. The three datasets are from the GPL21047, GPL16956 and GPL18180 platform, and contained data from 22 cases of colorectal cancer and 22 normal colorectal tissues. The batch effect was removed for the three data from different platforms, and then the data was merged together. The result is shown in Fig. 7.

ROC curves of 5 key lncRNAs were plotted based on lncRNA data in the validation set to evaluate their diagnostic efficiency (Fig. 8).

Table 2 MiRNAs interacting with LncRNAs

lncRNA	miRNA
CRNDE	hsa-miR-1277-5p、hsa-miR-620、hsa-miR-1270、hsa-miR-545-5p、hsa-miR-556-3p、hsa-miR-338-3p
EPIST	hsa-miR-6855-5p、hsa-miR-1908-5p、hsa-miR-1343-3p、
HMGA1P4	hsa-miR-301b-3p、hsa-miR-215-3p、hsa-miR-4782-3p
MT1JP	hsa-miR-3619-5p、hsa-miR-24-3p、hsa-miR-449c-5p、hsa-miR-1297

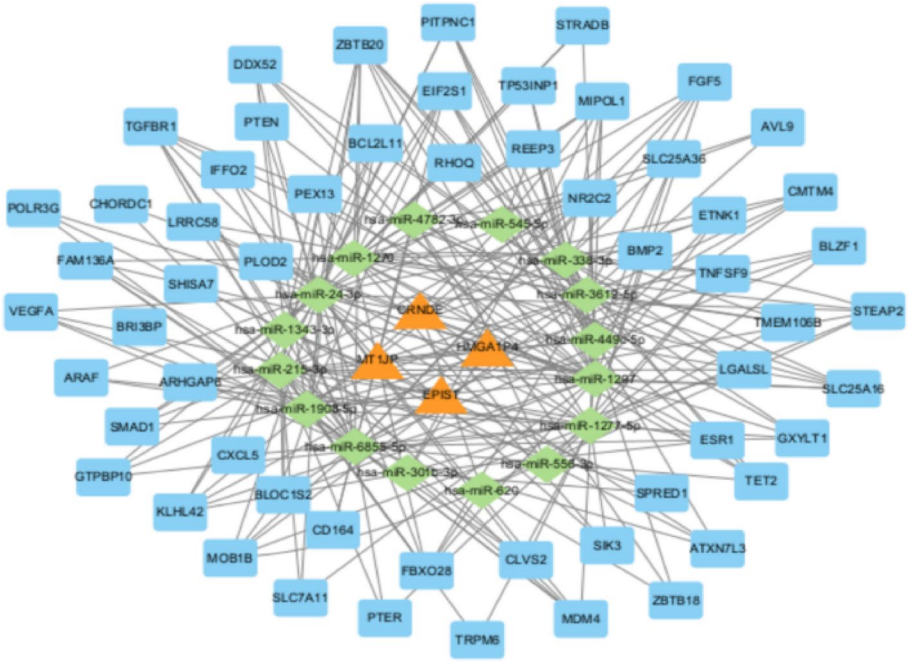

Fig. 9 lncRNA-miRNA-mRNA ceRNA regulatory network. Triangles represent lncRNAs, quadrilaterals represent miRNAs, and rectangles represent mRNAs

3.5 Predict the target genes interacting with lncRNA and construct the ceRNA network

It was predicted that 4 lncRNAs would interact with 16 miRNAs and 57 mRNAs (Table 2). The ceRNA network diagram is as follows (Fig. 9).

3.6 qRT-PCR verification of CRC key LncRNAs

3.6.1 Clinical characteristics of CRC patients

In this study, a total of 73 pairs of colorectal cancer tissues and adjacent tissues were collected for experimental verification according to inclusion and exclusion criteria. They ranged in age from 44 to 90, with an average age of 67. Among them, 39 were males, 34 were females, 52 were colon cancer, and 21 were rectal cancer (Table 3).

3.6.2 Expression of key LncRNAs in colorectal cancer tissues and neighboring tissues

The results of qRT-PCR showed that the relative expression levels of HMGA1P4, EPIST and MT1JP in colorectal cancer tissues were lower than those in neighboring tissues, and the difference was statistically significant ($p < 0.05$). The relative expression of NCAL1 in colorectal cancer tissues was higher than that in neighboring tissues, and the difference was statistically significant ($p < 0.05$) (Table 4).

Table 3 The clinical characteristics of CRC patients

Characteristics		*n* (73)
Age(year)	<65	33
	≥ 65	40
Gender	Male	39
	Female	34
Tumor size	<5 cm	15
	≥ 5 cm	58
TNM stage	I-II	50
	III-IV	23
Tumor position	Colon	52
	Rectum	21

Table 4 Expression of key LncRNAs in CRC tissues and adjacent tissues

LncRNA	Tumor		Normal		Z	*p*
	M	Q	M	Q		
NCAL1	2.750	3.974	1.094	2.035	−5.676	<0.001
HMGA1P4	0.552	0.476	1.039	0.819	−4.912	<0.001
EPIST	0.290	0.896	0.710	1.662	−3.593	<0.001
MT1JP	0.024	0.101	1.145	2.229	−6.935	<0.001

Table 5 Comparison of the expression levels of HMGA1P4 in CRC patients with different clinical characteristics

Characteristics	*n*	M	Q	Z	*p*
Age(year)					
<65	33	0.487	0.569	−0.532	0.595
≥ 65	40	0.566	0.436		
Gender					
Male	39	0.487	0.502	−1.377	0.169
Female	34	0.641	0.488		
Tumor size					
<5 cm	15	0.598	0.382	−0.614	0.539
≥ 5 cm	58	0.498	0.534		
TNM stage					
I-II	50	0.609	0.501	−2.001	0.045
III-IV	23	0.388	0.486		
Tumor position					
Colon	52	0.535	0.440	−0.810	0.418
Rectum	21	0.604	0.704		

3.6.3 Expression of key lncRNAs in colorectal cancer patients with different clinical characteristics

The expression of HMGA1P4 and MT1JP in cancer tissues of CRC patients with different clinical stages was significantly different ($p < 0.05$). The expression of HMGA1P4 and MT1JP in cancer tissues of CRC patients with clinical stage I to II was higher than that of CRC patients with stage III to IV. (Tables 5 and 6).

In addition, the results showed that NCAL1 and EPIST were not significantly correlated with gender, age, tumor location, tumor size, and clinical stage in patients with colorectal cancer ($p < 0.05$) (Tables 7 and 8).

Table 6 Comparison of the expression levels of MT1JP in CRC patients with different clinical characteristics

Characteristics	*n*	M	Q	Z	*p*
Age(year)					
<65	33	0.017	0.191	−0.781	0.435
≥ 65	40	0.280	0.072		
Gender					
Male	39	0.045	0.209	−0.697	0.486
Female	34	0.024	0.068		
Tumor size					
<5 cm	15	0.013	0.108	−0.430	0.667
≥ 5 cm	58	0.024	0.097		
TNM stage					
I-II	50	0.054	0.123	−2.066	0.039
III-IV	23	0.014	0.022		
Tumor position					
Colon	52	0.022	0.096	−0.896	0.370
Rectum	21	0.055	0.134		

Table 7 Comparison of the expression levels of NCAL1 in CRC patients with different clinical characteristics

Characteristics	*n*	M	Q	Z	*p*
Age(year)					
<65	33	3.148	3.370	−1.580	0.114
≥ 65	40	2.185	4.480		
Gender					
Male	39	2.897	3.133	−0.636	0.525
Female	34	2.467	6.128		
Tumor size					
<5 cm	15	2.769	5.631	−0.512	0.609
≥ 5 cm	58	2.672	4.033		
TNM stage					
I-II	50	2.672	3.172	−0.594	0.553
III-IV	23	3.271	5.610		
Tumor position					
Colon	52	2.953	5.621	−1.304	0.192
Rectum	21	2.361	2.634		

3.7 Evaluation of diagnostic value of NCAL1, HMGA1P4, EPIST and MT1JP in CRC

ROC curve analysis results showed that the AUC value of NCAL1 was 0.727, the sensitivity and specificity were 68.5%, and the Youden index was 0.370. The AUC value of HMGA1P4 was 0.763, the sensitivity was 67.1%, the specificity was 75.3%, and the Youden index was 0.424. The AUC value of EPIST is 0.691, the sensitivity is slightly lower (56.2%), but the specificity is higher (78.1%), and the Youden index is 0.342. It is worth noting that MT1JP exhibits a high AUC value of 0.919, with a sensitivity of 86.3%, a specificity of 83.6%, and a Youden index of 0.699. When the combined diagnosis method was adopted, the AUC value of the curve was further increased to 0.946, the sensitivity was as high as 90.4%, the specificity was 87.7%, and the Youden index reached 0.781 (Table 9). The AUC value, sensitivity, specificity and Jorden index of the combined diagnosis of four lncRNAs were higher than that of a single lncRNA; the AUC value, sensitivity, specificity and Youden index of MT1JP were all higher than those of other lncRNAs; the AUC value of NCAL1, HMGA1P4 and MT1JP were all above 0.7. These

Table 8 Comparison of the expression levels of EPIST in CRC patients with different clinical characteristics

Characteristics	*n*	M	Q	Z	*p*
Age(year)					
<65	33	0.287	0.359	−0.926	0.355
≥ 65	40	0.294	1.284		
Gender					
Male	39	0.203	0.514	−1.156	0.248
Female	34	0.334	1.287		
Tumor size					
<5 cm	15	0.203	0.192	−1.201	0.230
≥ 5 cm	58	0.325	1.308		
TNM stage					
I-II	50	0.289	0.545	−0.617	0.537
III-IV	23	0.344	1.322		
Tumor position					
Colon	52	0.290	0.511	−0.902	0.367
Rectum	21	0.262	4.446		

Table 9 The diagnostic value of LncRNA differential expression in CRC was analyzed by ROC curve

lncRNA	AUC	SE	Sensitivity (%)	Specificity (%)	*p*	95%CI	
						Lower	Upper
NCAL1	0.727	0.041	68.5	68.5	< 0.001	0.646	0.808
HMGA1P4	0.763	0.039	67.1	75.3	< 0.001	0.686	0.841
EPIST	0.691	0.045	56.2	78.1	< 0.001	0.604	0.779
MT1JP	0.919	0.022	86.3	83.6	< 0.001	0.875	0.963
Combined	0.946	0.018	90.4	87.7	< 0.001	0.911	0.982

results indicate that they have a relatively large role in distinguishing CRC tissue from adjacent non-cancerous tissues. The result of Fig. 10 is as follows.

4 Discussion

In recent years, significant progress has been made in the research of long non-coding RNAs related to colorectal cancer in terms of screening techniques, functional verification and clinical applications [17–22]. This study combines the Random Forest (RF) and the Least Absolute Contraction and Selection Operator (LASSO) algorithm. Five candidate lncRNAs (CRNDE, NCAL1, HMGA1P4, EPIST, MT1JP) that might be closely related to the occurrence and development of CRC were screened out from the CRC expression profile data in the GEO database. Verified by PCR experiments, we confirmed that the expression levels of EPIST, HMGA1P4, and MT1JP in CRC tissues were significantly downregulated compared with adjacent normal tissues, and the expression level of NCAL1 was significantly up-regulated compared with adjacent normal tissues in CRC tissues. They may be applied in clinical research as potential biomarkers for the diagnosis of colorectal cancer.

HMGA1P4 (high mobility groupA1▪ pseudogene4), located on chromosome 9. Studies have shown that HMGA1P4 plays an important role in hepatocellular carcinoma, gastric cancer and femoral atherosclerosis [23, 24], HMGA1P4 can also promote the resistance of gastric cancer cells to cisplatin [25]. However, so far, the relationship between HMGA1P4 expression in CRC and clinical features has not been reported. In this study, bioinformatics results showed that HMGA1P4 may interact with hsa-miR-301b-3p,

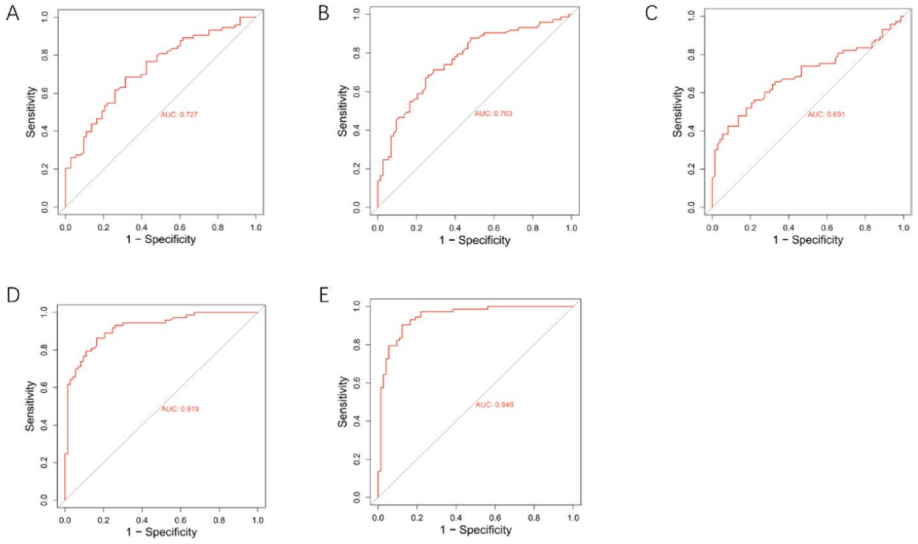

Fig. 10 ROC curves of diagnostic value of NCAL1 (**A**), HMGA1P4 (**B**), EPIST (**C**) and MT1JP (**D**) and combination (**E**) in CRC

hsa-miR-215-3p, and has-miR-4702-3p. Our bioinformatics analysis showed that HMGA1P4 expression was up-regulated in colorectal cancer tissues. Subsequently, we performed qRT-PCR on HMGA1P4, and the results showed that HMGA1P4 was a tumor suppressor gene of CRC ($p < 0.05$), which was contrary to the results of bioinformatics analysis. Possible reasons: On the one hand, this may be related to the fact that the experimental sample and the transcriptome sample are not the same batch of samples. For qRT-PCR verification, the RNA expression of different batches of samples will be different even if the treatment conditions are the same. Even for the same batch of samples, the RNA expression is specific in time and space. There may be significant differences in RNA expression at different time points or in different tissues at the same time point. In addition, the difference in preservation time and preservation method will also have an impact on RNA. In the future, the expression of HMGA1P4 in CRC can be further verified by increasing the sample size and conducting cell experiments. The correlation between HMGA1P4 and the clinical information of colorectal cancer patients showed that the expression of HMGA1P4 in colorectal cancer patients with different clinical stages had statistical significance ($p < 0.05$), and the expression level of HMGA1P4 in patients with colorectal cancer in stage I to II was higher than that in patients with colorectal cancer in stage III to IV. There was no significant difference in the expression of HMGA1P4 in colorectal cancer patients with different age, tumor size, sex and tumor site. ROC curve analysis of HMGA1P4 showed that the AUC value of HMGA1P4 was greater than 0.7, which helped distinguish colorectal cancer tissue from adjacent non-cancerous tissue.

Long non-coding RNA EPIST (also known as CTC-276P9.1 or C5orf66 antisense strand 1, c5orf66-AS1), located on chromosome 5, is a newly discovered tumor-associated lncRNA [26]. It has been reported to play an important regulatory role in gastric cancer, lung adenocarcinoma, cervical cancer, triple-negative breast cancer, osteosarcoma, oral squamous cell carcinoma, and hepatocellular carcinoma [26–34]. The results of this study showed that the expression of EPIST was down-regulated in colorectal cancer tissues, and it may interact with hsa-miR-6855-5p, hsa-miR-1908-5p, and

hsa-miR-1343-3p. The results of qRT-PCR showed that, EPIST was a tumor suppressor gene of CRC ($p < 0.05$), which was consistent with the analysis results. No correlation was found between EPIST and clinical stage, tumor size, age, tumor site, and sex of colorectal cancer patients. ROC curve results showed that the AUC value of EPIST = 0.691, indicating that EPIST had a little suggestive effect in distinguishing colorectal cancer tissues from adjacent non-cancerous tissues.

lncRNA MT1JP metallothionein 1 J, pseudogene (metallothionein 1 J, pseudogene, MT1JP) nucleotide chain length is 1348 bp, located in chromosome 16 16q13.

It plays a cancer-suppressing role in gastric cancer, retinoblastoma, liver cancer, lung cancer, osteosarcoma, breast cancer and glioma [35–44]. The analysis results in this study showed that the expression of MT1JP was down-regulated in colorectal cancer tissues. The results of qRT-PCR showed that MT1JP was a tumor suppressor gene of CRC ($p < 0.05$), which was consistent with the analysis results. MT1JP may interact with hsa-miR-3619-5p, hsa-miR-24-3p, hsa-miR-449c-5p, and hsa-miR-1297. There was statistical significance in the expression of MT1JP in CRC patients with different clinical stages ($p < 0.05$). The expression of MT1JP in cancer tissues of CRC patients with clinical stage I to II was higher than that of patients with stage III to IV. For CRC patients with different age, sex, tumor size and tumor site, there was no significant difference in the expression of MT1JP. ROC curve analysis results showed that the AUC value of MT1JP was 0.919, which was the highest prediction accuracy among the four lncRNAs, suggesting that MT1JP played a high role in separating colorectal cancer tissues from adjacent non-cancerous tissues.

NCAL1 (NK cell activity associated lncRNA 1), located on chromosome 2, is rarely studied at present. There have been no studies on the relationship between NCAL1 and CRC. For the first time, researchers found a new lncRNA-NK cell activity related lnCRNA-1 (NCAL1), and studied its function in NK cells. The anti-tumor effect of NK cells is closely related to the occurrence and development of tumors [45], but the molecular factors that determine the anti-tumor activity of NK cells remain to be identified. In this study, the analysis results showed that NCAL1 is an up-regulated gene in CRC. For the first time, qRT-PCR test found that the relative expression of NCAL1 in CRC tissues was significantly higher than that in adjacent tissues, suggesting that NCAL1 may be a cancer-promoting gene in CRC, which was consistent with the prediction results of bioinformatics. Statistical analysis of the clinical characteristics of NCAL1 and colorectal cancer patients found that there was no significant difference in the expression of NCAL1 in colorectal cancer patients with different clinical stages, tumor size, age, tumor site, and gender.

The RF-LASSO joint screening strategy adopted in this study effectively integrates the advantages of the two algorithms: RF can robustably evaluate the importance of features in high-dimensional data and reduce the risk of overfitting [46], while LASSO is good at precise feature selection and coefficient compression [47]. This study further confirmed its strong efficacy in mining key lncRNAs related to CRC, which is helpful for identifying the candidate molecules with the greatest biological and clinical significance from massive data. It should be pointed out that there are some limitations in this study. Firstly, the sample size in the experimental verification stage is relatively limited and comes from a single center. In the future, external verification needs to be conducted in a larger-scale, multi-center cohort to enhance the universality and reliability of the

results. Secondly, our research mainly focuses on the identification of expression differences of lncRNAs, and has not yet deeply explored the specific biological functions of lncRNAs and their exact molecular mechanisms in the occurrence and development of CRC through functional experiments (such as in vitro cell models or in vivo animal models). Furthermore, this study is mainly based on the chip data of GEO. In the future, integrating more sources (such as RNA-seq data of TCGA) and types of omics data (such as methylation, proteomics) will help to understand the functional networks of these lncRNAs more comprehensively.

5 Conclusions

1. Combined with bioinformatics methods, the random forest and LASSO regression algorithm were used to mine five CRC related lncRNAs: MT1JP, CRNDE, EPIST, NCAL1 and HMGA1P4, and through the construction of ceRNA networks, it was found that they may be involved in the occurrence and development of CRC.
2. qRT-PCR experiments confirmed that compared with para-cancer tissues, NCAL1 expression was up-regulated in CRC tissues, while MT1JP, EPIST and HMGA1P4 expression was down-regulated in CRC tissues, which is expected to become biomarkers for early diagnosis of CRC and potential targets for treatment.
3. The expression levels of HMGA1P4 and MT1JP are different in CRC patients with different clinical stages. The expression levels of HMGA1P4 and MT1JP in cancer tissues of stage I to II CRC patients are higher than those of stage III to stage IV CRC patients, which may be used as indicators of CRC progression and prognosis. ROC curve analysis showed that NCAL1, HMGA1P4 and MT1JP showed strong differential ability in distinguishing CRC tissue from adjacent non-cancerous tissue.

Acknowledgements
We acknowledge GEO database and online analysis tools for providing their platforms and contributors for uploading their meaningful datasets.

Author contributions
Yujia Zhao and Qian Li performed the qRT-PCR and completed the original draft manuscript, Zhiyu Zhang and Yong You performed the RF and LASSO analysis，Xiaowen Hou, Xintong Cui and Yan Wang conducted the data analysis，Xu Feng designed the study and revised the manuscript. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript. Yujia Zhao and Qian Li contributed equally to this article.

Funding
This work was supported by grants from the Liaoning Provincial Natural Science Foundation [No.2019-ZD-0324] and Science and Technology Innovation Fund Project for Postgraduates of Shenyang Medical College [No. Y20220521].

Data availability
RNA-Seq data were deposited into the Gene Expression Omnibus database under accession number GSE70880 and are available at the following URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi? acc=GSE70880.

Declarations

Ethics approval and consent to participate
All methods were carried out in accordance with the institutional guidelines and regulations. The patient data we used were acquired by publicly available datasets that were collected with patients' informed consent. For the collection of clinical samples, all patients signed the informed consent forms before the specimens were obtained. The study was reviewed and approved by the Ethics Committee of The Fourth People's Hospital of Shenyang (Approval Comment Number: 2021-kt-010). All authors agree to publication.

Competing interests
The authors declare no competing interests.

Received: 9 March 2025 / Accepted: 20 June 2025
Published online: 01 July 2025

References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229–63.
2. Weiser MR, Landmann RG, Kattan MW, et al. Individualized prediction of colon cancer recurrence using a nomogram[J]. J Clin Oncol. 2008;26(3):380–5.
3. Chi Y, Wang D, Wang J, et al. Long Non-Coding RNA in the pathogenesis of Cancers[J]. Cells. 2019;8(9):1015.
4. Clark MB, Johnston RL, Inostroza-Ponta M, et al. Genome-wide analysis of long noncoding RNA stability. Genome Res. 2012;22(5):885–98.
5. Deniz E, Erman B. Long noncoding RNA (lincRNA), a new paradigm in gene expression control [J]. Funct Integr Genomics. 2017;17(2/3):135–43.
6. Vance KW, Ponting CP. Transcriptional regulatory functions of nuclear long noncoding RNAs [J]. Trends Genet. 2014;30(8):348–55.
7. Zhang MQ, Yang BZ, Wang ZQ, et al. Fatty acid metabolism-related LncRNAs are potential biomarkers for survival prediction in clear cell renal cell carcinoma. Med (Baltim). 2024;103(8):e37207.
8. Shakeri F, Mohamadynejad P, Moghanibashi M. Identification of ASMTL-AS1 and LINC02604 LncRNAs as novel biomarkers for diagnosis of colorectal cancer. Int J Colorectal Dis. 2024;39(1):112.
9. Heidari R, Assadollahi V, Marashi SN, et al. Identification of novel lncRNAs related to colorectal cancer through bioinformatics analysis. Biomed Res Int. 2025;2025:5538575.
10. Liu J, Liu W, Li H, et al. Identification of key genes and pathways associated with cholangiocarcinoma development based on weighted gene correlation network analysis. PeerJ. 2019;7:e7968.
11. Qi Y. Randomforest for bioinformatics. In: Ensemble machine learning: methods and applications. New York: Springer; 2012. p. 307–23.
12. Tibshirani R. Regression shrinkage and selection via the Lasso. J Roy Stat Soc. 1996;58(1):267–88.
13. Zhang H, Chi M, Su D, et al. A random forest-based metabolic risk model to assess the prognosis and metabolism-related drug targets in ovarian cancer. Comput Biol Med. 2023;153:106432.
14. Lee TF, Chao PJ, Ting HM, et al. Using multivariate regression model with least absolute shrinkage and selection operator (LASSO) to predict the incidence of Xerostomia after intensity-modulated radiotherapy for head and neck cancer. PL O S One. 2014;9(2):e89700.
15. Huang X, Liu Y, Qian C, et al. CHSY3 promotes proliferation and migration in gastric cancer and is associated with immune infiltration. J Transl Med. 2023;21(1):474.
16. Chen DL, Cai JH, Wang CCN. Identification of key prognostic genes of triple negative breast Cancer by LASSO-Based machine learning and bioinformatics analysis. Genes (Basel). 2022;13(5):902.
17. Li Z, Liu Y, Yi H, et al. Identification of N6-methylandenosine related LncRNA signatures for predicting the prognosis and therapy response in colorectal cancer patients. Front Genet. 2022;13:947747.
18. Talebi A, Rokni P, Kerachian MA. Transcriptome analysis of colorectal cancer liver metastasis: the importance of long noncoding RNAs and fusion transcripts in the disease pathogenesis. Mol Cell Probes. 2022;63:101816.
19. Zhou C, Qiu Q, Liu X, et al. Novel exosome-associated LncRNA model predicts colorectal cancer prognosis and drug response. Hereditas. 2025;162(1):79.
20. Yuan L, Zhao J, Sun T, et al. A machine learning framework that integrates multi-omics data predicts cancer-related LncRNAs. BMC Bioinformatics. 2021;22(1):332.
21. Samadi P, Soleimani M, Nouri F, et al. An integrative transcriptome analysis reveals potential predictive, prognostic biomarkers and therapeutic targets in colorectal cancer. BMC Cancer. 2022;22(1):835.
22. Raza A, Khan AQ, Inchakalody VP, et al. Dynamic liquid biopsy components as predictive and prognostic biomarkers in colorectal cancer. J Exp Clin Cancer Res. 2022;41(1):99.
23. Zhang X, Zhang W, Jiang Y, et al. Identification of functional LncRNAs in gastric cancer by integrative analysis of GEO and TCGA data. J Cell Biochem. 2019;120(10):17898–911.
24. Wang K, Ye Y, Huang L, et al. The long Non-coding RNA AC148477.2 is a novel therapeutic target associated with vascular smooth muscle cells proliferation of femoral atherosclerosis. Front Cardiovasc Med. 2022;9:954283.
25. Qiao XL, Zhong ZL, Dong Y, et al. LncRNA HMGA1P4 promotes cisplatin-resistance in gastric cancer. Eur Rev Med Pharmacol Sci. 2020;24(17):8830–6.
26. Zhang H, Song J. Knockdown of LncRNA C5orf66-AS1 inhibits osteosarcoma cell proliferation and invasion via miR-149-5p upregulation. Oncol Lett. 2021;22(5):757.
27. Zhou Q, Li H, Jing J, et al. Evaluation of C5orf66-AS1 as a potential biomarker for predicting early gastric Cancer and its role in gastric carcinogenesis. Onco Targets Ther. 2020;13:2795–805.
28. Dong YY, Zhou Q, Li H, et al. Abnormally expressed LncRNAs as potential biomarkers for gastric Cancer risk: A diagnostic Meta-Bioinformatics analysis. Biomed Res Int. 2022;2022:6712625.
29. Guo W, Lv P, Liu S, et al. Aberrant methylation-mediated downregulation of long noncoding RNA C5orf66-AS1 promotes the development of gastric cardia adenocarcinoma. Mol Carcinog. 2018;57(7):854–65.
30. Rui X, Xu Y, Jiang X, et al. Long non-coding RNA C5orf66-AS1 promotes cell proliferation in cervical cancer by targeting miR-637/RING1 axis. Cell Death Dis. 2018;9(12):1175.
31. Luo W, Wang M, Liu J, et al. Identification of a six LncRNAs signature as novel diagnostic biomarkers for cervical cancer. J Cell Physiol. 2020;235(2):993–1000.
32. Zhu S, Sun J, Liu X, et al. CTCF-Induced LncRNA C5orf66-AS1 facilitates the progression of Triple-Negative breast Cancer via sponging miR-149-5p to Up-Regulate CTCF and CTNNB1 to activate Wnt/β-Catenin pathway. Mol Cell Biol. 2022;42(6):e0018821.
33. Feng L, Houck JR, Lohavanichbutr P, et al. Transcriptome analysis reveals differentially expressed LncRNAs between oral squamous cell carcinoma and healthy oral mucosa. Oncotarget. 2017;8(19):31521–31.
34. Lu T, Liu H, You G. Long non-coding RNA C5orf66-AS1 prevents oral squamous cell carcinoma through inhibiting cell growth and metastasis. Int J Mol Med. 2018;42(6):3291–9.
35. Yang J, Zhang Y, Liu P, et al. Decreased expression of long noncoding RNA MT1JP May be a novel diagnostic and predictive biomarker in gastric cancer[J]. Int J Clin Experimental Pathol. 2017;10(1):432–8.

36. Bi LL, Han F, Zhang XM, et al. LncRNA MT1JP acts as a tumor inhibitor via reciprocally regulating Wnt/β-Catenin pathway in retinoblastoma. Eur Rev Med Pharmacol Sci. 2018;22(13):4204–14.
37. Wu JH, Xu K, Liu JH, et al. LncRNA MT1JP inhibits the malignant progression of hepatocellular carcinoma through regulating AKT. Eur Rev Med Pharmacol Sci. 2020;24(12):6647–56.
38. Shan QL, Chen NN, Meng GZ, et al. Overexpression of LncRNA MT1JP mediates apoptosis and migration of hepatocellular carcinoma cells by regulating miR-24-3p. Cancer Manag Res. 2020;12:4715–24.
39. Mo W, Dai Y, Chen J, et al. Long noncoding RNA (lncRNA) MT1JP suppresses hepatocellular carcinoma (HCC) in vitro. Cancer Manag Res. 2020;12:7949–60.
40. Zhang S, Xu J, Chen Q, et al. LncRNA MT1JP-overexpression abolishes the Silencing of PTEN by miR-32 in hepatocellular carcinoma. Oncol Lett. 2021;22(2):604.
41. Ma J, Yan H, Zhang J, et al. Long-Chain Non-Coding RNA (lncRNA) MT1JP suppresses biological activities of lung Cancer by regulating miRNA-423-3p/Bim Axis. Med Sci Monit. 2019;25:5114–26.
42. Yang L, Liu G, Xiao S, et al. Long noncoding MT1JP enhanced the inhibitory effects of miR-646 on FGF2 in osteosarcoma. Cancer Biother Radiopharm. 2020;35(5):371–6.
43. Zhu D, Zhang X, Lin Y, et al. MT1JP inhibits tumorigenesis and enhances cisplatin sensitivity of breast cancer cells through competitively binding to miR-24-3p. Am J Transl Res. 2019;11(1):245–56.
44. Chen J, Lou J, Yang S, et al. MT1JP inhibits glioma progression via negative regulation of miR-24. Oncol Lett. 2020;19(1):334–42.
45. Niu C, Li M, Chen Y, et al. LncRNA NCAL1 potentiates natural killer cell cytotoxicity through the Gab2-PI3K-AKT pathway. Front Immunol. 2022;13:970195.
46. Geng R, Huang X, Li L, et al. Gene expression analysis in endometriosis: immunopathology insights, transcription factors and therapeutic targets. Front Immunol. 2022;13:1037504.
47. Zhang T, Chen Y, Xiang Z. Machine learning-based integration develops a disulfidptosis-related LncRNA signature for improving outcomes in gastric cancer. Artif Cells Nanomed Biotechnol. 2025;53(1):1–13.

Publisher's note