Biochemical Genetics
https://doi.org/10.1007/s10528-023-10516-4

ORIGINAL ARTICLE

Construction of Immune Infiltration-Related LncRNA Signatures Based on Machine Learning for the Prognosis in Colon Cancer

Zhe Liu[1] · Olutomilayo Olayemi Petinrin[1] · Muhammad Toseef[1] · Nanjun Chen[1] · Ka-Chun Wong[1]

Received: 10 July 2023 / Accepted: 5 September 2023

Abstract

Colon cancer is one of the malignant tumors with high morbidity, lethality, and prevalence across global human health. Molecular biomarkers play key roles in its prognosis. In particular, immune-related lncRNAs (IRL) have attracted enormous interest in diagnosis and treatment, but less is known about their potential functions. We aimed to investigate dysfunctional IRL and construct a risk model for improving the outcomes of patients. Nineteen immune cell types were collected for identifying house-keeping lncRNAs (HKLncRNA). GSE39582 and TCGA-COAD were treated as the discovery and validation datasets, respectively. Four machine learning algorithms (LASSO, Random Forest, Boruta, and Xgboost) and a Gaussian mixture model were utilized to mine the optimal combination of lncRNAs. Univariate and multivariate Cox regression was utilized to construct the risk score model. We distinguished the functional difference in an immune perspective between low- and high-risk cohorts calculated by this scoring system. Finally, we provided a nomogram. By leveraging the microarray, sequencing, and clinical data for immune cells and colon cancer patients, we identified the 221 HKLncRNAs with a low cell type-specificity index. Eighty-seven lncRNAs were up-regulated in the immune compared to cancer cells. Twelve lncRNAs were beneficial in improving performance. A risk score model with three lncRNAs (*CYB561D2*, *LINC00638*, and *DANCR*) was proposed with robust ROC performance on an independent dataset. According to immune-related analysis, the risk score is strongly associated with the tumor immune microenvironment. Our results emphasized IRL has the potential to be a powerful and effective therapy for enhancing the prognostic of colon cancer.

Keywords Colon cancer · Immune-related lncRNA (IRL) · Prognosis · Machine learning · Risk score system · Cell type-specificity index

✉ Ka-Chun Wong
kc.w@cityu.edu.hk

[1] Department of Computer Science, City University of Hong Kong, Hong Kong, China

Published online: 04 October 2023

Abbreviations

AS	Alternative splicing
BLCA	Bladder urothelial carcinoma
ccRCC	Clear cell renal cell carcinoma
C-index	Harrell's consistency index
ceRNA	Competitive endogenous RNA
COAD	Colon adenocarcinoma
DCA	Decision curve analysis
GEO	Gene expression omnibus
GSEA	Gene set enrichment analysis
HCC	Hepatocellular carcinoma
HKLncRNA	House-keeping LncRNAs
HR	Hazard ratio
ICI	Immunological checkpoint inhibitor
IRL	Immune-related LncRNAs
LASSO	Least absolute shrinkage and selection operator
LIHC	Liver hepatocellular carcinoma
LncRNA	Long-noncoding RNA
ncRNA	Non-coding RNA
OS	Overall survival
NSCLC	Non-small cell lung cancer
RF	Random forest
RMA	Robust multiarray average
ROC	Receiver operating characteristic curve
SAM	Significance analysis of microarrays
SE	Standard error
siRNA	Short-interfering RNA
TCGA	The Cancer Genome Atlas
TMB	Tumor mutation burden
tROC	Time-dependent ROC curve
VIM	Variable importance measure

Introduction

Strong heterogeneity and aggressiveness are the two main characteristics of colon cancer which has a high prevalence and fatality rate (Liu et al. 2022a). Colon tumor continues to be ranked the second-leading cause of death and the third-leading cause of incidence despite improvements made in recent decades (Sung et al. 2021). Female colorectal cancer accounts for 10% of all cancers in women. Colon cancer is the second leading cause of cancer death, accounting for 9.4% of all death (Sung et al. 2021). Thus, it is crucial to provide a model for identifying biomarkers that benefit cancer therapy in the early stage and distinguishing poor prognosis groups.

Tumor-infiltrating immune cells are important for the development and aggressiveness of cancer, according to the expanding body of research on the tumor microenvironment (TME) (Jochems and Schlom 2011; Bense et al. 2017; Barnes and

Amir 2018). There is evidence that distinct kinds of immune cells related to clinical outcomes are abundant in the microenvironment of colon cancer (Mola et al. 2020; Ooki et al. 2021; Liang et al. 2022; Vitorino et al. 2022). As a result, the quantitative molecular signature of immune cells that infiltrate tumors is being recognized as a class of prognostic biomarkers that may help patients better manage and choose their own treatment. Numerous lncRNAs have been demonstrated to be crucial in controlling transcription, translation, and protein modification, among other cellular and biological processes in cancer (Peng et al. 2017). LncRNAs have recently been found in a variety of immune cells and have been identified as essential regulators of immune cell growth and differentiation (Turner et al. 2014; Elling et al. 2016; Chen et al. 2017). LncRNAs are also linked to the immunological control of cancer, including immune activation, immune escape, dendritic cell (DC), T cell, regulatory T cell, B cell, and macrophage penetration into cancer tissues (Denaro et al. 2019; Egranov et al. 2020). The tumor infiltration immune-related lncRNA signatures have been established in glioblastoma and non-small cell lung cancer (NSCLC) (Sun et al. 2020a, b; Zhang et al. 2022a, b).

Immunological checkpoint inhibitors (ICIs) are a cutting-edge type of tumor immunotherapy that works by focusing on immune checkpoint proteins (Mahoney et al. 2015). However, only a tiny percentage of patients have thus far seen a significant improvement after receiving ICI treatment (Robert 2020). Hence, researchers need to develop a score for splitting all patients into poor and good immune response groups. Clinical doctors give personalized treatment strategies for colon cancer patients based on molecular characteristics. Long-noncoding RNAs (lncRNAs) are a class of non-coding RNAs (ncRNAs) with a length of more than 200 nt, which don't encode proteins but directly play a role in the formation of RNAs (Liu et al. 2022a). LncRNAs regulate the expression of protein-coding genes at the transcriptional and post-transcriptional levels and participate in the life processes (Park et al. 2022). Notably, recent research has shown that lncRNAs have critical roles in immune response, immune cell formation, differentiation, function, the tumor immunological microenvironment, and cancer immunotherapy (Coker and Wood 1986; Najafi et al. 2022). In addition, the expression specificity of immune-related lncRNA makes it can be a promising biomarker. Wu et al. reported eight immune-related lncRNAs classifier was applied to predict recurrent bladder cancer (Wu et al. 2020b). Four lncRNAs have been identified by Li et al. as potential independent prognostic variables for triple-negative breast cancer. They also confirmed that the high-risk group has strong immune responses (Li et al. 2021). A systematic and exhaustive strategy to find lncRNAs linked to immunological prognosis in colon cancer is currently lacking. Thus, we used the lncRNAs as risk factors for constructing the risk model.

Recently, machine learning-based algorithms have been widely used to mine prognostic factors in cancer research. Machine learning-based technology can be used to identify genes, CT-scan features, or clinical characteristics that are associated with patients' survival. Then, the prognostic model built by these genes or clinical characteristics was utilized to infer the risk score, which is an index value that evaluates the effect of therapy. According to the prediction results, researchers give treatment suggestions. It is a constructive way that can apply to many kinds of cancer. For example, Liu et al. defined risk genes in colorectal cancer based on the importance score of genes from the RF algorithm (Liu et al. 2023). Liu et al. selected alternative splicing (AS) events without co-linear correlation tested by the LASSO algorithm to feed into the Cox regression model to predict the survival time of the bladder urothelial carcinoma (BLCA) cohort (Liu et al. 2022b). Similarly, Zheng et al. reported a CT-based nomogram in clear cell renal cell carcinoma (ccRCC) by considering the 20 features filtered from 1316 radiomics features using LASSO logistic regression (Zheng et al. 2021). However, there is no integrated method proposed considering the advantages of various machine-learning algorithms. Here, we selected important lncRNA features based on the occurrence twice among five machine learning algorithms.

The birth of high-throughput sequencing technology can be said to be a landmark event in the field of genomics research (Pareek et al. 2011; Slatko et al. 2018). This technology makes the single-base cost of nucleic acid sequencing drop sharply compared with the first-generation sequencing technology. Before the advent of deep sequencing technology, the primary method for high-throughput measurement of different gene expression levels was the gene microarray (Hung and Weng 2017; Nurk et al. 2022). On this basis, the differences and patterns of gene expression in different tissues or different developmental stages could be analyzed. With the successful completion of the human genome project in 2003, sequencing technology has dramatically improved (Collins et al. 2003). These advances provide researchers and medical diagnosticians an excellent platform for further understanding phenotypic changes and disease development caused by genomic variation. GEO and TCGA are two data resources for providing us with lots of data (Edgar et al. 2002; Barrett et al. 2013), including microarray, RNA-seq, clinical information, and so on. We constructed a risk model based on microarray and clinical traits data, and then we validated this model by RNA-seq data.

In this study, we constructed a risk score model using tumor immune infiltration-related and prognostic lncRNAs. The risk score model has high specificity and sensitivity across training and testing datasets. We identified three lncRNAs, which can illustrate the mechanism of the tumor progression, improve the prognosis, and design new drug targets for colon cancer.

Materials and Methods

Immune Cell Types Data

We collected 19 different immune cell types (B cell activated, CD4 T cell activated, CD4 T cell resting, CD8 T cell activated, CD8 T cell resting, Dendritic cells activated, Dendritic cells resting, Eosinophils, Immature dendritic cells, Mast cells activated, Monocytes, Myeloid dendritic cells, NK activated, NK resting, NKT activated, Neutrophils, Plasmacytoid dendritic cells, T gamma delta, and T helper 17) for 115 samples from the GEO database with the accession numbers: GSE13906, GSE23371, GSE25320, GSE27291, GSE27838, GSE28490, GSE28698, GSE28726, GSE37750, GSE39889, GSE42058, GSE49910, GSE51540, GSE59237, GSE6863, and GSE8059. Detailed information about the immune cell microarray data is shown in Supplementary Table 1. The platform of these data is HG-U133_Plus_2.

Data Collection and Pre-processing of Colon Cancer

Raw ".cel" format microarray data for colon cancer was downloaded by the GEOquery R package (version: 2.64.2) with the accession number GSE39582 and platform number GPL570 (HG-U133_Plus_2) (Edgar et al. 2002). The Robust Multiarray Average (RMA) algorithm was selected for background correction, quantile normalization, and log2 transformation using the "affy" R package (version: 1.74.0) (Gautier et al. 2004). We saved the clinical information for each patient. Gene annotation was performed by matching the probe id to the gene symbol from NetAffx (https://sec-assets.thermofisher.com/TFS-Assets/LSG/Support-Files/HG-U133_Plus_2-na36-annot-csv.zip).

RNA-seq data of colon adenocarcinoma (COAD) was collected from the TCGA project via TCGAbiolinks (version: 2.25.0) R package (Colaprico et al. 2016). We saved the corresponding clinical data for patients. The gene type information came from GENCODE (https://www.gencodegenes.org/, version: GRCh38/hg38). The lncRNA gene information was from the GENCODE database (https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_21/gencode.v21.long_noncoding_RNAs.gtf.gz).

The detailed clinical characteristics of the training and testing datasets can be found in Table 1.

The Identification of HKLncRNAs

In this study, we aimed to explore immune-related and can be used for immunotherapy lncRNAs in colon cancer. In general, the expression levels of lncRNAs are lower than mRNAs, so it is difficult for researchers to detect them (Park et al. 2022). Thus, we performed the following two analyses: (A) identify highly expressed lncRNAs. (B) identify lncRNAs that are present in all immune cells.

Capture Highly Expressed lncRNAs

We determined the expression level of each lncRNA in each immune cell based on the average value of this lncRNA across all cells belonging to this immune cell. The top 30% expressed lncRNAs in each immune cell in descending order were merged into a gene list, and then we defined this gene list as highly expressed lncRNAs.

Capture Widely Expressed lncRNAs

The cell type-specificity index was used to evaluate the expression pattern of lncRNAs in immune cells as follows (Smith 1992; Yanai et al. 2005):

Table 1 The clinical traits of colon cancer patients in each dataset

Characteristics	Training dataset GSE39582	Testing dataset TCGA
Platform	HG-U133_Plus_2	IlluminaHiSeq
Patients	566	461
Age		
<60	150	126
≥ 60	416	333
NA	0	2
Gender		
Female	256	216
Male	310	243
NA	0	2
Stage		
Stage 0	4	0
Stage I	33	76
Stage II	264	178
Stage III	205	129
Stage IV	60	65
NA	0	13
Survival		
Dead	191	102
Alive	371	357
NA	4	2

Note that we only used colon cancer samples in our study

NA not applicable, *TCGA* The Cancer Genome Atlas

$$Specificity\ Index_{lncRNA} = \frac{\sum_{i=1}^{N}\left(1 - x_{lncRNA,i}\right)}{N - 1}$$

where N indicates the number of cell types, and the x_{lncRNA} indicates the normalized expression levels of lncRNA by max expression level. Cell type-specificity index values were calculated and sorted in ascending order. The top 30% lncRNAs with lower cell type-specificity index values were defined as HKLncRNA. And the bottom 30% lncRNAs with higher cell type-specificity index values were defined as cell type-specificity lncRNAs.

Capture lncRNAs That Are Up-Regulated in the Immune Compared to Cancer Cells

Significance analysis of microarrays (SAM) algorithm was chosen for capturing lncRNAs that are upregulated in immune compared to cancer cells. The cutoff of the significant level was set to 0.05 based on FDR correction. In this study, samr (version: 3.0) was utilized for performing DEG analysis (Tusher et al. 2001; Tibshirani 2006; Li and Tibshirani 2013; Tzeng 2021).

The Selection of Optimal lncRNAs

Before the construction of the risk score model, four machine learning-based methods (LASSO, Random Forest, Boruta, and Xgboost), univariate Cox regression, and a Gaussian mixture model (GMM) were utilized to mine the optimal combination of lncRNAs.

LASSO

LASSO was performed because it has powerful advantages in handling high-dimensional data and solving multicollinearity problems (Liu et al. 2023).

$$Q(\beta) = \|y - X\beta\|^2 + \lambda\|\beta\|_1$$

$$\Leftrightarrow \arg min\|y - X\beta\|^2 s.t. \sum \beta_j^2 \leq s$$

Here, $Q(\beta)$ represents the error vector, and we expected it to be as small as possible (so that we lose as little sample information as possible). β represents the features (lncRNAs). λ is generally obtained by cross-validation (CV).

Random Forest

Random Forest (RF) has the ability to analyze complex interaction classification features, has good robustness to noise data and data with missing values, and has a fast-learning speed (Toth et al. 2019; Zhang et al. 2022b).

$$GI_q^{(i)} = \sum_{c=1}^{|C|} \sum_{c\prime \neq c} p_{qc}^{(i)} p_{qc\prime}^{(i)} = 1 - \sum_{c=1}^{|C|} \left(p_{qc}^{(i)}\right)^2$$

Here, GI (Gini) value was used to calculate the contribution of each feature (lncRNA). i represents the decision tree, and q represents the node in the decision tree. C reflects the classified category. p_{qc} indicates the percentage of node q in C.

The variable importance measure (VIM) stands for the importance of each feature (lncRNA). The VIM of the feature in node q and tree i can be calculated by the following formula:

$$VIM_{jq}^{(Gini)(i)} = GI_q^{(i)} - GI_l^{(i)} - GI_r^{(i)}$$

where $GI_l^{(i)}$ and $GI_r^{(i)}$ represent the GI values of the two nodes before and after the branch, respectively.

We assume Q is the collection of all nodes that feature X_j present in the decision tree i, then the VIM of the feature X_j in decision tree i can be calculated by the following formula:

$$VIM_j^{(Gini)(i)} = \sum_{q \in Q} VIM_{jq}^{(Gini)(i)}$$

We assume *I* is the collection of all trees in RF, then the *VIM* of the feature can be calculated by the following formula:

$$VIM_j^{(Gini)} = \sum_{i=1}^{I} VIM_j^{(Gini)(i)}$$

Finally, we normalized the *VIM* of the feature:

$$normalized\, VIM_j^{(Gini)} = \frac{VIM_j^{(Gini)}}{\sum_{j\prime}^{J} VIM_{j\prime}^{(Gini)}}$$

In this study, the randomForestSRC R package (version: 3.2.1) was used for feature selection based on the RF algorithm.

Boruta

In general, the goal of feature selection in machine learning is to filter out the features that minimize the cost function of the current model. However, Boruta's feature selection aims to filter out all features correlated with the dependent variable (Wallentin et al. 2021). The significance of the Boruta algorithm is that it can help us understand the influencing factors of the dependent variable more comprehensively so as to perform feature selection better and more efficiently.

Real features:

$$Z_{score} = \frac{average\, feature\, important}{SE(feature\, important)}$$

Shuffled features:

$$MZSA = \max\{Z_{score}\}$$

$$\begin{cases} feature\, is\, important\, if\, Z_{score} \geq MZSA \\ feature\, is\, not\, important\, if\, Z_{score} < MZSA \end{cases}$$

In this study, the Boruta R package (version: 8.0.0) was used for performing feature selection based on the Boruta algorithm.

Xgboost

Xgboost is an extreme gradient boosting algorithm based on GBDT (Chai et al. 2021; Jiang et al. 2021; Hu et al. 2022). It has the characteristics of high efficiency, flexibility, and portability, which makes it has been widely used in data mining, recommendation system, and other fields.

The contribution degree of the feature is defined by the following formula:

$$V(k) = \frac{\sum_{t=1}^{T}\sum_{i=1}^{N(t)} I(\beta(t,i) = k)H_{\gamma(t,i)}}{\sum_{t=1}^{T}\sum_{i=1}^{N(t)} I(\beta(t,i) = k)}$$

Here, k represents the feature, T indicates the total number of trees, $N(t)$ represents the number of non-leaf nodes, $\beta(t,i)$ represents the divide signature of the i non-leaf node in t tree. $H_{\gamma(t,i)}$ represents the sum of second derivatives for the i non-leaf node in t tree from all samples. $I(\beta(t,i) = k)$ is an indicator function.

In this study, the Xgboost R package (version: 1.7.5.1) was used for performing feature selection based on the Xgboost algorithm.

Gaussian Mixture Model

Gaussian mixture model (GMM) refers to the linear combination of multiple Gaussian distribution functions (Zhang et al. 2022a). The IRL presented at least twice among four machine learning-based models were fed into GMM.

The distribution of GMM is defined as:

$$P_M(x) = \sum_{k=1}^{K} \pi_k \bullet N\left(x|\mu_k, \Sigma_k\right)$$

$$\sum_{k=1}^{K} \pi_k = 1, \pi_k > 0$$

Here, the distribution consists of K mixture components. μ is an n dimensional mean vector, Σ is a $n \times n$ covariance matrix, and π is the corresponding mixture coefficient.

GMM was used for clustering analysis. We assumed that the sample data obeys the mixed Gaussian distribution, the parameters of the mixed Gaussian distribution are deduced from the sample dataset, and which Gaussian distribution each sample is most likely to belong to. In our study, there are 4095 ($2^{12} - 1$) models consisting of 12 prognostic-related HKLncRNA signatures associated with immune regulation, up regulated in immune compared to cancer cells, and benefit for the prediction of outcomes of colon cancer patients.

The criteria for selecting the optimal IRL is based on the best prediction performance with minimum consumption.

The Construction of Risk Scoring System

The combination of IRL with the highest AUC and least number was used to construct the final risk score model. To predict the outcome of colon cancer patients, a predictive model based on the expression levels of IRL and clinical information as follows:

$$Risk\,Score = \sum\nolimits_{i=1}^{N} Exp(LncRNA) \times i\beta$$

where N represents the number of IRL, and $i\beta$ represents correlation coefficient values for its corresponding lncRNA in the multivariate Cox regression model.

This multivariable Cox regression model was built by coxph() and step() functions in the survival R package (version: 3.5.1).

Performance of Risk Score Model

The receiver operating characteristic curve (ROC) is an analysis tool for assessing the sensitivity and specificity of our model. The range of area under the curve (AUC) is the indicator of ROC. The closer it is to 1, the better the model performs.

After constructing the risk score model, the risk score model was validated on an independent dataset in terms of C-index value, tROC, and ROC. TCGA-COAD was treated as an independent dataset for validating the robustness of our risk score model. Survival analysis was performed using the survival R package (version: 3.3.1). Time-dependent ROC analysis was performed by the timeROC R package (version: 0.4).

Univariate and Multivariate Cox Regression

The univariate Cox regression model was utilized for the selection of prognostic IRL. Obtained IRL served as individual factors associated with colon cancer patients' outcomes by combining clinical characteristics. Then, multivariate Cox regression was utilized to identify independent prognostic factors among clinical traits and IRLs.

Two R packages, survival (version: 3.1.1) and forestplot (version: 3.1.1), were used to select independent prognostic factors and visualize.

Other Statistical and Bioinformatics Analysis

All statistical analysis was performed in R language (version: 4.2.2). The gsva() function in the GSVA R package (version: 1.44.5) was applied to immune infiltration analysis, which calculates the enrichment score of each immune cell for each patient using the ssGSEA algorithm. Immune cell types and gene sets in each immune cell type were collected from Pan-Cancer research, which includes 28 immune cell types B cell, CD4 T cell, CD8 T cell, dendritic cell, CD56 bright natural killer cell, CD56 dim natural killer cell, Central memory CD4 T cell, Central memory CD8 T cell, Effector memory CD4 T cell, Effector memory CD8 T cell, Eosinophil, Gamma delta T cell, Immature B cell, Immature dendritic cell, MDSC, Macrophage, Mast cell, Memory B cell, Monocyte, Natural killer T cell, Natural killer cell, Neutrophil, Plasmacytoid dendritic cell, Regulatory T cell, T follicular helper cell, Type 1 T helper cell, Type 17 T helper cell, Type 2 T helper cell 23 (Charoentong et al. 2017). To investigate the tumor immune microenvironment, we compare the expression levels of lncRNA in the risk model across five immune subtypes, including C1 (wound healing), C2 (IFN-gamma dominant), C3 (inflammatory), C4 (lymphocyte depleted), and C6 (TGF-beta dominant) (Thorsson et al. 2019).

Wilcoxon and ANOVA test was utilized for statistically significant analysis.

Results

The Expressed lncRNAs in Human Immune Cells

The overall schematic workflow is shown in Fig. 1. After the lncRNA annotation process (match probe id to lncRNA gene symbol), 1422 lncRNAs were kept. In order to determine the expression pattern of lncRNAs in human immune cells, we ranked lncRNAs according to their expression levels, from high to low, for each immune cell type. The top 30% of highly expressed lncRNAs for each immune cell type were merged and removed duplicates. Then 737 lncRNAs were treated as immune-related lncRNAs and kept for the following analysis.

The top 30% of highly expressed lncRNAs were obtained as the candidate IRL for each immune cell type. The specificity of expression of a candidate IRL with respect to different immune cell types was calculated using the specificity index. Those HKLncRNA, which are significantly up-regulated in immune samples compared to colon cancer samples, were selected as IRL. We analyzed GEO dataset GSE39582 and TCGA-COAD with four machine learning algorithms and a Gaussian mixture model to screen out the optimal combination of lncRNAs: LINC00638, CYB561D2, and DANCR. A prognostic signature was constructed using the linear combination of the expression values of the prognostic IRL, weighted by their estimated regression coefficients in the multivariate Cox regression analysis. The model has a satisfactory performance and was validated by an independent dataset, C-index, ROC, and tROC. Finally, we explored the difference in immune cell types between low- and high-risk score groups.

HKLncRNAs in Human Immune Cells

We calculated these 737 lncRNAs' expression levels across 19 immune cell types. By introducing a tissue specificity index value and setting cutoff values, we identified 221 HKLncRNAs (Supplementary Table 2) and 221 cell type-specific lncRNAs in immune cells (Supplementary Table 3). Supplementary Fig. 1 shows the heatmap of housekeeping and cell type-specific genes' expression profile across all immune cell types.

HKLncRNAs in immune cells is a type of constitutive gene which have an essential role in the maintenance of cellular immune function.

LncRNAs That Up-Regulated in the Immune Compared to Colon Cancer Cells

By combining microarray data in immune and cancer cells, we conducted DEG analysis using SAM() function in samr R package. There are 87 HKLncRNAs

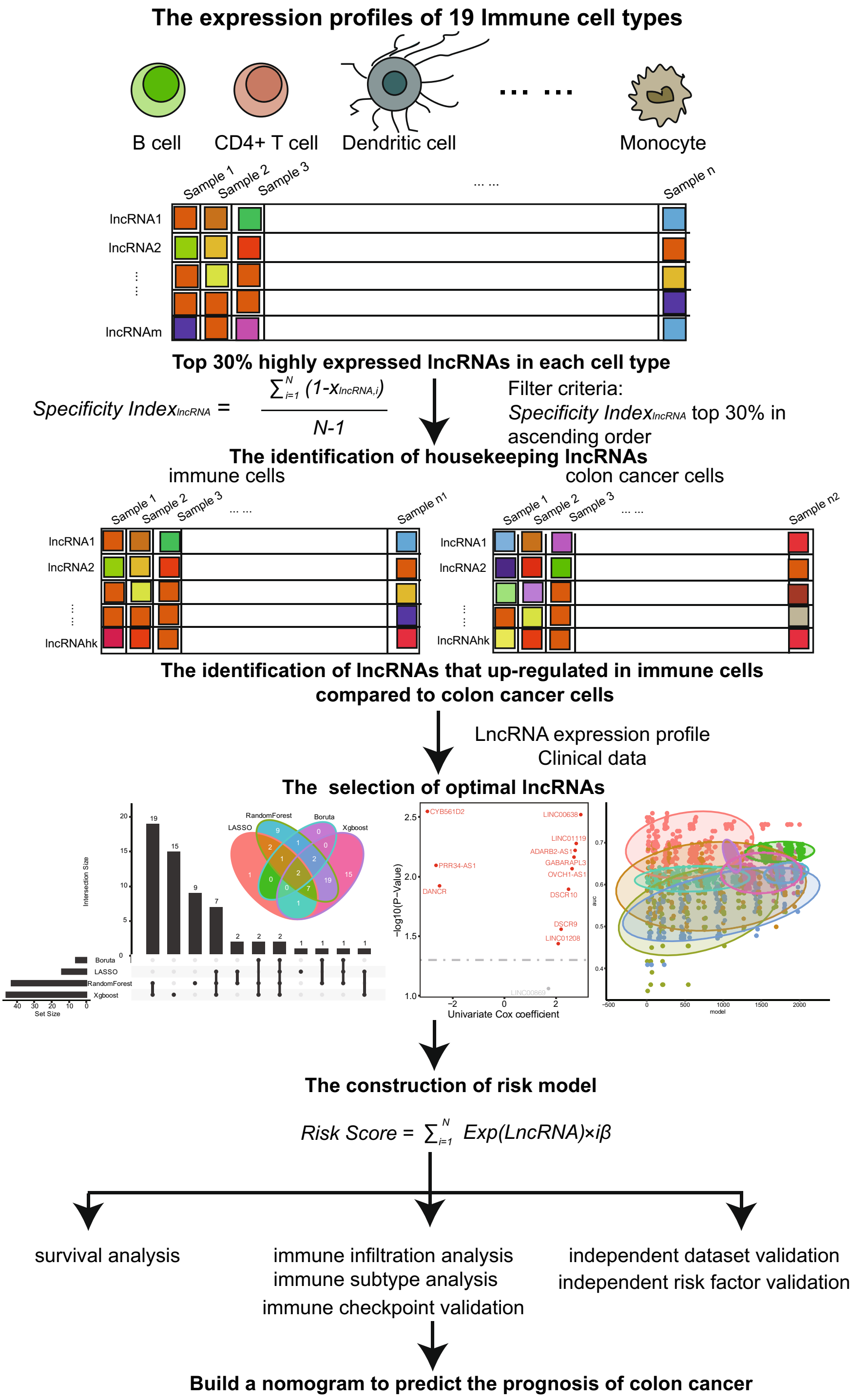

The expression profiles of 19 Immune cell types
B cell
CD4+ T cell
Dendritic cell
... ...
Monocyte
Sample 1
Sample 2
Sample 3
... ...
Sample n
lncRNA1
lncRNA2
lncRNAm
Top 30% highly expressed lncRNAs in each cell type
Specificity Index_lncRNA = ∑_{i=1}^{N} (1-x_lncRNA,i) / N-1
Filter criteria:
Specificity Index_lncRNA top 30% in ascending order
The identification of housekeeping lncRNAs
immune cells
colon cancer cells
Sample n1
Sample n2
lncRNAhk
The identification of lncRNAs that up-regulated in immune cells compared to colon cancer cells
LncRNA expression profile
Clinical data
The selection of optimal lncRNAs
Intersection Size
Set Size
Boruta
LASSO
RandomForest
Xgboost
Univariate Cox coefficient
−log10(P−Value)
CYB561D2
PRR34-AS1
DANCR
LINC00638
LINC01119
ADARB2-AS1
GABARAPL3
OVCH1-AS1
DSCR10
DSCR9
LINC01208
LINC00868
auc
model
The construction of risk model
Risk Score = ∑_{i=1}^{N} Exp(LncRNA)×iβ
survival analysis
immune infiltration analysis
immune subtype analysis
immune checkpoint validation
independent dataset validation
independent risk factor validation
Build a nomogram to predict the prognosis of colon cancer

Fig. 1 The schematic workflow for identifying IRLs enhances colon cancer patients' prognosis

significantly up-regulated in the immune compared to colon cancer cells that were kept for the following feature selection analysis (Supplementary Table 4).

The Prognostic and Optimal lncRNAs

14, 43, 6, and 91 prognostic lncRNAs were mined by four machine learning-based models, LASSO, RF, Boruta, and Xgboost, separately (Supplementary Fig. 2). Finally, 12 lncRNAs identified twice (Fig. 2A, Supplementary Table 5). 11 lncRNAs were confirmed that there are associated with the prognosis of colon cancer patients using univariate Cox regression analysis (Fig. 2B).

GMM was utilized for identifying the optimal combination of lncRNAs. Three lncRNAs, *CYB561D2*, *LINC00638*, and *DANCR*, are identified as optimal lncRNAs that are related to the prognosis of colon cancer with the maximum AUC = 0.770 (Fig. 2C). There is another combination of lncRNAs, *CYB561D2*, *LINC00638*, *DANCR*, and *LINC01208*, which have the same maximum AUC value. However, the number of lncRNAs is equal to four. Considering that the more detection index, the more cost will be taken, we choose three lncRNAs instead of four as the optimal combination.

A Scoring System Based on Immune-Related and Prognostic lncRNAs

A scoring system based on immune-related and prognostic lncRNAs that can be used to detect the prognosis of colon cancer patients is constructed by multivariate Cox regression. Table 2 gives the detailed coefficient value of the above three lncRNAs. The immune-related and prognostic lncRNA signature can be calculated by the following formula:

$$\begin{aligned} lncRNA\ signature = & -0.356 \times Exp(CYB561D2) \\ & + 0.830 \times Exp(LINC00638) \\ & - 0.170 \times Exp(DANCR) \end{aligned}$$

Risk Score Model Is an Evaluation Indicator for Clinical Outcome

Multivariate Cox regression was constructed to distinguish the patients into two groups when setting the mean value of all risk scores as the cutoff. We defined the patient as high-risk if the score of the patient is larger or equal to the average value of all patients' risk scores. On the contrary, we defined the patient as low-risk if the score of the patient is lower than the average value of all patients' risk scores. Figure 3A is the distribution of patients' age, gender, tumor stage, and survival status in low- and high-risk groups. It demonstrated that there were significant differences in tumor stage and survival status between the low- and high-risk groups, but there were no significant differences in age and gender between the low- and high-risk groups. Figure 3B shows a significant difference between low- and high-risk groups (P-value < 0.05), which hints that the low-risk group has a longer overall survival

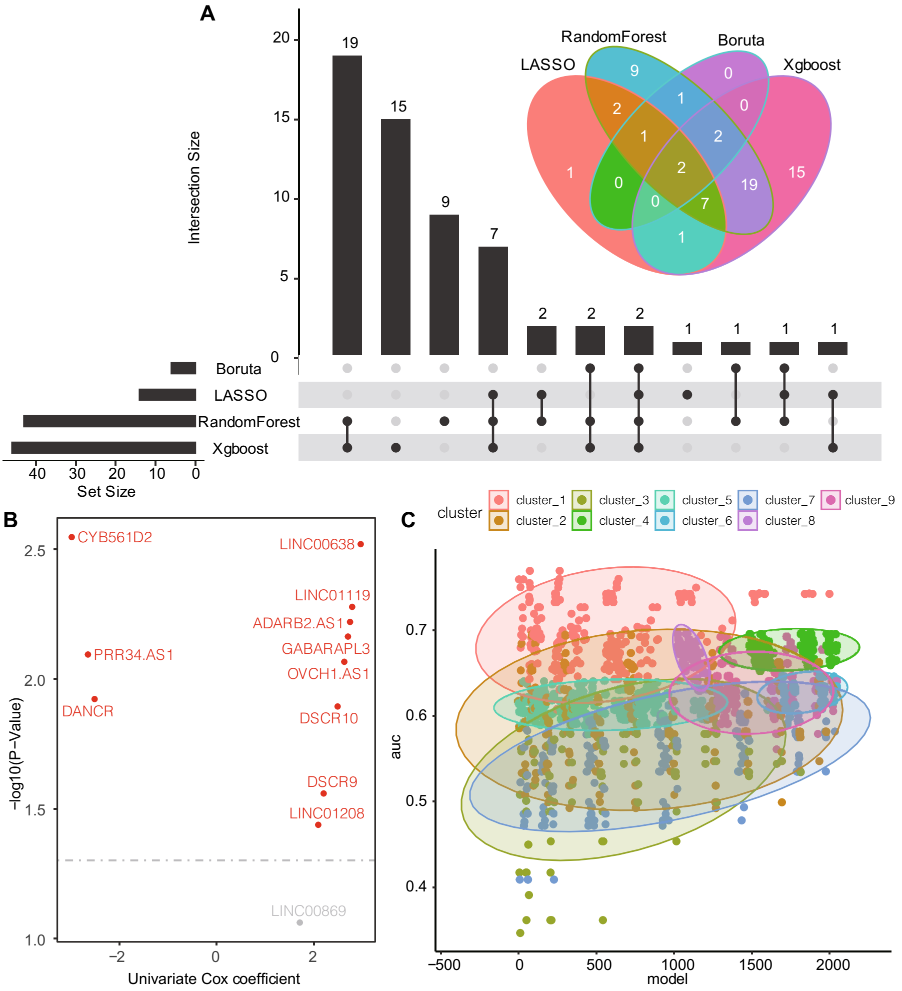


Fig. 2 The identification of prognostic and optimal lncRNAs. **A** The overlap lncRNAs among four different machine learning-based models, LASSO, RF, Boruta, and Xgboost. **B** Prognosis-related lncRNAs inferred by univariate Cox coefficient regression. **C** GMM model was conducted to identify the optimal combination of lncRNAs

Table 2 The coefficient of the three lncRNAs in the multivariate Cox regression model

LncRNA	Coefficient value	HR	HR.95L	HR.95H
CYB561D2	– 0.356	0.700	0.508	0.965
LINC00638	0.830	2.294	1.275	4.129
DANCR	– 0.170	0.844	0.670	1.063

The first column indicates the immune-related and prognostic lncRNA, the second column indicates the coefficient value corresponding to its lncRNA, the third column indicates the HR value and the last two columns show the HR value's low and high 95% confidence interval

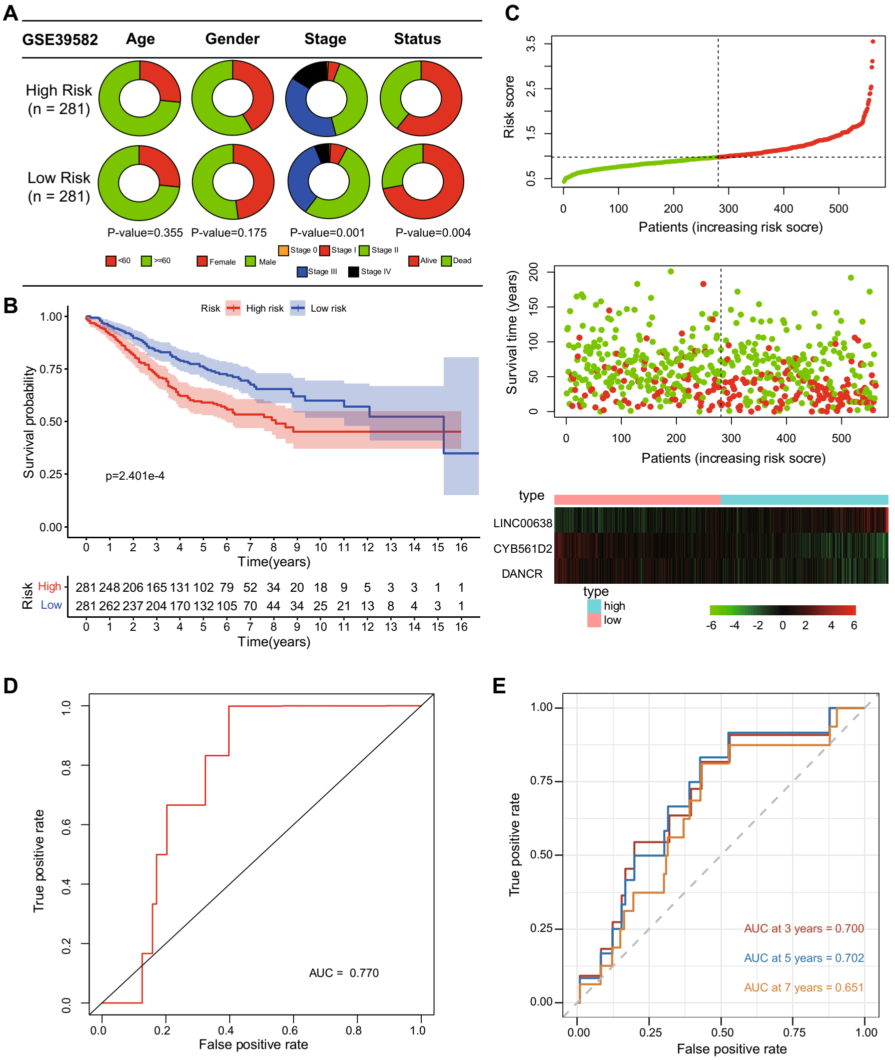


Fig. 3 The application of immune-related and prognostic lncRNA signature into the training dataset. **A** The pie chart displayed the difference in patients' age, gender, tumor stage, and survival status between low- and high-risk groups. **B** The survival curve reflected that there is a significant difference between low- and high-risk cohorts. **C** The risk plot demonstrated the relationship among risk score, survival time, and lncRNA expression levels. **D** The ROC curve of this scoring system. The AUC is equal to 0.770. **E** The time-dependent ROC results for 3-, 5-, and 7-years of this scoring system. AUC $_{\text{3-years}}$ = 0.700, AUC$_{\text{5-years}}$ = 0.702, and AUC$_{\text{7-years}}$ = 0.651

time than the high-risk group. Figure 3C is the visualization of the relationships among the risk score of patients, survival time, and the expression levels of three lncRNAs in the risk score model. From the risk plot, we concluded that high-risk score patients are associated with shorter survival time compared to low-risk score patients. In addition, the expression level of *LINC00638* is positively related to the risk score, while the expression levels of *CYB561D2* and *DANCR* are negatively related to the risk score. AUC is an index for describing the performance of the risk score model. Our scoring model has a high correction ratio (Fig. 3D). The high accuracy indicates our model can be well used to distinguish the prognostic effect of training samples. Also, the time-dependent ROC (tROC) supported that our provided scoring system has a satisfactory performance at 3-, 5-, and 7-years (Fig. 3E).

The Risk Score Model Has a Good Robustness

TCGA-COAD dataset was treated as an independent validating dataset. Our provided scoring system was also applied to this independent dataset. The C-index and SE (C-index) values of the training dataset (GSE39582, self-validation) are 0.599 and 0.022, respectively. And the C-index and SE (C-index) values of the testing dataset (TCGA, independent validation) are 0.592 and 0.032, respectively (Fig. 4A). Thus, our scoring system has good robustness and the ability to avoid overfitting problems. Figure 4B indicates there is a significant difference between low- and high-risk score groups, and the survival time of high-risk score patients is significantly lower than low-risk score patients. The risk plot shows that the high-risk score group tends to have a shorter survival time compared to the low-risk score group (Fig. 4C). The heatmap reflects that the expression level of *LINC00638* is positively related to the risk score, while the expression levels of *CYB561D2* and *DANCR* are negatively related to the risk score (Fig. 4C). ROC and tROC were calculated to evaluate the effect of the model prediction. Figure 4D and E exhibited that this scoring model has a high accuracy of prediction.

Immune Infiltration-Related lncRNA Signature Is an Independent Prognostic Factor

To determine which clinical characteristics are associated with survival time, each clinical trait is compared individually with survival time and survival status. The results showed that patients' age, TNM_t, TNM_n, TNM_m, and risk score are significantly related to survival, while patients' gender is not significantly related to survival (Fig. 5A). Further, we simultaneously considered all clinical factors to survival for identifying independent prognostic factors. Finally, patients' age, gender, TNM_t, TNM_n, TNM_m, and risk score can serve as six independent prognostic factors in colon cancer (P-value < 0.05), while TNM_stage can be represented by other clinical factors (Fig. 5B).

Immune-Related and Prognostic lncRNA Signature Is Associated with Immune Cell Infiltration

Further, we explored this kind of disorder in the immune infiltration levels. We considered 28 different kinds of immune cell types in the analysis. We found that patient risk groups stratified by IRL signature showed distinct immune infiltration patterns. As shown in Fig. 6A, patients in the low-risk group were enriched with six immune

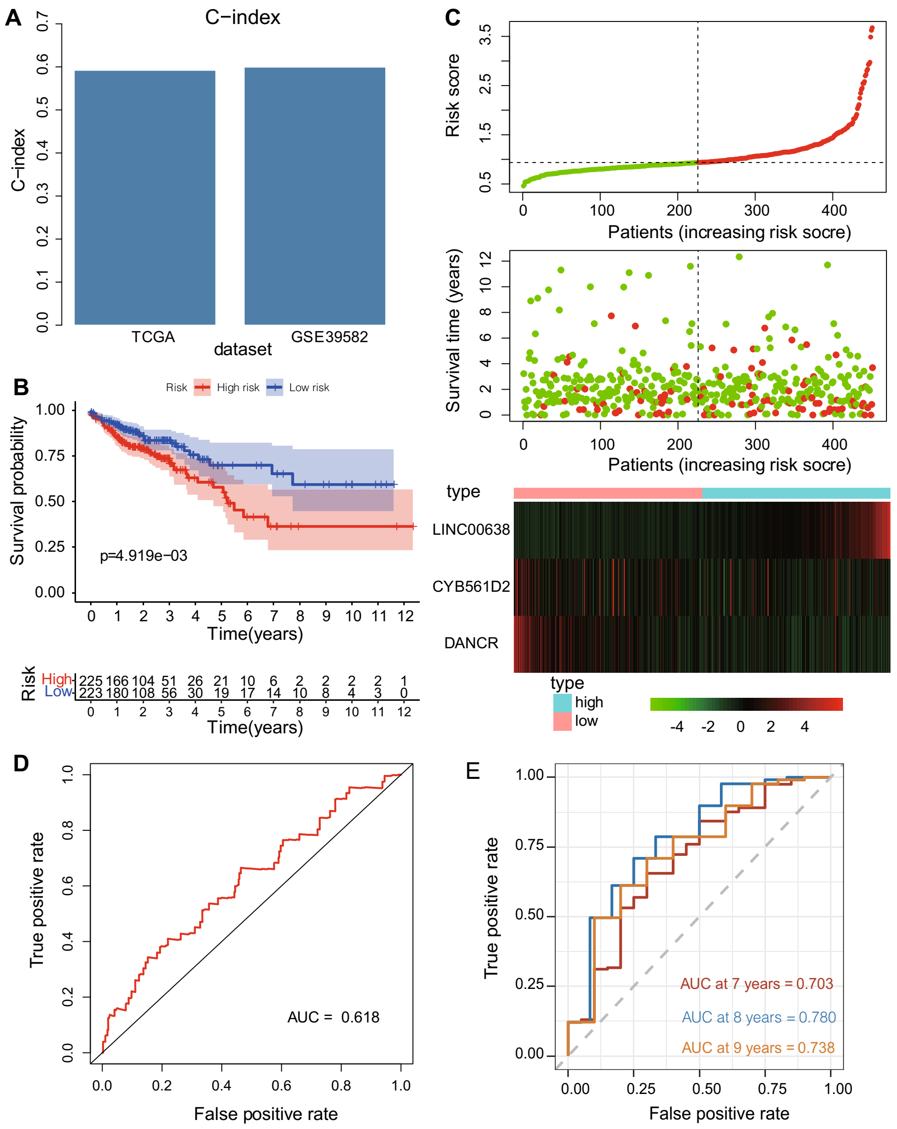
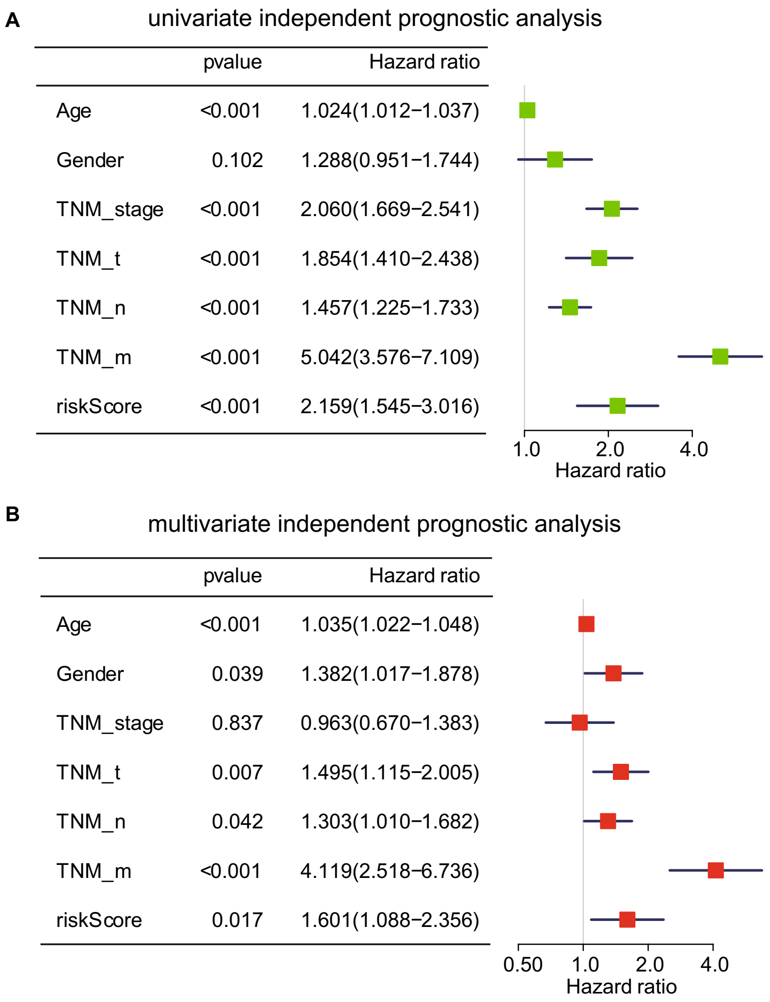


Fig. 4 The application of immune-related and prognostic lncRNA signature into the testing dataset. **A** The bar plot displayed the C-index values in TCGA (testing) and GSE39582 (training) datasets. **B** The survival curve reflected that there is a significant difference between low- and high-risk cohorts. **C** The risk plot demonstrated the relationship among risk score, survival time, and lncRNA expression levels. **D** The ROC curve of this scoring system. The AUC is equal to 0.618. **E** The time-dependent ROC results for 7-, 8-, and 9-years of this scoring system. $\mathrm{AUC}_{7\text{-years}}=0.703$, $\mathrm{AUC}_{8\text{-years}}=0.780$, and AUC $_{9\text{-years}}=0.738$

subpopulations, while only two immune subpopulations were enriched in patients with high risk (P-value < 0.01). These results suggested that the higher score of IRL corresponded to less immune cell infiltration and poor outcome, while lower score of



	pvalue	Hazard ratio
Age	<0.001	1.024(1.012−1.037)
Gender	0.102	1.288(0.951−1.744)
TNM_stage	<0.001	2.060(1.669−2.541)
TNM_t	<0.001	1.854(1.410−2.438)
TNM_n	<0.001	1.457(1.225−1.733)
TNM_m	<0.001	5.042(3.576−7.109)
riskScore	<0.001	2.159(1.545−3.016)

	pvalue	Hazard ratio
Age	<0.001	1.035(1.022−1.048)
Gender	0.039	1.382(1.017−1.878)
TNM_stage	0.837	0.963(0.670−1.383)
TNM_t	0.007	1.495(1.115−2.005)
TNM_n	0.042	1.303(1.010−1.682)
TNM_m	<0.001	4.119(2.518−6.736)
riskScore	0.017	1.601(1.088−2.356)

Fig. 5 Univariate and multivariate Cox regression analysis. **A** Univariate independent prognostic analysis. **B** Multivariate independent prognostic analysis

IRL corresponded to greater immune cell infiltration and better outcome. Figure 6A demonstrated significantly positive relationships between risk scores and immune cell types, including memory B cells, natural killer cells, macrophages, mast cells, etc. We further examined the risk score distribution among five immune subtypes reported by a recent study (Thorsson et al. 2019). Risk score showed a notable difference among five different immune subtypes (Fig. 6B). *LINC00638* (Fig. 6C) showed a notable difference among five kinds of immune subtypes (P-value < 0.05), while *CYB561D2* (Fig. 6D) and *DANCR* (Fig. 6E) didn't show a notable difference among five types of immune subtypes (P-value > 0.05). Because C1, C2, C3, C4, and C6 classification systems were from pan-cancer research (Charoentong et al. 2017),

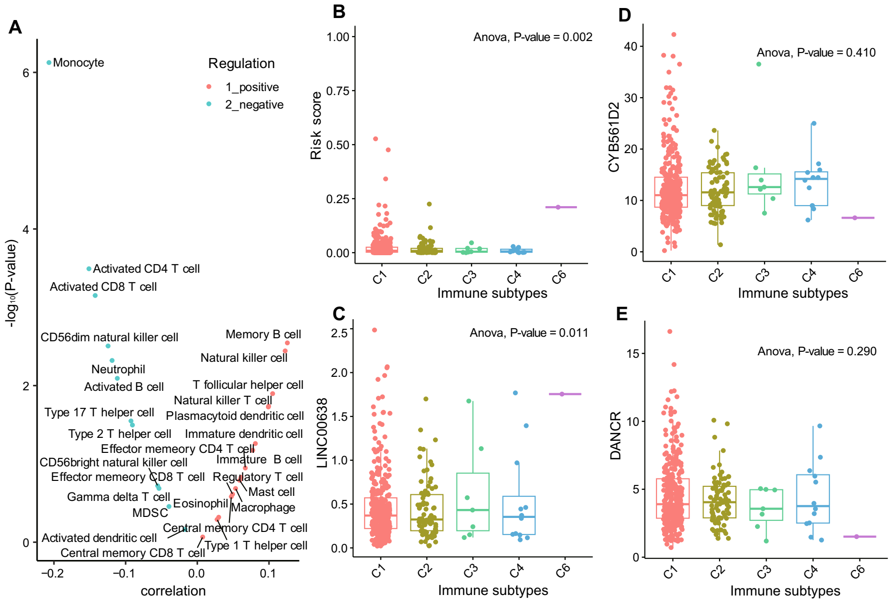


Fig. 6 The validation of immune-related and prognostic lncRNA signature in immune cell types. **A** The volcano plot represents the enrichment of immune cell types for colon cancer with positive and negatively associated risk scores calculated by normalized enrichment score (NES) score from gene set enrichment analysis (GESA). **B** The distribution of risk scores among five different kinds of immune cell types. The expression pattern of three lncRNAs, *LINC00638* (**C**), *CYB561D2* (**D**), and *DANCR* (**E**), in five immune cell subtypes. C1 = Wound healing; C2 = IFN-γ dominant; C3 = Inflammatory; C4 = Lymphocyte depleted; C6 = TGF-β dominant

we defined them as tumor immune environments. The combination of *LINC00638*, *CYB561D2*, and *DANCR* is closely associated with the tumor immune microenvironment. However, each lncRNA in the model might not be closely associated with the tumor immune microenvironment. The enrichment of 28 immune cell types with positive and negatively associated the expression levels of three IRL, *LINC00638*, *CYB561D2*, and *DANCR*, calculated by normalized enrichment score (NES) score from gene set enrichment analysis (GESA) was shown in Supplementary Fig. 3.

Immunotherapy Response Prediction

Researchers designed immunotherapy drugs based on the targeted immune checkpoint proteins. *PD-1* is an immune checkpoint receptor in T cells, which serves as a “switch-off” (Wang et al. 2022). When the *PD-1* binds to the *PD-L1* of tumor cells, T cells will not attract tumor cells (Yu et al. 2022). By developing an inhibitor for *PD-1*, we can extensively block the combination between *PD-1* and *PD-L1*, further enhancing immune response. *CTLA-4* (also known as *CD152*) is constitutively expressed in regulatory T cells (Wang et al. 2022). In cancer cells, *CTLA-4* is upregulated after the immune system activation (Iranzo et al. 2022).

Thus, we explored the expression levels of *PD-1*, *PD-L1*, and *CTLA-4* between low- and high-risk score groups. We didn't identify a significant difference between low- and high-risk score groups for *PD-1* (Fig. 7A) and *CTLA-4* (Fig. 7E). However, a significant difference exists between low- and high-risk score groups for *PD-L1* was observed (Fig. 7C). Recently, researchers reported that high PD-L1 expression on tumor cells indicates the presence of an anti-tumor immune response (Sorensen et al. 2016; Shah et al. 2022). This is consistent with our conclusion. This phenomenon indicates that the model helps to propose drugs related to immunotherapy.

Nomogram-Based Survival Prediction

A comprehensive model, including immune-related and prognostic lncRNA signature and clinical characteristics, was developed and displayed as the nomogram (Fig. 8A). Its prognosis reliability was established by the calibration examination (Fig. 8B, D, and F). The nomogram demonstrated viability for 3-, 5-, and 10-year survival under control, as indicated by the Decision Curve Analysis (DCA) curve (Fig. 8C, E, and G). These results demonstrated that the nomogram achieved favorable predictive performance.

Discussion

Colon cancer, also known as colorectal cancer, is one kind of cancer that develops from colon or rectum (Labianca et al. 2010; Terzić et al. 2010). The mechanism of colon cancer development and progression is still unclear. Recent studies have found that tumor immune cell infiltration is associated with cancer development and may adversely affect cancer prognosis (Kong et al. 2022; Wei et al. 2022). Accumulating evidence demonstrated that lncRNAs play essential roles in the immune response by participating in cancer progression; for example, Wu et al. revealed that most marker genes of immune cells showed a significant correlation with *LINC00665* (Wu et al. 2020a). Especially the expression of *LINC00885* has a positive relationship with marker genes of M2 macrophages (Wu et al. 2020a). However, a systematic model for identifying immune-related lncRNAs is currently lacking. Therefore, we aim to develop a risk score model to mine regulatory lncRNAs in the colon cancer immune microenvironment.

Firstly, we obtained the top 30% highly expression lncRNAs from gene expression profiles of 19 immune cell types. Two-hundred and twenty-one HKLncRNAs and Two-hundred and twenty-one cell type-specific lncRNAs in immune cells were screened by calculating the cell type specificity index. The results can be validated by manually reviewing publications. For example, *CYB561D2* and *EHIH* are two HKLncRNAs in immune cells. *CYB561D2* encoded cytochrome B561 family member D2 participating in ion metabolism and stress defense pathways (Sananmuang et al. 2020). Sordillo et al. reported that oxidative stress is a major underlying reason for inflammatory dysfunction (Sordillo and Aitken 2009). Sun et al. validated that *EHIH* is a diagnostic and prognostic biomarker in pan-cancer, and it is involved in an immune-oncogenic system combined with *YBX3*, particularly for colon cancer

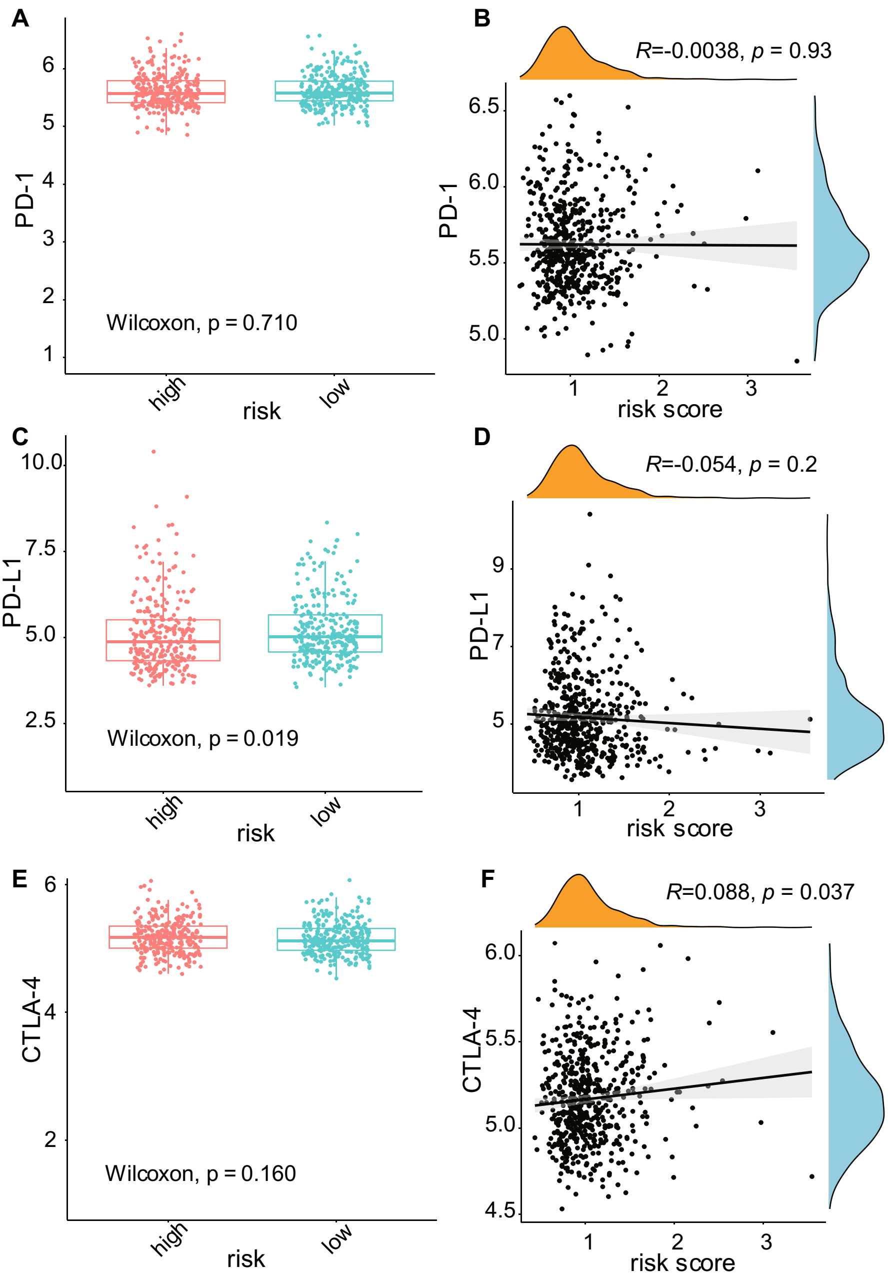


Fig. 7 The relationship between immune checkpoint genes (PD-1, PD-L1, and CTLA-4) and risk score. The distribution of normalized expression levels of the PD-1 (**A**), PD-L1 (**C**), and CTLA-4 (**E**) across low- and high-risk score groups. The correlation between the normalized expression levels of three immune checkpoint genes [PD-1 (**B**), PD-L1 (**D**), CTLA-4 (**F**)] and risk score

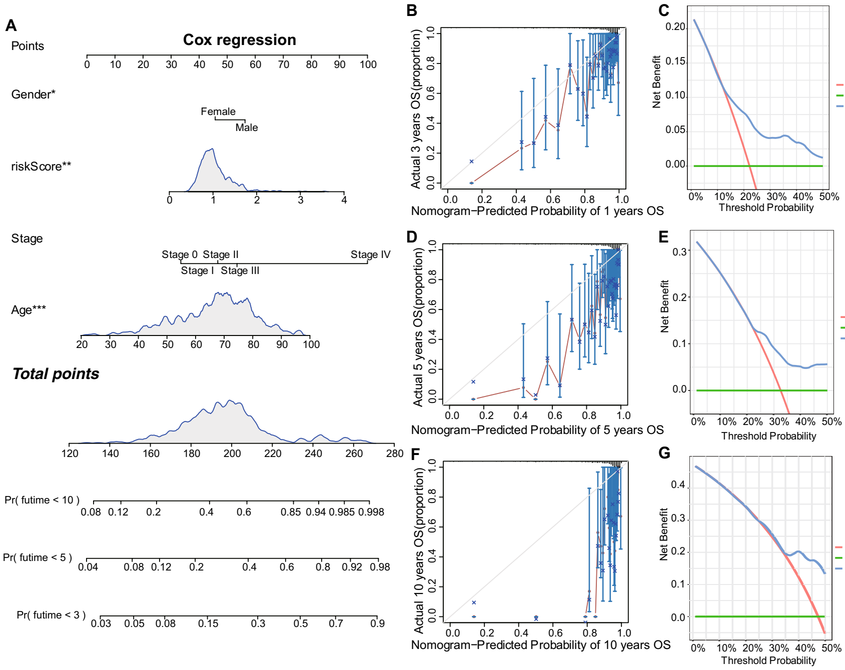


Fig. 8 The construction and validation of a nomogram for predicting the survival of colon cancer patients. **A** The nomogram is displayed to estimate the results for CRC patients. The prognostic risk score model is presented in a visualization pattern. The nomogram is shown to assess the outcome of CRC patients. It is a readable style of the prognostic risk score model. The patient survival calibration curve at 3- (**B**), 5- (**D**), and 10-year (**F**) years. The x-axis displays the OS probability predicted by the nomogram at that time, while the y-axis displays the actual data at different timepoint. The optimum prediction is shown by the 45° grey line. The training dataset is represented by the dots, while the validation dataset is represented by the curve line. The figure also included a label for the 95% CI. Nomogram DCAs for the OS at the 3- (**C**), 5- (**E**), and 10-year (**G**) intervals

(Sun et al. 2022). Then, we identified 87 lncRNAs that are up-regulated in immune samples and down-regulated in colon cancer samples, which demonstrated their expression specificity to immune cells rather than tumor cells. We aimed to identify lncRNAs that can be used as biomarkers to improve colon cancer's prognosis and patients' immunotherapy response. These lncRNAs were treated as specificity expression in immune cells compared to tumor cells. Combining with clinical traits information, we got lncRNAs that were significantly associated with the survival time of colon cancer patients. These lncRNAs were incorporated into four machine learning-based algorithms, LASSO regression analysis, RF, Boruta, and Xgboost. Twice lncRNAs identified from four methods are *CYB561D2*, *PRR34-AS1*, *DANCR*, *LINC00638*, *LINC01119*, *ADARB2-AS1*, *GABARAPL3*, *OVCH1-AS1*, *DSCR10*, *DSCR9*, *LINC00869*, and *LINC01208*. Only *LINC00869* was neglected because there is no significant relationship between expression level and prognosis of colon patients. For maximum prediction accuracy and minimum cost, three immune-related and prognostic lncRNAs, *LINC00638*, *CYB561D2*, and *DANCR*, were used to create the risk score model.

Then, multivariate regression analysis was conducted to construct a risk score model. Our linear mathematically combinational model was validated by survival analysis, ROC, tROC, C-index, risk plot analysis, and independent dataset analysis. In addition, there is an immune checkpoint gene named *PD-L1* that is expressed differently between low- and high-risk score groups, which suggests that our model can be used to assess patient's immune response. Tao et al. reported *CYB561D1* is up-regulated in glioma samples compared to normal samples (Tao et al. 2021). And they concluded that the over-expression of *CYB561D1* is associated with a short survival time of high-grade glioma (Tao et al. 2021). The mechanism is that the over-expression of *CYB561D1* increased the expression of *CCL2* and *PD-L1* and triggered immunosuppression in T cells by activating the *STAT3* signaling pathway (Tao et al. 2021). At the same time, *LINC00638/miR-4732-3p/ULBP1* is a lncRNA-related competitive endogenous RNA (ceRNA) network, which is highly associated with immune filtration and tumor mutation burden (TMB) in hepatocellular carcinoma (HCC) (Qi et al. 2021). In HCC with elevated TMD, *LINC00638/miR-4732-3p/ULBP1* is a prognostic predictor and controls immunological escape via PD-L1 (Qi et al. 2021). *LINC00638*/hsa-miR-29b-3p/CDCA4 is a candidate regulatory network in liver hepatocellular carcinoma (LIHC). Tumor immune evasion and anti-tumor immunity may play a role in *CDCA4*-mediated LIHC carcinogenesis (Wang et al. 2023). The prognosis of LIHC patients is dramatically improved by low *CDCA4* expression, and *CDCA4* is a promising novel biomarker for predicting LIHC prognosis. Tumor immune evasion and anti-tumor immunity may play a role in *CDCA4*-mediated LIHC carcinogenesis (Wang et al. 2023). In recent years, more and more studies have been done on using *DANCR* as a biomarker to predict colon cancer prognosis (Yang et al. 2018; Shi et al. 2020; Sun et al. 2020b). *DANCR* was extensively expressed in colon cancer tissue and cell lines (Sun et al. 2020b). Sun et al. reported that higher levels of *DANCR* were associated with a poorer prognosis and shorter patient survival time for colon cancer (Sun et al. 2020b). Cell proliferation and colony formation were drastically reduced when *DANCR* was silenced by short-interfering RNA (siRNA) (Yang et al. 2018). Although immune checkpoint inhibitors targeting *PD-1*, *PD-L1*, and *CTLA-4* have been developed to treat cancer and improve survival time (Sun et al. 2020a, b). However, the immune responses of different patients are not the same due to the heterogeneity of the tumor immune environment. Our findings revealed that the lncRNAs have complex crosstalk between tumor cells and immune cells. Low-risk group patients have high expression of *PD-L1* and longer survival time compared to high-risk group patients.

Compared to published studies (Toth et al. 2019; Wallentin et al. 2021; Jiang et al. 2021; Chai et al. 2021; Zhang et al. 2022b, a; Hu et al. 2022), our framework has the advantage of considering multiple machine learning-based methods. We proposed that immune-related and prognostic lncRNAs have a great potential to predict the survival of colon cancer patients based on the linear regression model. The feature selection process is achieved by combing four machine learning methods, keeping prognosis-related lncRNAs, and selecting the optimal combination of lncRNAs from 4095 combinations using GMM. We provided a nomogram with the maximum performance and minimum cost. This framework is very helpful for distinguishing patients into two groups and gives different treatment strategies not only for colon cancer but also can be used for other cancers. However, this study also has some limitations. First of all, this study is a retrospective experiment. Second, the mechanism of *DANCR* is not fully explained in the previous research. Further experiments should be conducted to validate our model.

In conclusion, we give an immune-related and prognostic lncRNA signature by combining transcriptome data and clinical data. This signature can be validated by ROC, tROC, C-index, independent dataset, and literature. It has a good potential to predict the outcome of colon cancer patients. Applying this model to colon cancer patients, we can discover that the tumor immune microenvironment is different between low- and high-risk score groups, which is beneficial for immunotherapy and precision medicine.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10528-023-10516-4.

Acknowledgements The author would like to thank the computer science laboratory (CSLab) from City University of Hong Kong (CityU) for the help with the computational resource, such as, high throughput CPU cluster (HPCC) support.

Author Contributions All authors contributed to the study conception and design. ZL: Resources, Visualization, Writing-original draft, Writing-review, and editing. OOP: Writing-original draft, Writing-review, and editing. MT: Visualization and Writing-original draft. NC: Writing-review. and KCW: Conception, Supervision, Fund acquisition. All authors read and approved the final manuscript.

Funding This research was substantially sponsored by the research projects (Grant Nos. 32170654 and 32000464) supported by the National Natural Science Foundation of China and was substantially supported by the Shenzhen Research Institute, City University of Hong Kong. The work described in this paper was substantially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region [CityU 11203723]. This project was substantially funded by the Strategic Interdisciplinary Research Grant of the City University of Hong Kong (Project No. 2021SIRG036). The work described in this paper was partially supported by a Grant from the City University of Hong Kong (CityU 9667265).

Data Availability The transcriptome RNA-seq and clinical colon carcinoma data was generated by TCGA Research Network: https://www.cancer.gov/tcga. Microarray colon cancer data and its corresponding clinical data were downloaded from the GEO database (https://www.ncbi.nlm.nih.gov/geo/) with the accession number GSE39582. Nineteen immune cell types data were downloaded via GSE13906, GSE23371, GSE25320, GSE27291, GSE27838, GSE28490, GSE28698, GSE28726, GSE37750, GSE39889, GSE42058, GSE49910, GSE51540, GSE59237, GSE6863, and GSE8059.

Code Availability All code for data cleaning and analysis associated with the current submission is available at https://github.com/liuzhe93/ImmuneInfiltration_Related_LncRNA.

Declarations

Competing interest The authors have no relevant financial or non-financial interests to disclose.

Ethical Approval The microarray, RNA sequencing, and clinical characteristics data are publicly available from TCGA and GEO databases. We strictly followed the user guideline from the data source.

Consent for Publication Not applicable.

References

Barnes TA, Amir E (2018) HYPE or HOPE: the prognostic value of infiltrating immune cells in cancer. Br J Cancer 118:e5. https://doi.org/10.1038/bjc.2017.417

Barrett T, Wilhite SE, Ledoux P et al (2013) NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res 41:D991–D995. https://doi.org/10.1093/nar/gks1193

Bense RD, Sotiriou C, Piccart-Gebhart MJ et al (2017) Relevance of tumor-infiltrating immune cell composition and functionality for disease outcome in breast cancer. J Natl Cancer Inst. https://doi.org/10.1093/jnci/djw192

Chai H, Zhou X, Zhang Z et al (2021) Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. Comput Biol Med 134:104481. https://doi.org/10.1016/j.compbiomed.2021.104481

Charoentong P, Finotello F, Angelova M et al (2017) Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. Cell Rep 18:248–262. https://doi.org/10.1016/j.celrep.2016.12.019

Chen YG, Satpathy AT, Chang HY (2017) Gene regulation in the immune system by long noncoding RNAs. Nat Immunol 18:962–972. https://doi.org/10.1038/ni.3771

Coker R, Wood PB (1986) Changing patterns of Kaposi's sarcoma in N.E. Zaire. Trans R Soc Trop Med Hyg 80:965–966. https://doi.org/10.1016/0035-9203(86)90274-9

Colaprico A, Silva TC, Olsen C et al (2016) TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res 44:e71. https://doi.org/10.1093/nar/gkv1507

Collins FS, Morgan M, Patrinos A (2003) The human genome project: lessons from large-scale biology. Science 300:286–290. https://doi.org/10.1126/science.1084564

Denaro N, Merlano MC, Lo Nigro C (2019) Long noncoding RNAs as regulators of cancer immunity. Mol Oncol 13:61–73. https://doi.org/10.1002/1878-0261.12413

Edgar R, Domrachev M, Lash AE (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30:207–210. https://doi.org/10.1093/nar/30.1.207

Egranov SD, Hu Q, Lin C, Yang L (2020) LncRNAs as tumor cell intrinsic factors that affect cancer immunotherapy. RNA Biol 17:1625–1627. https://doi.org/10.1080/15476286.2020.1767455

Elling R, Chan J, Fitzgerald KA (2016) Emerging role of long noncoding RNAs as regulators of innate immune cell development and inflammatory gene expression. Eur J Immunol 46:504–512. https://doi.org/10.1002/eji.201444558

Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) affy–analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20:307–315. https://doi.org/10.1093/bioinformatics/btg405

Hu C, Li L, Huang W et al (2022) Interpretable machine learning for early prediction of prognosis in sepsis: a discovery and validation study. Infect Dis Ther 11:1117–1132. https://doi.org/10.1007/s40121-022-00628-6

Hung J-H, Weng Z (2017) Analysis of microarray and RNA-seq expression profiling data. Cold Spring Harb Protoc. https://doi.org/10.1101/pdb.top093104

Iranzo P, Callejo A, Assaf JD et al (2022) Overview of Checkpoint inhibitors mechanism of action: role of immune-related adverse events and their treatment on progression of underlying cancer. Front Med (lausanne) 9:875974. https://doi.org/10.3389/fmed.2022.875974

Jiang Y-Q, Cao S-E, Cao S et al (2021) Preoperative identification of microvascular invasion in hepatocellular carcinoma by XGBoost and deep learning. J Cancer Res Clin Oncol 147:821–833. https://doi.org/10.1007/s00432-020-03366-9

Jochems C, Schlom J (2011) Tumor-infiltrating immune cells and prognosis: the potential link between conventional cancer therapy and immunity. Exp Biol Med (maywood) 236:567–579. https://doi.org/10.1258/ebm.2011.011007

Kong A, Dong H, Zhang G et al (2022) Correlation between genes of the ceRNA network and tumor-infiltrating immune cells and their biomarker screening in kidney renal clear cell carcinoma. J Oncol 2022:4084461. https://doi.org/10.1155/2022/4084461

Labianca R, Beretta GD, Kildani B et al (2010) Colon cancer. Crit Rev Oncol Hematol 74:106–133. https://doi.org/10.1016/j.critrevonc.2010.01.010

Li J, Tibshirani R (2013) Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res 22:519–536. https://doi.org/10.1177/0962280211428386

Li Y-X, Wang S-M, Li C-Q (2021) Four-lncRNA immune prognostic signature for triple-negative breast cancer running title: immune lncRNAs predict prognosis of TNBC. Math Biosci Eng 18:3939–3956. https://doi.org/10.3934/mbe.2021197

Liang Z, Xie H, Shen W et al (2022) The synergism of natural compounds and conventional therapeutics against colorectal cancer progression and metastasis. Front Biosci 27:263. https://doi.org/10.31083/j.fbl2709263

Liu Z, Liu L, Weng S et al (2022a) Machine learning-based integration develops an immune-derived lncRNA signature for improving outcomes in colorectal cancer. Nat Commun 13:816. https://doi.org/10.1038/s41467-022-28421-6

Liu Z, Liu X, Liu F et al (2022b) The comprehensive and systematic identification of BLCA-specific SF-regulated, survival-related AS events. Gene 835:146657. https://doi.org/10.1016/j.gene.2022.146657

Liu Z, Georgakopoulos-Soares I, Ahituv N, Wong K-C (2023) Risk scoring based on DNA methylation-driven related DEGs for colorectal cancer prognosis with systematic insights. Life Sci 316:121413. https://doi.org/10.1016/j.lfs.2023.121413

Mahoney KM, Rennert PD, Freeman GJ (2015) Combination cancer immunotherapy and new immunomodulatory targets. Nat Rev Drug Discov 14:561–584. https://doi.org/10.1038/nrd4591

Mola S, Pandolfo C, Sica A, Porta C (2020) The macrophages-microbiota interplay in colorectal cancer (CRC)-related inflammation: prognostic and therapeutic significance. Int J Mol Sci. https://doi.org/10.3390/ijms21186866

Najafi S, Khatami SH, Khorsand M et al (2022) Long non-coding RNAs (lncRNAs); roles in tumorigenesis and potentials as biomarkers in cancer diagnosis. Exp Cell Res 418:113294. https://doi.org/10.1016/j.yexcr.2022.113294

Nurk S, Koren S, Rhie A et al (2022) The complete sequence of a human genome. Science 376:44–53. https://doi.org/10.1126/science.abj6987

Ooki A, Shinozaki E, Yamaguchi K (2021) Immunotherapy in colorectal cancer: current and future strategies. J Anus Rectum Colon 5:11–24. https://doi.org/10.23922/jarc.2020-064

Pareek CS, Smoczynski R, Tretyn A (2011) Sequencing technologies and genome sequencing. J Appl Genet 52:413–435. https://doi.org/10.1007/s13353-011-0057-x

Park E-G, Pyo S-J, Cui Y et al (2022) Tumor immune microenvironment lncRNAs. Brief Bioinform. https://doi.org/10.1093/bib/bbab504

Peng W-X, Koirala P, Mo Y-Y (2017) LncRNA-mediated regulation of cell signaling in cancer. Oncogene 36:5661–5667. https://doi.org/10.1038/onc.2017.184

Qi F, Du X, Zhao Z et al (2021) Tumor mutation burden-associated LINC00638/miR-4732-3p/ULBP1 axis promotes immune escape via PD-L1 in hepatocellular carcinoma. Front Oncol 11:729340. https://doi.org/10.3389/fonc.2021.729340

Robert C (2020) A decade of immune-checkpoint inhibitors in cancer therapy. Nat Commun 11:3801. https://doi.org/10.1038/s41467-020-17670-y

Sananmuang T, Puthier D, Nguyen C, Chokeshaiusaha K (2020) Novel classifier orthologs of bovine and human oocytes matured in different melatonin environments. Theriogenology 156:82–89. https://doi.org/10.1016/j.theriogenology.2020.06.029

Shah M, Hubbard RA, Mamtani R et al (2022) Very high PD-L1 expression as a prognostic indicator of overall survival among patients with advanced non-small cell lung cancer receiving anti-PD-(L)1 monotherapies in routine practice. Pharmacoepidemiol Drug Saf 31:1121–1126. https://doi.org/10.1002/pds.5487

Shi H, Li K, Feng J et al (2020) LncRNA-DANCR interferes with miR-125b-5p/HK2 axis to desensitize colon cancer cells to cisplatin vis activating anaerobic glycolysis. Front Oncol 10:1034. https://doi.org/10.3389/fonc.2020.01034

Slatko BE, Gardner AF, Ausubel FM (2018) Overview of next-generation sequencing technologies. Curr Protoc Mol Biol 122:e59. https://doi.org/10.1002/cpmb.59

Smith MC (1992) Is all knowing personal knowing? Nurs Sci Q 5:2–3. https://doi.org/10.1177/089431849200500102

Sordillo LM, Aitken SL (2009) Impact of oxidative stress on the health and immune function of dairy cattle. Vet Immunol Immunopathol 128:104–109. https://doi.org/10.1016/j.vetimm.2008.10.305

Sorensen SF, Zhou W, Dolled-Filhart M et al (2016) PD-L1 expression and survival among patients with advanced non-small cell lung cancer treated with chemotherapy. Transl Oncol 9:64–69. https://doi.org/10.1016/j.tranon.2016.01.003

Sun Y, Li Z, Wang W et al (2022) Identification and verification of YBX3 and its regulatory gene HEIH as an oncogenic system: a multidimensional analysis in colon cancer. Front Immunol 13:957865. https://doi.org/10.3389/fimmu.2022.957865

Sun J, Zhang Z, Bao S et al (2020a) Identification of tumor immune infiltration-associated lncRNAs for improving prognosis and immunotherapy response of patients with non-small cell lung cancer. J Immunother Cancer. https://doi.org/10.1136/jitc-2019-000110

Sun Y, Cao B, Zhou J (2020b) Roles of DANCR/microRNA-518a-3p/MDMA ceRNA network in the growth and malignant behaviors of colon cancer cells. BMC Cancer 20:434. https://doi.org/10.1186/s12885-020-06856-8

Sung H, Ferlay J, Siegel RL et al (2021) Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 71:209–249. https://doi.org/10.3322/caac.21660

Tao B, Shi J, Shuai S et al (2021) CYB561D2 up-regulation activates STAT3 to induce immunosuppression and aggression in gliomas. J Transl Med 19:338. https://doi.org/10.1186/s12967-021-02987-z

Terzić J, Grivennikov S, Karin E, Karin M (2010) Inflammation and colon cancer. Gastroenterology 138:2101–2114. https://doi.org/10.1053/j.gastro.2010.01.058

Thorsson V, Gibbs DL, Brown SD et al (2019) The immune landscape of cancer. Immunity 51:411–412. https://doi.org/10.1016/j.immuni.2019.08.004

Tibshirani R (2006) A simple method for assessing sample sizes in microarray experiments. BMC Bioinf 7:106. https://doi.org/10.1186/1471-2105-7-106

Toth R, Schiffmann H, Hube-Magg C et al (2019) Random forest-based modelling to detect biomarkers for prostate cancer progression. Clin Epigenetics 11:148. https://doi.org/10.1186/s13148-019-0736-8

Turner M, Galloway A, Vigorito E (2014) Noncoding RNA and its associated proteins as regulatory elements of the immune system. Nat Immunol 15:484–491. https://doi.org/10.1038/ni.2887

Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 98:5116–5121. https://doi.org/10.1073/pnas.091062498

Tzeng I-S (2021) Modified significance analysis of microarrays in heterogeneous diseases. J Pers Med. https://doi.org/10.3390/jpm11020062

Vitorino M, Alpuim Costa D, Vicente R et al (2022) Local breast microbiota: a "new" player on the block. Cancers (basel). https://doi.org/10.3390/cancers14153811

Wallentin L, Eriksson N, Olszowka M et al (2021) Plasma proteins associated with cardiovascular death in patients with chronic coronary heart disease: a retrospective study. PLoS Med 18:e1003513. https://doi.org/10.1371/journal.pmed.1003513

Wang Y, Zhang H, Liu C et al (2022) Immune checkpoint modulators in cancer immunotherapy: recent advances and emerging concepts. J Hematol Oncol 15:111. https://doi.org/10.1186/s13045-022-01325-0

Wang B, Chen L, Li Q et al (2023) LINC00638/hsa-miR-29b-3p axis-mediated high expression of CDCA4 correlates with tumor immune infiltration and hepatocellular carcinoma progression. Transl Cancer Res 12:1290–1306. https://doi.org/10.21037/tcr-23-569

Wei S, Gu X, Zhang W (2022) Development and validation of a novel fibroblast scoring model for lung adenocarcinoma. Front Oncol 12:905212. https://doi.org/10.3389/fonc.2022.905212

Wu M, Shang X, Sun Y et al (2020a) Integrated analysis of lymphocyte infiltration-associated lncRNA for ovarian cancer via TCGA. Gtex and GEO Datasets Peerj 8:e8961. https://doi.org/10.7717/peerj.8961

Wu Y, Zhang L, He S et al (2020b) Identification of immune-related LncRNA for predicting prognosis and immunotherapeutic response in bladder cancer. Aging 12:23306–23325. https://doi.org/10.18632/aging.104115

Yanai I, Benjamin H, Shmoish M et al (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 21:650–659. https://doi.org/10.1093/bioinformatics/bti042

Yang X-J, Zhao J-J, Chen W-J et al (2018) Silencing long non-coding RNA, differentiation antagonizing non-protein coding RNA promotes apoptosis and inhibits tumor growth in colon cancer. Oncol Lett 16:2865–2872. https://doi.org/10.3892/ol.2018.9034

Yu X, Fang C, Zhang K, Su C (2022) Recent advances in nanoparticles-based platforms targeting the PD-1/PD-L1 pathway for cancer treatment. Pharmaceutics. https://doi.org/10.3390/pharmaceutics14081581

Zhang H, Zhang N, Wu W et al (2022a) Machine learning-based tumor-infiltrating immune cell-associated lncRNAs for predicting prognosis and immunotherapy response in patients with glioblastoma. Brief Bioinform. https://doi.org/10.1093/bib/bbac386

Zhang L, Huang T, Xu F et al (2022b) Prediction of prognosis in elderly patients with sepsis based on machine learning (random survival forest). BMC Emerg Med 22:26. https://doi.org/10.1186/s12873-022-00582-z

Zheng Z, Chen Z, Xie Y et al (2021) Development and validation of a CT-based nomogram for preoperative prediction of clear cell renal cell carcinoma grades. Eur Radiol 31:6078–6086. https://doi.org/10.1007/s00330-020-07667-y

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