A 15-lncRNA signature predicts survival and functions as a ceRNA in patients with colorectal cancer

Xunino Wang1
Jiango Zhou2
MaoLin Xu1
Yongfei Yan3
Liang Huang1
Yanshen Kuang1
YuanSheng Liu4
Peng Li3
Wei Zheng3
Hongyi Liu3
Baoqing Jia3

1Department of General Surgery, Chinese PLA Medical School, Beijing 100853, People’s Republic of China; 2Division of Thoracic Oncology, Department of Oncology, Affiliated Hospital of Zunyi Medical University, Zunyi City 563000, Guizhou Province, People’s Republic of China; 3Department of General Surgery, Chinese PLA General Hospital, Beijing 100853, People’s Republic of China; 4School of Medicine, Nankai University, Tianjing, People’s Republic of China

Purpose: Colorectal cancer (CRC) is one of the most common malignant tumors worldwide. This study aimed to explore the prognostic value of lncRNAs in CRC.

Material and methods: We performed gene expression profiling to identify differentially expressed lncRNAs between 51 normal and 646 tumor tissues from The Cancer Genome Atlas database. Cox regression and robust likelihood-based survival models were used to find prognosis-related lncRNAs. An lncRNA signature was developed to predict the overall survival of patients with CRC. In addition, a receiver operating characteristic curve analysis was performed to identify the optimal cutoff with the best Youden index to divide patients into different groups based on risk level.

Results: Eighty survival-related lncRNAs were identified and a 15-lncRNA signature was developed on the basis of a risk score to comprehensively predict the overall survival of patients with CRC. The prognostic value of the 15-lncRNA risk score was validated using the internal testing set and total set. The risk indicator was shown to be an independent prognostic factor (hazard ratio =2.92; 95% CI: 1.73–4.94; P<0.001). Notably, all 15 lncRNAs (AC024581.1, FOXD3-AS1, AC012531.1, AC003101.2, LINC01219, AC083967.1, AL590483.1, AC105118.1, AC010789.1, AC067930.5, AC105219.2, LINC01354, LINC02474, LINC02257, and AC079612.1) were newly found to correlate with the prognosis of patients with CRC. Furthermore, the function of 15 lncRNAs was explored through the ceRNA network. These lncRNAs regulated coding genes that were involved in many key cancer pathways.

Conclusion: A 15-lncRNA expression signature was discovered as a prognostic indicator for patients with CRC, which may act as competing endogenous RNA (ceRNAs) to play a crucial role in the modulation of cancer-related pathways. These findings may allow a better understanding of the prognostic value of lncRNAs.

Keywords: long noncoding RNA, colorectal cancer, survival, biomarker, competing endogenous RNA, ceRNA

Introduction

Colorectal cancer (CRC) is one of the most common malignant tumors of the gastrointestinal tract worldwide, as well as the fourth leading cause of cancer-related death owing to its prevalence and mortality.1 Studies have shown that CRC is caused by several genetic factors, including changes in chromosomal copy number, aberrant gene methylation, and dysregulated gene expression.1,2 Considerable progress has been made in the diagnosis and treatment of CRC in the last several decades. However, the current prognostic factors for patients with CRC do not meet clinical needs, making it necessary to identify novel biomarkers in a sensitive and accurate way to better predict overall survival.
lncRNAs, usually >200 nucleotides in length, are a class of RNAs that do not code for proteins.\(^4\) lncRNAs used to be considered “transcript junk,” but have recently emerged as key molecules in multiple complex biological processes (BP),\(^1\,^5\) including proliferation, cell cycle progression, and survival.\(^6\) Several reports have shown that lncRNAs serve as modulators of carcinogenesis and affect the rates of invasion and metastasis in several types of cancer.\(^6\) However, the biological function and prognostic value of many lncRNAs remain unknown. Interestingly, it has been shown that numerous lncRNAs can act as competing endogenous RNAs (ceRNAs) to regulate the expression of coding genes\(^7\) that have common miRNA response elements (MREs). In this study, the predictive value of lncRNAs in patients with CRC was explored. Furthermore, the function of these lncRNAs was investigated using the ceRNA network.

**Materials and methods**

**Data processing and computational analysis**

Figure 1 shows the overall workflow of this study. The data of 697 RNA expression profiles (level 3), including 51 normal tissues and 646 tumor tissues, were downloaded from The Cancer Genome Atlas (TCGA) data portal (dated to September 18, 2017). This study met the publication guidelines provided by TCGA (http://cancergenome.nih.gov/publications/publicationsguidelines). According to TCGA guidelines, RNA expression profiles can be studied in three forms: HT-seq raw read count, Fragments per Kilobase of transcript per Million mapped reads (FPKM), and FPKM-UQ (upper quartile normalization). Here, HT-seq raw read count was chosen. IncRNAs general feature format file (Gencode.v27) was used as the lncRNA annotation reference.\(^8\) The expression profiles of lncRNAs were analyzed by edgeR.\(^9\,^10\) Differentially expressed lncRNAs were selected according to \(P\)-value (\(\leq 0.01\)) and absolute fold change (\(\geq 2\)).

**Identification of lncRNAs related to patient prognosis**

Samples were filtered by removing cases without complete survival data to yield 616 samples that were included in our analysis. All samples were randomly divided into either training set (308 samples) or validation set (308 samples) groups. The clinical and demographic characteristics of the study population are shown in Table 1. There was no statistical difference between the two sets. To determine the feasibility and reliability of survival-associated lncRNAs as prognostic markers in CRC, univariate Cox proportional hazards regression was applied to identify overall survival-related lncRNAs. The robust likelihood-based survival model, using the R package analysis method (Rbsurv), was then applied to further identify prognosis-related lncRNAs.\(^11\) The protocol of this method was as follows: first, the model randomly put \(N(1-p)\) samples into the training set and \(Np\) cases into the validation set. Here, we chose \(p=1/3\). Second, the model added a gene to the training set and obtained the parameter for the gene. The loglik was evaluated for each parameter and validated within the internal validation samples. The procedure was repeated 1,000 times to select the best prognosis-related lncRNAs with the smallest mean negative loglik. Next, the Akaike information criterion (AIC) was computed and used as an estimator of the relative quality of statistical models for a given set of data, and the optimal model was chosen with the smallest AIC. \(P<0.05\) was considered statistically significant.

**Establishment and validation of the risk formula**

lncRNAs chosen from the previous step were inserted into the multiply Cox proportional model to calculate the coefficients in the training set, thereby establishing the risk
formula. Risk scores for each sample were calculated using this formula. All patients were classified into either the high-risk or the low-risk group on the basis of the median of their risk score. The Kaplan–Meier method and the log-rank test were applied to analyze the overall survival of the two groups using the R package survival analysis.12,13 A time-dependent receiver operating characteristic curve (ROC) was constructed to evaluate the prediction value of the model (version 1.0.3),14 and the figures were plotted by ggplot2 (version 2.2.1)15 and ggfortify (version 0.4.1).16,17 All data were processed and analyzed by perl 5 version 24, excel 2010, and R (version 3.4.1).

Determination of lncRNA function

The function of the lncRNAs was explored using the triple ceRNA (lncRNA–miRNA–mRNA) network. The sequences of the identified lncRNAs were obtained from Ensembl18 and inputted into the miRDB19,20 database to predict their miRNA targets. The corresponding coding genes were then identified using miRDB,19,20 miRTarBase,21 and TargetScan.22 The triple ceRNA network was visualized and constructed by Cytoscape v3.5.1.23 The Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis coding genes were annotated by the R package of clusterProfiler.24 The cutoff P-value was 0.05.

Results

Differential expression of lncRNAs

A total of 1,103 differentially expressed lncRNAs were identified in patients with CRC. These lncRNAs are listed in Table S1. Eighty lncRNAs that were associated with overall survival were identified through our univariate Cox regression analysis in the total set (Table S2).

Identification of a 15-lncRNA signature

The 20 lncRNAs with the lowest P-value were selected (Table 2) and analyzed with the robust likelihood-based survival model. Fifteen lncRNAs were selected with the lowest AIC values. The risk coefficients for these lncRNAs were calculated using the multivariable Cox proportional hazards model. The risk formula used to calculate the risk score was as follows: $(0.238 \times AC024581.1) + (0.053 \times FOXD3-AS1) + (0.067 \times AC012531.1) + (0.221 \times AC003101.2) + (0.357 \times LINC01219) + (0.082 \times AC083967.1) + (0.113 \times AL590483.1) + (0.060 \times AC01051-18.1) + (0.031 \times AC010789.1) + (0.126 \times AC067930.5) + (0.161 \times AC105219.2) + (0.317 \times LINC01354) + (0.139 \times LINC02474) + (0.034 \times LINC02257) + (0.269 \times AC079612.1).

| Table 1 Clinical covariates for TCGA colorectal cancer |
|---|---|---|---|---|
| Covariate | Total set (n=616) | Training set (n=308) | Validation set (n=308) | P-value* |
| Age (years), n | | | | |
| ≥65 | 368 | 175 | 193 | 0.162 |
| <65 | 248 | 133 | 115 | |
| Gender | | | | |
| Male | 329 | 164 | 165 | 1.000 |
| Female | 287 | 144 | 143 | |
| Pathological stage, n | | | | |
| I + II | 330 | 166 | 164 | 0.805 |
| III + IV | 267 | 131 | 136 | |
| Not report | 19 | 11 | 8 | |

Note: χ² test.

| Table 2 Top 20 survival-related lncRNAs |
|---|---|---|
| lncRNAs | HR | P-value |
| AC093895.1 | 1.207063 | 0.000129 |
| AC012531.1 | 1.201873 | 0.000225 |
| AC020891.2 | 1.27243 | 0.000854 |
| AC002076.1 | 1.34785 | 0.001103 |
| AC016027.1 | 0.682998 | 0.001165 |
| AC105118.1 | 1.329465 | 0.001222 |
| LINC02474 | 1.10204 | 0.001489 |
| AC079612.1 | 0.743541 | 0.00256 |
| AC083967.1 | 1.213143 | 0.002645 |
| AC067930.5 | 1.20642 | 0.002924 |
| AC100789.1 | 1.158292 | 0.003233 |
| AL590483.2 | 0.832073 | 0.004144 |
| LINC01219 | 1.258921 | 0.004814 |
| AL590483.1 | 0.832461 | 0.004836 |
| AC030101.2 | 1.265178 | 0.005871 |
| FOXD3-AS1 | 1.188585 | 0.006167 |
| AC105219.2 | 1.182066 | 0.006402 |
| LINC02257 | 1.159131 | 0.006745 |
| AC024581.1 | 1.289277 | 0.00682 |
| LINC01354 | 1.199976 | 0.009076 |
Additionally, the risk scores were calculated for each patient in the training set. The patients were divided into two groups on the basis of the median of the risk scores (Figure 2A). Figure 2B shows the distribution of patient survival status and survival time. Survival, assessed with the Kaplan–Meier method and log-rank test, indicated that patients with a high-risk score had a shorter survival time ($P<0.001$) (Figure 2C). In our analysis, survival time was negatively correlated with risk score.

**Validation of the prognostic value of the IncRNAs**

To assess prognostic value, ROC was conducted for the 15- IncRNA signature (Figure 3A). For our analysis, the area under curve was 0.708. 2.027 was chosen as the best optimal cutoff, taking into account the maximal sensitivity and specificity of our survival prediction. Patients from the data sets (total set and validating set) were further divided into high-risk or low-risk groups. Figure 3B and 3C shows the Kaplan–Meier survival curves for the testing set and the total set, respectively, where the results were all consistent with our model.

**Determination of IncRNA function**

The 15 IncRNAs identified in our study were inputted into the miRDB database to predict their miRNA targets (yielding a total of 222 miRNAs), and the coding genes for these miRNAs were then predicted (yielding 1,179 genes). Figure 4A shows an overview of the triple ceRNA (IncRNA–miRNA–mRNA) network. The detailed interactions of the ceRNA network are shown in Table S3. The functional enrichment assay identified 691 GO terms in BP, 46 GO terms in cellular components, 81 GO terms in molecular function (Table S4), and 46 pathways (Table S5). It also showed that these genes are involved in multiple BP, such as regulation of cell morphogenesis, and Wnt-mediated cell signaling. The top ten GO results are shown in Figure 4B. The top 20 KEGG pathways are shown in Figure 4C. KEGG was enriched in several cancer-related pathways, including the p53 and Wnt signaling pathways. IncRNA AC012531.1 was not only related to the mTOR signal pathway by regulating hsa-mir-424-5p, and hsa-mir-138-5p, played an important role in the Wnt signal pathway. IncRNA AC079612.1 interacted with hsa-mir-760 targeting PPIP5K1 to involve in the phosphatidylinositol signal. Furthermore, these four IncRNAs were also involved in other pathways. However, the rest of the IncRNAs in this study have not been found involved in pathways through interaction with miRNAs.

**Discussion**

Recently, much attention has been given to the clinical significance of IncRNAs, which account for the majority of transcriptional products in the cell. Many IncRNAs have tissue-specific expression patterns and play crucial roles in the progression of diseases, such as gastric cancer and breast cancer.

Those IncRNAs expressed in CRC were comprehensively analyzed, and 1,103 differentially expressed IncRNAs were identified. Then, 80 IncRNAs that were correlated with the overall survival of patients with CRC were selected using the univariate Cox regression model. The robust likelihood-based survival model was then applied, and the 20 IncRNAs with the lowest $P$-value selected to identify a 15-IncRNA signature that predicts the 5-year overall survival of patients with CRC. This model showed excellent performance and consistency throughout the training set, testing set, and total set. These results imply that the 15-IncRNA signature identified in our study may be used as a biomarker to predict patient prognosis in clinical practice.

A literature search in PubMed and Google Scholar indicates this is the first time these 15 IncRNAs are reported to be correlated with CRC.

Previous studies have shown that there is signaling “crosstalk” between different transcriptional products. Many cancer-related phenotypes are driven by IncRNAs, either directly or indirectly, by modulating the stability of various molecules, including DNA, proteins, and miRNAs. The hypothesis of ceRNA is that transcriptional products that share common MREs with target genes communicate with different genes through miRNAs. Furthermore, any transcriptional product that has MREs can act as a ceRNA. These transcriptional products, which share common MREs, including IncRNAs, circular RNAs, and pseudogenes, regulate corresponding genes through miRNAs that function in RNA posttranscriptional silencing by binding the 3’-untranslated region to influence transcript stability. Thus, IncRNAs may act as ceRNAs to indirectly regulate coding genes through miRNAs. It is therefore necessary to explore the role of IncRNAs as ceRNAs.
A novel biomarker of 15 IncRNAs in colorectal cancer

Figure 2 Risk score of IncRNAs in the training set.
Notes: (A) The risk score of patients in the training set based on risk formula. (B) The distribution of patient survival status and survival time. (C) Survival curve of the low-risk and high-risk groups based on median risk score using the Kaplan–Meier method.
In this study, a triple ceRNA (lncRNA–miRNA–mRNA) network was constructed. Bioinformatics analyses of this ceRNA network revealed that 15 lncRNAs may function as ceRNAs to regulate genes that participate in cancer-associated signaling, including p53 and Wnt signaling. Furthermore, this ceRNA network may be involved in other types of cancer because KEGG analysis results of ceRNA showed this network was associated with many cancer-related pathways. For example, the TP53 signaling pathway participates in multiple tumor genesis.

Taken together, these results suggest that 15 differentially expressed lncRNAs play an important role in oncogenesis and may be used as a prognostic biomarker in clinical practice. However, there were still some limits to our study. Our results are based on a bioinformatics analysis and were validated using in vitro or in vivo experimentation. In addition, as the binding affinity between miRNAs and their RNA targets is influenced by the matching between MRE and the seeds regions (as well as other factors), we could not adequately assess the exact function of each ceRNA. Future studies will assess the biological functions of these lncRNAs by measuring their effects on cell proliferation and apoptosis and will further evaluate these lncRNAs as prognostic biomarkers.
Conclusion
In summary, we identified 1,103 lncRNAs that were differentially expressed in CRC.

A 15-lncRNAs’ risk formula was developed that correlated with the overall survival of patients with CRC using a robust likelihood-based survival model, and the function of these newly identified survival-associated lncRNAs was explored. Our results justify further study of the transcriptional regulatory network of lncRNAs in CRC and provide a new resource to discover novel prognostic biomarkers.

Acknowledgments
We thank everyone who supported this study. We also thank Dr Yuan Tang for giving us important advice pertaining to this article.

Disclosure
The authors report no conflicts of interest in this work.

References