An eight-long non-coding RNA signature as a candidate prognostic biomarker for lung cancer

ZHENBO TU1, DU HE2*, XINZHOU DENG1, MENG XIONG1, XIAOXING HUANG1, XINRAN LI1, LING HAO1, QIANSAN DING3 and QIUPING ZHANG1

1Department of Immunology, Wuhan University School of Basic Medical Sciences, Wuhan, Hubei; 2Department of Oncology, The Central Hospital of Enshi Autonomous Prefecture, Enshi Clinical College of Wuhan University, Enshi, Hubei; 3Department of Gastroenterology, Renmin Hospital of Wuhan University, Wuhan, Hubei, P.R. China

Received January 13, 2016; Accepted February 22, 2016

DOI: 10.3892/or.2016.4817

Correspondence to: Professor Qiuping Zhang, Department of Immunology, School of Basic Medical Sciences, Wuhan University, 185 Donghu Road, Wuchang, Wuhan, Hubei 430071, P.R. China
E-mail: qpzhang@whu.edu.cn

Dr Qianshan Ding, Department of Gastroenterology, Renmin Hospital of Wuhan University, Wuhan, Hubei, P.R. China
E-mail: iamdqs@163.com

*Contributed equally

Abbreviations: lncRNA, long non-coding RNA; GEO, Gene Expression Omnibus; GSEA, Gene Set Enrichment Analysis; KEGG, Kyoto Encyclopedia of Genes and Genomes

Key words: long non-coding RNA, lung cancer, survival, gene signature

Abstract. Cumulative evidence suggests that long non-coding RNAs (lncRNAs) may be good biomarkers in various types of tumors. In the present study, we mined lncRNA expression profiling in 739 lung cancer patients from Gene Expression Omnibus (GEO) datasets. A risk score model was constructed based on the expression data of these eight lncRNAs in the training dataset (GSE30219). The validation for the association was performed in three independent testing sets (GSE31210, GSE37745 and GSE19188). Finally, a set of eight lncRNA genes (AK021595, BC030759, AK000053, AK124307, BC020384, AK022024, CR615992 and AF085995) were identified by the random survival forest algorithm. Using a risk score based on the expression signature of these lncRNAs, we separated the patients into low-risk and high-risk groups with significantly different survival times in the training set. This finding was validated in the other three testing sets. Further study revealed that the eight-lncRNA expression signature was independent of age and gender. Gene Set Enrichment Analysis (GSEA) suggested that lncRNAs were involved in cell cycle and DNA replication signaling pathways. Therefore, the eight lncRNAs may be candidate prognostic biomarkers for lung cancer patients.

Introduction

Lung cancer is one of the most commonly diagnosed cancers and remains the leading cause of cancer-related death worldwide (1,2). It has been traditionally subdivided into two principal groups, namely, small cell lung cancer and non-small cell lung cancer. The latter type is more common than the former. Despite diverse treatment methods including surgery, radiation therapy and chemotherapy, the overall 5-year survival rate remains ~18.2% (3). The high mortality rate of lung cancer is partly due to the lack of effective prognostic biomarkers. Therefore, the identification of novel prognostic factors as biomarkers that may be used in the early detection of lung cancer is critical.

Long non-coding RNAs (lncRNAs) are mRNA-like transcripts with more than 200 nucleotides that lack significant protein-coding abilities (4,5). Increasing evidence suggests that lncRNAs are a new class of players involved in the development and progression of cancer (6). More and more research suggests that these transcripts are frequently aberrantly expressed in cancers, and some have been implicated in the diagnosis and prognostication (7) in neuroblastoma (8), prostate (9), breast (10-12), ovarian (13,14), gastric (15) and colorectal cancer (16,17), and multiple myeloma (18). Due to the specific expression of lncRNAs in cancer, lncRNAs could become biomarkers by which to diagnosis cancer or predict patient survival. Thus, identification of various lncRNAs which are specifically expressed in lung cancer may have predictive and prognostic value for lung cancer patients.

Currently, massive lncRNA-specific probes are presented on microarray platforms (Affymetrix U133 Plus 2.0); thus, we are able to use previously published gene expression microarray data from the Gene Expression Omnibus (GEO) database to identify various prognostic signature lncRNAs. Furthermore, bioinformatic analysis was used to identify the signaling pathways that involve lncRNAs by Gene Set Enrichment Analysis (GSEA).
Materials and methods

Lung cancer datasets and patient information. Lung cancer datasets were downloaded from the GEO database. A total of 739 patients were utilized in the present study after filtering out samples without clinical survival information. It included 293 patients from GSE30219 (19), 226 patients from GSE31210 (20), 168 patients from GSE37745 (21), and 52 patients from GSE19188 (22). We selected these datasets that included >50 patients with survival status information. We followed the strategy of using the largest dataset (GSE30219) as training set. Three independent datasets (GSE31210, GSE37745 and GSE19188) were included in the present study as testing sets.

Microarray processing and IncRNA profile mining. All the microarray raw data (CEL files) of four lung cancer cohorts were processed using the robust multichip average (RMA) algorithm for background adjustment (23). GATExplorer was used to process microarrays on a local computer for gene expression of IncRNAs (24). IncRNA mapper was obtained from GATExplorer, which included the probes that do not map to any coding region but that were mapped to a database for non-coding RNAs of human and mouse (derived from RNAdb (25). The coding potential analysis of the IncRNAs was carried out by CNCI to classify protein-coding or non-coding transcripts (26). Each IncRNA included at least a minimum of three probes mapping in the corresponding IncRNA entity. We created a risk-score formula according to the expression of these eight IncRNAs for survival prediction. Patients having higher risk scores were expected to have poorer survival outcomes.

Statistical analysis. The association between the IncRNA gene expression and patient survival was assessed by univariable Cox proportional hazards regression analysis along with a permutation test using BRB-ArrayTools (Biometric Research Branch) package (27) in the training set. We identified expression of several IncRNAs that were strongly correlated with survival. Considering that a smaller number of genes in the model would make the model more practical, we performed the random survival forests variable hunting (RSFVH) algorithm (28). Using a smaller number of genes selected fitted in a multivariable Cox regression model; we constructed a formula to predict survival in the training set. Each patient was assigned a risk score that is a linear combination of the expression levels of the significant IncRNAs weighted by their respective Cox regression coefficients (29). According to this risk score, patients in the training set were divided into low-risk and high-risk groups using the median risk score as the cut-off. The Kaplan-Meier method was used to estimate survival time, and other three independent testing groups were performed for validation. Differences in survival times between the low-risk and high-risk groups in each set were compared by the two-sided log-rank test, respectively.

Bioinformatic analysis of IncRNA gene function. GSEA was performed by the JAVA program (http://www.broadinstitute.org/gsea) using MSigDB C2 CP: canonical pathway gene set collection (1,320 gene sets available). Gene sets with a false discovery rate (FDR) value <0.05 after performing 1,000 permutations were considered to be significantly enriched (30). Cytoscape (version 2.8.2) and the Enrichment Map software were used to visualize the GSEA results (31). Encyclopedia of Genes and Genomes (KEGG) enrichment analyses of the co-expressed protein-coding genes with prognostic IncRNAs were performed to predict the biological function of prognostic IncRNAs using the DAVID Bioinformatics Tool (version 6.7) (32). Enrichment analysis was carried out using the functional annotation clustering options, and was limited to KEGG pathways in the ‘Biological Process’ categories.

Results

Identification of prognostic IncRNA genes from the training set. As summarized in the workflow (Fig. 1), all analyses were performed in the training set (GSE30219) and validated in the testing set (GSE31210, GSE37745 and GSE19188). The training set (n=293) was analyzed for the detection of prognostic IncRNA genes. By subjecting the IncRNA expression data derived from the training set to univariable Cox proportional hazards regression analysis using the BRB-ArrayTools, we identified a set of 36 IncRNAs that were strongly correlated with patient overall survival (p<0.001 and FDR <0.001) from a total of 5,635 IncRNAs. Based on the random forest survival model (see Materials and methods), the eight IncRNAs (Table I) were selected as predictors (Fig. 2). In Table I a list of these eight IncRNAs is shown with their obtained coefficient and variable importance values. Based on these results, BC030759 was the most relevant with overall survival in the training set (HR=3.513); the positive coefficients of the IncRNAs (AK021595, BC030759, AK000053, BC020384 and AK022024) indicated that their higher levels of expression were associated with shorter survival, and the negative coefficients of the other IncRNAs (AK124307, CR615992 and AF085995) indicated that their higher levels of expression were associated with longer survival. All of the eight IncRNAs have been verified in the ncRNA Expression Database (www.nred.matticklab.com) and these eight transcripts were classified as ncRNAs in this website (33). As coding potential analysis is commonly used to classify whether a transcript is of coding potential or not (34), we used CNCI to test those eight transcripts (26). This tool also suggested that all the eight transcripts were non-coding transcripts with no coding potential.

An eight-IncRNA signature predicts survival of lung cancer patients in the training set. To investigate whether the eight-IncRNA signature could provide an accurate prediction of survival in lung cancer patients, we created a risk-score formula according to the expression of these eight IncRNAs for survival prediction in the training set GSE30219 (n=293), as follows: risk score, 0.212*AK021595+0.416*BC030759 +0.322*AK000053-0.165*AK124307+0.301*BC020384+0.423*AK022024-0.084*CR615992-0.459*AF085995. Then, we calculated the eight-IncRNA signature risk score for each patient in the training set. Patients were divided into a low-risk or high-risk group using the median risk score as cut-off value. Patients in the high-risk group had a shorter survival time than
Figure 1. Diagram of the study. The order of analyses to obtain the large number of lncRNAs and develop the risk score model to predict prognostic outcomes by validation in the test data sets.

Figure 2. Error rate for the data as a function of trees (A) and out-of-bag importance values for predictors (B).
Figure 3. Kaplan-Meier estimates of the survival of GEO patients using the eight-lncRNA signature. The Kaplan-Meier plots were used to visualize the survival probabilities for the low-risk vs. high-risk group of GEO patients determined on the basis of the median risk score from the training set. (A) Kaplan-Meier curves for GSE30219 training-set patients (n=293). (B) Kaplan-Meier curves for GSE31210 testing-set patients (n=226). (C) Kaplan-Meier curves for GSE37745 testing-set patients (n=168). The longest survival time in the model was 250 months, thus the patients whose survival time was >300 months was removed. (D) Kaplan-Meier curves for GSE19188 testing-set patients (n=52). The tick marks on the Kaplan-Meier curves represent the censored subjects. The differences between the two curves were determined by the two-sided log-rank test.

Figure 4. lncRNA risk score analysis of the GSE30219 patients. (A) Patient overall survival status. (B) Heatmap of the lncRNA expression profiles. Rows represent lncRNAs and columns represent patients.
The association of the eight-lncRNA risk score and survival was also significant when it was evaluated as a continuous variable in the univariate Cox regression model (Table II).
Validation of the eight-lncRNA signature for survival prediction in the testing sets. To confirm our findings, we calculated the risk score in the testing sets including GSE31210 (n=226), GSE37745 (n=168) and GSE19188 (n=52). Similar to the training set findings, patients in the high-risk group had a shorter survival time than patients in the low-risk group (Fig. 3B-D). Meanwhile, patient survival throughout the follow-up in the low-risk group was better when compared to survival in the high-risk group. In the univariable Cox regression model, the risk score was similar with the high-risk group which had a
shorter overall survival. The patient survival status (Fig. 4A) and IncRNA values (Fig. 4B) were analyzed independently in the training set. Some of the clinical information (stage and subtype) was not available for a substantial proportion of cases, thus we performed multivariate Cox regression analysis concerning age and gender. The result showed that the eight-IncRNA expression signature was independent of age and gender. Eight-IncRNA risk score, age (available in GSE31210 and GSE37745) and gender (available in GSE31210, GSE37745 and GSE19188) were defined as covariates. These results showed that risk score was an independent predictor of lung cancer patient survival (Table II).

Identification of eight-IncRNA signature-associated biological pathways and processes. GSEA was carried out to identify the associated biological processes and signaling pathways (30). We compared the gene expression profile of lung cancer patients in the low-risk and high-risk groups classified by the eight-IncRNA gene signature in the training set (GSE 30219). The gene sets with significantly different expression (FDR <0.01; p<0.005) were picked up, which implied that the signature may be involved in the cell cycle and DNA replication-related pathways (Fig. 5A and B), and it was visualized as an interaction network with Cytoscape (Fig. 5C). These related pathways were reported to affect cancer cell proliferation (35-37).

Discussion

As a new class of ncRNAs, IncRNAs were demonstrated to be dysregulated in a variety of diseases, particularly in cancers (38). Numerous studies of abnormal IncRNA expression in various types of cancer suggest that they play an important role in tumorigenesis, and IncRNAs may serve as independent biomarkers for diagnosis and prognosis (39,40). In lung cancer, numerous studies have investigated IncRNAs to predict lung cancer patient survival (41-44). Nevertheless, a single factor to predict the prognosis of tumors is not accurate, since high specificity and sensitivity are lacking for most IncRNAs. Currently, research has found that IncRNA expression profiles can be obtained from publicly available, custom-designed DNA microarrays by re-annotating the array probes (12,17,34,45,46).

In the present study, in order to construct a risk score model, we downloaded four datasets (GSE30219, GSE31210, GSE37745 and GSE19188) from GEO databases, and obtained the IncRNA profiling of lung cancer patients. We identified a prognostic, eight-IncRNA signature from the training set. Furthermore, examination of associated molecular pathways revealed that the eight-IncRNA signature was more likely to involve the cell cycle and DNA replication signaling pathways. Cell cycle disorder and DNA replication induce cell proliferation and affect genome instability, further increasing the possibility of canceration of unstable cells, which participates in tumor occurrence and development (35,47-49). Thus, our findings suggest that IncRNA signatures may provide an efficient classification tool for the clinical prognosis of lung cancer.

Zhou et al (46) also identified an eight-IncRNA signature which may be an effective independent prognostic molecular biomarker in the prediction of non-small lung cancer patient survival, and our findings support the characteristics of the eight-IncRNA. The overexpression of IncRNAs (AK021595, BC030759, AK000053, BC020384 and AK022024) were found to be correlated with shorter survival while other IncRNAs (AK124307, CR615992 and AF085995) were downregulated in the high-risk group compared to the low-risk group. Most importantly, the functional study in cancer of these eight IncRNAs has not been reported to date.

The limitations should be acknowledged for the present study. First, in the present study, only 5,635 (out of 15,000+) human IncRNAs were included. The prognostic IncRNAs identified here may not represent all the IncRNA candidates that were potentially correlated with lung cancer overall survival. Secondly, the longest survival time in the model was 250 months, Thus, the patients in GSE37745 whose survival time was >300 months were removed. Thirdly, stage was not included in the present study, since this information was not available for a substantial proportion of cases. Meanwhile, the functions of these eight IncRNAs were inferred by bioinformatics analysis, and these biological roles in tumorigenesis were not clear and should be investigated in experimental studies.

In summary, we identified a signature of a set of eight IncRNAs, which predicted the overall survival in three independent testing sets. Further bioinformatic analysis revealed that the prognostic value was independent of age and gender. Moreover, these IncRNAs are involved in cell cycle and DNA replication signaling pathways. These IncRNAs may have clinical implications as diagnostic markers. However, the biological roles of these eight IncRNAs in tumorigenesis require further study.

Acknowledgements

The present study was supported by the National Natural Science Foundation of China (nos. 81400121, 81270607, 81541027 and 81501352).

References


