Pathology – Research and Practice 212 (2016) 549–554

Contents lists available at ScienceDirect

Pathology – Research and Practice

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

Original article

RNA-Seq analysis of non-small cell lung cancer in female never-smokers reveals candidate cancer-associated long non-coding RNAs

Jun Li, Lintao Bi*, Zhangzhen Shi, Yanxia Sun, Yumei Lin, Hui Shao, Zhenxing Zhu

Department of Hematology and Oncology, China-Japan Union Hospital of Jilin University, Changchun, Jilin 130031, People's Republic of China

ARTICLE INFO

Article history:
Received 13 August 2015
Received in revised form 7 March 2016
Accepted 18 March 2016

Keywords:
Non-small cell lung cancer (NSCLC)
Long non-coding RNAs (lncRNAs)
Co-expression network
Gene ontology (GO)
Pathway enrichment analysis

ABSTRACT

We aimed to elucidate the potential mechanisms of long non-coding RNAs (lncRNAs) in the progression of non-small cell lung cancer (NSCLC). The microarray datasets of GSE37764, including 3 primary NSCLC tumors and 3 matched normal tissues isolated from 6 Korean female never-smokers, were downloaded from Gene Expression Omnibus database. The differentially expressed lncRNAs and mRNA in NSCLC samples were identified using NOISeq package. Co-expression network of differentially expressed lncRNAs and mRNA was established. Gene Ontology (GO) and pathway enrichment analysis were respectively performed. Finally, lncRNAs related to NSCLC were predicted by blasting the differentially expressed lncRNAs with all predicted lncRNAs related to NSCLC. A total of 182 and 539 differentially expressed lncRNAs and mRNA (109 up- and 73 down-regulated lncRNAs; 307 up- and 232 down-regulated mRNA) were respectively identified. Among them, 4 up-regulated lncRNAs, like lnc-geranylgeranyl diphosphate synthase 1 (GGPS1), lnc-zinc finger protein 793 (ZNF793) and lnc-serine/threonine kinase 4 (STK4), and 4 down-regulated lncRNAs including lnc-LOC284440 and lnc-peptidylprolyl isomerase E-like pseudogene (PPIEL), and lnc-zinc finger protein 461 (ZNF461) were predicted related to NSCLC. lncSSPS1, lnc-ZNF793 and lnc-STK4 were co-expressed with linker for activation of T cells (LAT) and Lck interacting transmembrane adaptor 1 (LIME1). Lnc-LOC284440, lnc-PPIEL and lnc-ZNF461 were co-expressed with Src-like-adaptor 2 (SLA2) and defensin beta 4A (DEFB4A). Our study indicates that immune response may be a crucial mechanism involved in NSCLC progression. Lnc-GGPS1, lnc-ZNF793, lnc-STK4, lnc-LOC284440, lnc-PPIEL, and lnc-ZNF461 may be involved in immune response for promoting NSCLC progression via co-expressing with LAT, LIME1, SLA2 and DEFB4A.

1. Introduction

Lung cancer is the leading cause of cancer-related mortality around the world [1], in which non-small cell lung cancer (NSCLC) accounts for 80–85% of all lung cancers [2]. Smoking is the main cause of lung cancer, however, prevalence of NSCLC in female never-smoker patients has been observed, particularly in Asian countries [2,3]. These epidemiological data make non-smoking-associated lung cancer becoming a distinct disease entity, where specific genetic and molecular characteristics of tumors are being recognized [2]. Despite the recent advances in NSCLC therapies, the high mortality of NSCLC patients has not significantly decreased over the years [4]. Therefore, it is urgent to explore more effective and safe treatment strategies and it is of great importance to elucidate the mechanisms involved in NSCLC at molecular levels.

Recently, long non-coding RNAs (lncRNAs) are emerging as drivers of tumor suppressive and oncogenic functions in various prevalent cancers, such as lung cancer [5,6]. LncRNAs are mRNA-like transcripts ranging in length from 200 nt to 100 kb with lack of significant open reading frames, therefore, they do not function as templates for protein synthesis [7,8]. In spite of this, accumulating epidemiological studies have suggested that misregulated lncRNA expression may be a major contributor to tumorigenesis across numerous cancer types [8,9]. For example, the lncRNA metastasis associated lung adenocarcinoma transcript 1 (MALAT1) is thought to enhance cell migration of NSCLS cells *in vitro* by influencing the expression of motility-related genes [10,11]. lncRNA HOTAIR is associated with short disease-free survival in human NSCLC, and forced expression of HOTAIR enhances lung cancer cell growth and migration [12]. Knockdown of H19 expression can impair lung

* Corresponding author.
E-mail address: biilitao@163.com (L. Bi).

http://dx.doi.org/10.1016/j.prp.2016.03.006
0344-0338/

cancer cell growth and clonogenicity in model systems *in vitro* [13]. Therefore, lncRNAs have been considered as key regulators underlying various and are increasingly becoming a new cancer diagnostic and therapeutic gold mine. However, several major lncRNAs related to NSCLC and their roles in the molecular pathogenesis of NSCLC remain unclear.

Substantial advances in next generation sequencing technologies have revolutionized omics and biomedical studies, especially in the field of cancer research [14]. Deep sequencing techniques provide a comprehensive understanding of cancer progression at the molecular level [14]. In previous study, GSE37764 was used to explore the DNA copy number variations in female never-smoker patients with NSCLC for dissecting the molecular nature of NSCLC via integration with array comparative genomic hybridization (array-CGH) study [15]. In the present study, to investigate key lncRNAs related to NSCLC and to elucidate the roles of lncRNAs in NSCLC progression, lncRNA profiling by high throughput sequencing (RNA-seq) was used to screen the differentially expressed lncRNAs in female never-smokers with NSCLC. Then lncRNAs related to NSCLC and co-expressed mRNAs were identified using comprehensive bioinformatics approaches. Our study will yield new insights into the pathogenesis of NSCLC in female never-smokers.

2. Material and methods

2.1. Sources of data

The array data of GSE37764 [15], including 3 primary NSCLC tumors and 3 matched normal tissues isolated from 6 Korean female never-smokers, was downloaded from Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/), which was sequenced on Illumina Genome Analyzer IIx (Homo sapiens) platform. Sequencing strategy was paired-end reads and reads length was 78 nt.

2.2. Raw read filtering

The raw reads were firstly converted into the fastq format using fastq-dump program in sratoolkit [16], then dirty raw reads were removed prior to analyzing the data. Three criteria were utilized to filter out dirty raw reads: Remove reads with sequence adaptors; remove reads with more than 5% 'N' bases; remove low-quality reads, which have more than 10% QA $\leq$ 20 bases. Finally, clean reads were acquired for all subsequent analyses.

2.3. Sequence alignment and transcriptome assembly

The University of California Santa Cruz (UCSC) Genome Browser (http://genome.ucsc.edu) is an online public tool providing access to a growing database of genomic sequence and annotations of various organisms for visualization, comparison and analysis [17]. TopHat and Cufflinks [18] are open-source software tools for gene discovery and comprehensive expression analysis of high-throughput RNA sequencing (RNAseq) data.

Clean reads were aligned to the reference genome downloaded from the UCSC website (version hg19) using bowtie1 in Tophat. The runtime parameters of bowtie1 in the alignment for each read were set as follows:—read-mismatches = 2,—mate-inner-dist = 77, the others run as default parameters.

According to the reference transcript annotation information in UCSC website (version hg19), transcriptome of each read was assembled by Cufflinks. Then the assembled results of each read were merged using cuffmerge in Cufflinks.

2.4. Prediction of lncRNAs

Step1, the assembled transcripts smaller than 200 nt were removed.

Step2, the protein-coding potential of each transcript from step1 was assessed using CPC [19]. Then the transcripts which do not encode proteins but function as lncRNAs were identified.

Step3, lncRNAs were also screened via blasting transcripts from step1 with human lncRNAs extracted from the NONCODEv4 [20] database using blastn in BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi). The parameters of blastn program were sets as follow: Expectation value (E) [Real] = 20; −m = 8.

Step4, common lncRNAs from Step2 and Step3 were considered as predicted lncRNAs.

2.5. Identification of differentially expressed lncRNAs and mRNA

NOISeq [21] package was used to identify differentially expressed lncRNAs and mRNA in primary NSCLC tumors compared to normal controls. q = 0.99 was considered as the cut-off value for screening.

2.6. Co-expression network construction

The absolute value of Pearson correlation coefficient was used as co-expression similarity measure [22]. Cytoscape [23] is an open software for visualizing complex networks and integrating these with any type of attribute data. Therefore, the differentially co-expressed lncRNAs-mRNA pairs with Pearson correlation coefficient > 0.85 were screened, then, the co-expression network of these pairs was established using Cytoscape.

2.7. Gene ontology (GO) and pathway enrichment analysis

GO database is a collection of gene annotation terms for large-scale genomic or transcriptomic data [24]. Database for Annotation Visualization and Integrated Discovery (DAVID) [25] is an online tool used for systematically relating the functional terms with large gene or protein lists. We performed GO-BP (biological process) enrichment analysis of differentially co-expressed mRNA with lncRNAs using DAVID online too. The p-value < 0.05 was defined as the cutoff value.

Web-based Gene Set Analysis Toolkit (WebGestalt) [26] is web-based popular software for the efficient functional enrichment analysis of gene lists derived from large scale genomic, transcriptomic, and proteomic studies. In this paper, pathway enrichment analysis of differentially co-expressed mRNA with lncRNAs was further performed by WebGestalt. The rawR < 0.01 was set as the threshold value.

2.8. Prediction of lncRNAs related to NSCLC

Firstly, the differentially expressed lncRNAs were screened based on hg19 reference genome information in UCSC website again.

Secondly, lncRNAs related to NSCLC were exacted from LncRNADisease. LncRNADisease [27] is a publicly accessible lncRNAs and disease association database, which collect and curate approximately 480 entries of lncRNA-disease associations by experiment validation, including 166 diseases.

Thirdly, lncRNAs were obtained via blasting the above differentially expressed lncRNAs with lncRNAs related to NSCLC using blastn in BLAST. The parameters of blastn program were as follow: Expectation value (E) [Real] = 10; —m = 8. Then lncRNAs were removed based on following criteria: lncRNA was smaller than 200 nt and blast similarity was less than 90.