Clin. Lab. 2017;63:XXX-XXX
©Copyright

ORIGINAL ARTICLE

Comprehensive Analysis of lncRNA and mRNA Expression Profiles in Lung Cancer

Ran Chen [1, 4, *], Wen-Xing Li [2, 3, *], Yan Sun [1, *], Yong Duan [1], Qing Li [1], Ai-Xing Zhang [1, 4], Jia-Lun Hu [4], Yu-Ming Wang [4], Yue-Dong Gao [2, 5]

** These authors contributed equally to this work*
[1] Department of Clinical Laboratory, The First Affiliated Hospital of Kunming Medical University, Kunming 650032, Yunnan, China
[2] State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, Yunnan, China
[3] Institute of Health Sciences, Anhui University, Hefei 230601, Anhui, China
[4] Department of Clinical Laboratory, The Second Affiliated Hospital of Kunming Medical University, Kunming 650032, Yunnan, China
[5] Kunming Biological Diversity Regional Center of Instruments, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, Yunnan, China

SUMMARY

Background: **Lung cancer is one of the most common malignancies across the world. Long noncoding RNAs (lncRNAs) play an important role in the pathology. This study was to compare the lncRNA and mRNA of lung cancer. The aim of this study was to compare the lncRNA and mRNA expression profiles, their related biological functions, and pathways in lung cancer.**

Methods: **Human lung cancer mRNA expression datasets were searched and downloaded from NCBI-GEO. LncRNA expression profiles in lung cancer were detected by the Agilent Human LncRNA Microarray V3.0. Differential expression analysis was conducted between cases and controls in mRNA and lncRNA. The starBase web server v2.0 was used to decipher lncRNA-protein interactions. DAVID Bioinformatics Resources 6.7 was used to perform GO Biological Processes and KEGG pathway enrichment analysis of these dysregulated mRNA and lncRNA target genes.**

Results: **The study showed that differentially expressed lncRNA target genes included almost all of the differential expression genes from mRNA. Furthermore, these dysregulated lncRNAs reflected more comprehensive impairment in functional enrichment than dysregulated mRNAs. In addition, five top downregulated lncRNAs had more remarkable fold changes than top downregulated mRNAs, especially FENDRR.**

Conclusions: **Amount of lncRNAs and mRNAs were differentially expressed in lung cancer. The degrees of difference in lncRNAs were more than mRNAs. It may provide valuable help for an effective strategy in diagnosis and prevention.**

(Clin. Lab. 2017;63:xx-xx. DOI: 10.7754/Clin.Lab.2016.160812)

Correspondence:
Yu-Ming Wang
Department of Clinical Laboratory
The Second Affiliated Hospital of
Kunming Medical University
Kunming 650032, Yunnan
China
Phone: +86 0871 65351281
Fax: +86 0871 65361017
Email: wangym992011@163.com

Yue-Dong Gao
State Key Laboratory of Genetic Resources and Evolution
Kunming Institute of Zoology
Chinese Academy of Sciences
Kunming 650223, Yunnan
China
Email: gaoyd@mail.kiz.ac.cn

Manuscript accepted August 29, 2016

KEY WORDS

lung cancer, lncRNA, mRNA, gene expression, integrated analysis

INTRODUCTION

Lung cancer is one of the most common malignancies worldwide. Early diagnosis and effective treatment are difficulties which have been supported by the death-statistics from the World Health Organization (WHO) database between 2000 and 2010. After age-adjustment in 60 countries, the state of cancer mortality has been assembled in several of the most common cancer types: lung cancer, gastric cancer, breast cancer, colon cancer, kidney cancer, and prostate cancer, etc. Globally, a total of 696,000 of those patients died over past few years (including 426,000 male and 271,000 female) [1]. Among them, the mortality of liver cancer and female lung cancer has been demonstrated with a rising tendency. Within developed countries, many people have been nagged by social risk factors (low income, low education level, ethnic minorities, divorced family) which led to risk scores increasing from 0 to 3, and even higher. Unfortunately, they have to face more social pressure which results in various cancers, tobacco-related cancer, and lung cancer [2]. For another, the average census data in 2012 of Chinese population (ASIRC) and the world's average population (ASIRW) showed that new incidence of cancers has increased to 264.85/100,000 (289.30/100,000 in males, and 239.15/100,000 in females). This new cancer incidence of China has accounted for 21.82% of global accumulation in 0 - 74 age range. Chinese cancer mortality has increased to 161.49/100,000 (198.99/100,000 in males, 122.06/100,000 in females). It represented 12.61% cumulative mortality rate worldwide in the same age bracket. Lung cancer patients are a large group within representing 77.4% of new cancer cases, as well as 84.5% of deaths [3].

Early diagnosis and proper treatments can effectively improve lung cancer prognosis. A novel chromosome 5p15.33, based on the high- or intermediate-quality data on death certificates according to the WHO mortality database, showed the evidence of polymorphism in five high-throughput sequencing projects [4]. ERK and KRAS mutations, EGFR gene fusion, and PTEN gene disorder which have been identified to be closely related with lung cancer [5-7]. Nevertheless, the interactive molecular mechanisms of lung cancer still remain poorly understood.

Long noncoding RNAs (lncRNAs) which refer to a group of RNAs with length more than 200 nucleotides limit protein-coding potential. LncRNAs regulate the expression of genes and play an important role in physiological processes at epigenetic, transcriptional, and post-transcriptional levels. They have various functions such as RNA processing [8], structural scaffolds [9], modulation of apoptosis and invasion [10], marker of cell fate [11], reprogramming of induced pluripotent stem cells [12], and chromatin modification [13]. In addition, lncRNAs can act as antisense transcripts or as decoys for splicing factors leading to splicing malfunctioning and as cancer-driving RNA regulators in gastric cancer, bladder cancer, and breast cancer [14-16]. Lnc RNAs are dysregulated in human cancers, playing either an oncogenic or tumor-suppressive role. Also, as a vital role in the development and prognosis of lung cancer, they can be identified as a kind of early diagnostic biomarker and therapeutic target. It is clear from the foregoing that there is increasing data demonstrating that the lncRNA regulation mechanism is of great significance, though lung cancer is considered to be a degradative disease.

However, it is not yet clear whether lncRNAs have better veracity and specificity than mRNAs in the diagnosis as bio-marker in lung adenocarcinoma. High expression of NEAT1 in non-small cell lung cancer was discovered in the diagnosis area of a curve (AUC) was 0.684 (95% CI: 0.619 - 0.750, $p < 0.001$), which revealed that lncRNA can be a clinical diagnostic biomarker [17]. Therefore, the aim of this study was to explore the expression profiles, related biological functions, and pathways of mRNA and lncRNA in lung cancer, as well as evaluate the diagnostic efficiency between significantly differentially expressed mRNA and lncRNA in lung cancer.

MATERIALS AND METHODS

Study subject

This study recruited 5 lung cancer patients for microarray analysis of lncRNAs. All patients were from Kunming, China. The diagnosis of lung cancer was confirmed by histopathology. This study was approved by the Ethics Committee of the First Affiliated Hospital of Kunming Medical University, Kunming, China. All patients who participated from 2012 to 2015 gave informed consent.

LncRNA detect and preprocessing

Total RNA was extracted from all the tumor tissues and adjacent normal tissues using Trizol reagent (Invitrogen, CA, USA) according to the manufacturer's instruction. The quantity of RNA was assessed with the nucleic acid/protein spectrophotometer (OD 260 nm, Eppendorf, Germany), and RNA integrity was assessed using standard denaturing agarose gel electrophoresis. The Agilent Human LncRNA Microarray V3.0 is designed for global profiling of lncRNAs in the human genome. The microarray hybridization and collection of expression data were performed by CapitalBio, Beijing, China.

Agilent Feature Extraction software (version 11.0.1.1) was used to handle the tiff format image data after hybrid scanning to obtain the raw data files. Data normal-

ization and QC analysis was done for each sample. We detected a total of 37,000 lncRNAs.

Microarray data collection and preprocessing

Human lung cancer microarray datasets were searched and downloaded from the NCBI-GEO database (http://www.ncbi.nlm.nih.gov/geo) in January 2016. We used the keyword of "lung cancer", "lung adenocarcinoma" and "lung tumor" to perform an accurate search. The data selection criteria were: (1) all datasets were genome-wide; (2) the samples of each data set must include lung cancer patients and controls; (3) the tissue must be lung; (4) non cell line samples and (5) raw data were available. Based on the above criteria, we finally chose 6 datasets for our integrated analysis (GSE7670, GSE18842, GSE19804, GSE31547, GSE31552, and GSE33532). The integrated datasets included 246 lung cancer patients and 234 controls. The details of all datasets are listed in Supplementary Table 1.

R statistical software version 3.2.2 was used to perform data preprocessing. We used the Robust Multichip Average (RMA) algorithm in the oligo package [18] to normalize the raw expression data and generate normalized gene expression intensity. Gene annotation, integration, and renormalization of different datasets were carried out using custom written Python code. For details see the previous publication [19]. The distribution of RMA processed and global renormalized gene expression values across all studies were shown in Supplementary Figure 1 and 2.

Bioinformatics analysis

Differential expression genes analysis used R statistical software version 3.2.2 and Bioconductor Library. The empirical Bayes algorithm (function "eBayes") in the limma package [20] was used to detect differentially expressed genes between lung cancer patients and controls. GeneSpring GX software v12.0 was used to calculate of differentially expressed lncRNAs. Up-regulated genes and lncRNAs were considered as a logarithmic transformed fold-change (log(FC)) ≥ 1 and a false discovery rate (FDR) adjusted p-value ≤ 0.05. Down-regulated genes and lncRNAs were considered as a log (FC) ≤ -1 and a FDR-p-value ≤ 0.05. The starBase web server v2.0 (http://starbase.sysu.edu.cn/index.php) was used to decipher lncRNA-protein interactions. We used DAVID Bioinformatics Resources 6.7 [21] to perform GO Biological Processes and KEGG pathway enrichment analysis. The input parameters were differential expression genes, differential expression lncRNA target genes, and their overlapped genes. The enrichment significance level was set as a p-value ≤ 0.01. To compare the results between differential expression genes and differential expression lncRNA target genes, we used the Venn diagram in the VennDiagram package to show the up- and down-regulated genes as well as the GO Biological Processes results.

RESULTS

Overview of differential lncRNAs and mRNAs

In total, we detected 1484 differentially expressed lncRNAs in this study, including 535 up-regulated lncRNAs and 949 down-regulated lncRNAs. The lnc RNA-protein interaction results showed that 292 target genes were up-regulated and 500 down-regulated. Additionally, the integrated results showed there were 187 up-regulated genes and 406 down-regulated genes. Both results showed that the number of down-regulated genes were more than up-regulated genes. The Venn diagram of these differential expression genes is shown in Figure 1. The differentially expressed lncRNA target genes included almost all of the differential expression genes from mRNA (the mapped up- and down-regulated genes were 186 and 402, respectively).

Enrichment analysis in lncRNA and mRNA

Figure 2 showed the GO Biological Processes enrichment results. The enrichment results of differential expression of lncRNA target genes included almost all enrichment results of the differential expression genes from mRNA. The most significant Biological Processes in three groups were response to wounding (p = 4.60E-15, 4.40E-18, and 2.80E-18 for lncRNA target genes, mRNA, and overlapped genes, respectively). Furthermore, the top 5 GO Biological Processes were the same in the mRNA and overlapped genes. In addition to response to wounding, the remaining are inflammatory response (p = 1.60E-10 and 1.20E-10 for mRNA and overlapped genes, respectively), vasculature development (p = 9.20E-10 and 3.20E-9 for mRNA and overlapped genes, respectively), regulation of cell proliferation (p = 2.10E-9 and 1.50E-9 for mRNA and overlapped genes, respectively), and regulation of response to external stimulus (p = 2.20E-9 and 1.90E-9 for mRNA and overlapped genes, respectively). These Biological Processes were all significantly enriched in lncRNA target genes.

The KEGG Pathway enrichment results are shown in Figure 3. The common KEGG Pathways in three groups were ECM-receptor interaction, complement and coagulation cascades, cell cycle, cytokine-cytokine receptor interaction, cell adhesion molecules (CAMs), PPAR signaling pathway, focal adhesion, and TGF-β signaling pathway. Three pathways were only significantly enriched in lncRNA target genes (chemokine signaling pathway, leukocyte transendothelial migration, and p53 signaling pathway). The renin-angiotensin system was only significantly enriched in mRNA. The gene expression profiles of these 12 pathways are shown in Figure 4. All these pathways showed serious effects. In p53 signaling pathway, ECM-receptor interaction and cell cycle showed a large number of up-regulated genes. Other pathways mostly showed down-regulated genes. Compared to the enrichment results from the lncRNA target genes and mRNA, the enriched KEGG Pathways from lncRNA target genes presented a more compre-

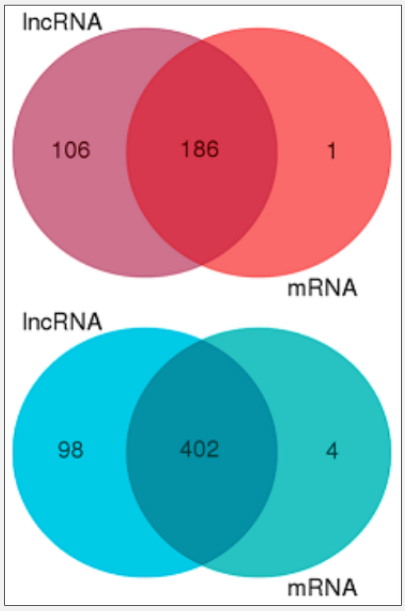


Figure 1. Venn Diagram of differentially expressed lncRNA target genes and mRNAs in lung cancer.

The red cycle showed the up-regulated genes and the blue cycle showed the down-regulated genes.

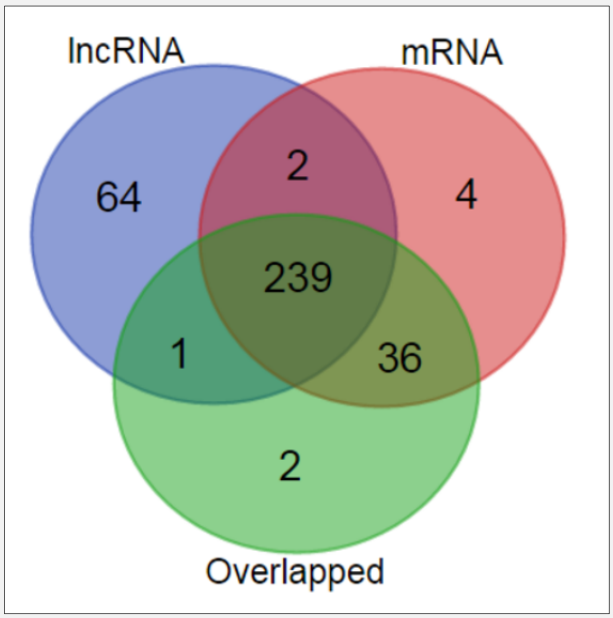


Figure 2. Venn Diagram of enriched GO Biological Process (GOBP) in lung cancer.

The blue cycle represents the enriched GOBP using differentially expressed lncRNA target genes, the red cycle represents the enriched GOBP using differentially expressed mRNAs, and the green cycle represents the enriched GOBP using overlapped genes.

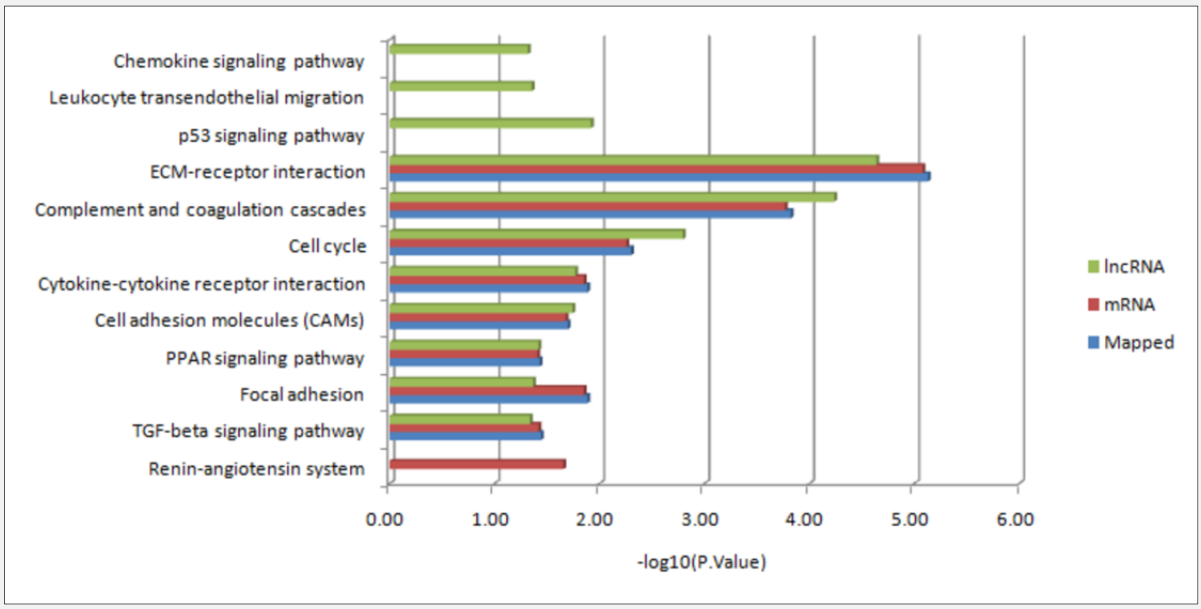


Figure 3. Enriched KEGG pathways in lung cancer.

The green column represents the enriched KEGG pathways using differentially expressed lncRNA target genes, and the red column represents the enriched KEGG pathways using differentially expressed mRNAs, and the green cycle represents the enriched KEGG pathways using overlapped genes.

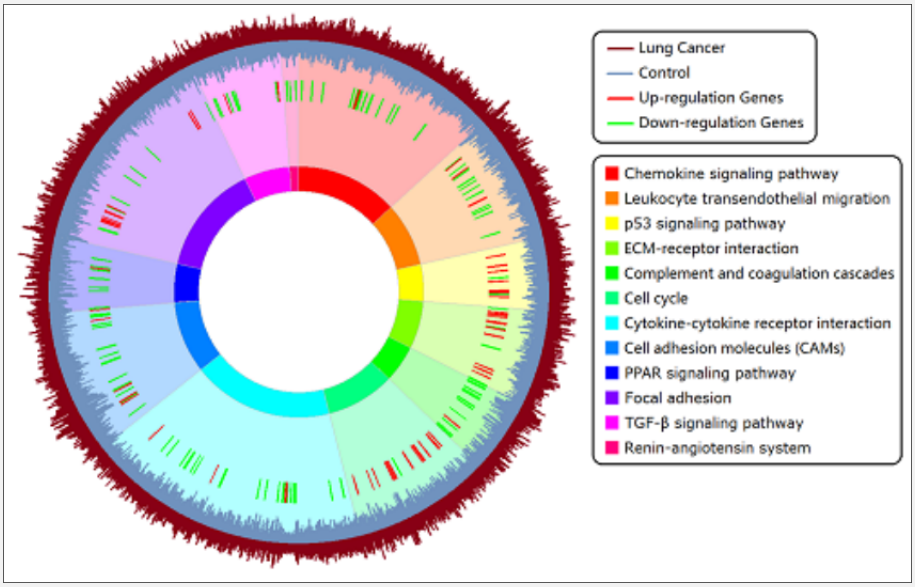


Figure 4. Gene expression profiles of enriched KEGG pathways in lung cancer.

12 enriched KEGG pathways are represented by different rainbow colors. The length of the first layer lines outside the circle represents the expression value in lung cancer patients and the length of the second layer lines within the circle represents the expression value in controls. The up- and down-regulation genes are marked as red and green lines, respectively, in the third layer.

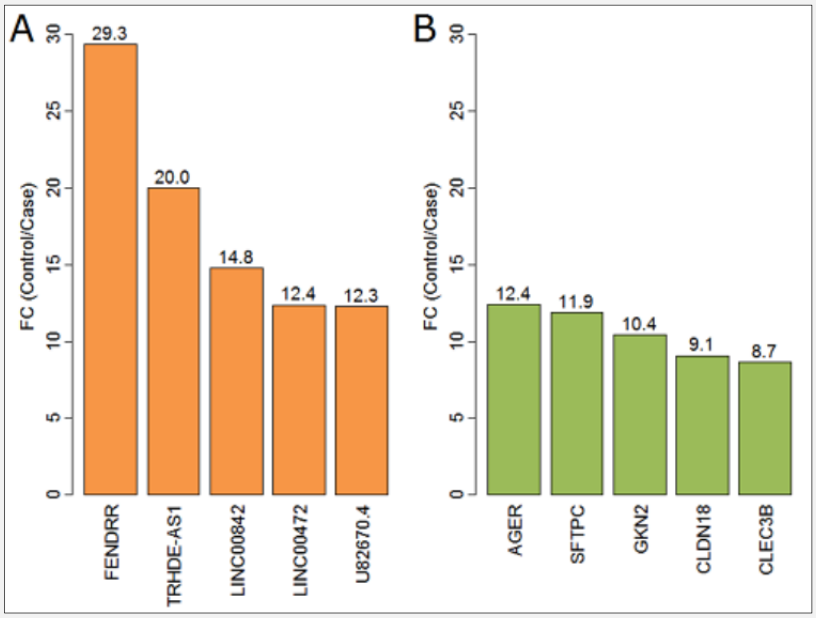


Figure 5. Bar plot of the top five differentially expressed lncRNAs and mRNAs.

Panel A showed the fold-change of lncRNA. Panel B showed the fold-change of mRNA. All five lncRNA and mRNA were down-regulated. The value of the fold-change showed in the top of the column.

hensive picture of the situation of the disease.

Top 5 differential expressed lncRNAs and mRNAs

Figure 5 was the fold-change (control/case) of the top 5 differential expressed lncRNAs and mRNAs. All these lncRNAs and mRNAs were down-regulated. The most significant down-regulated lncRNA was FENDRR (FC = 29.3, p = 0.021). And the most down-regulated mRNA was AGER (FC = 12.4, p = 1.0E-70). However, the FC of the AGER was lower than the half of the FC of FENDRR. This value (12.4) was the same as the FC of LINC00472, the fourth lncRNA in the top 5 lncRNAs.

DISCUSSION

This study showed that differentially expressed lncRNAs give a more comprehensive view in related genes, biological functions, and pathways than differentially expressed mRNAs. In addition, the FC of the top 5 differentially expressed lncRNAs was more than the top 5 mRNAs, especially FENDRR.

In GO Biological Processes analysis, the response to wounding, inflammatory response, vasculature development, regulation of cell proliferation, regulation of response to external stimulus and other 234 biological processes were both enriched in differentially expressed lncRNA target genes and differentially expressed mRNAs. Additionally, we have identified several pathways such as the p53 signaling pathway, TGF-β signaling pathway, PPAR signaling pathway, ECM receptor interaction, complement and coagulation cascade, cell cycle, and 6 other pathways that are closely related to lung cancer. Among them, reports have demonstrated that the mutant p53 protein in lung cancer can induce specific humoral responses to enforce the existence of the evaluation of serum antibody [22]. Epithelial-mesenchymal transition (EMT) gives a resistance to complement dependent cytotoxicity (CDC) and promotes lung cancer prognosis. TGF-β induced EMT and CD59 expression to produce tumor immune escape. Together, CD59 inhibition can be used as adjuvant to enhance antibody mediated therapeutic effect and the inhibition of lung cancer metastasis [23]. Additionally, reducing DNA repair response and interfering with the cell cycle signal to inhibit apoptosis is significant for NSCLC resistance treatment [24]. Furthermore, interstitial extracellular matrix (ECM) is involved in tumor microenvironment such as collagen type I (CoI) and promotes lncRNA HOTAIR expression in non-small cell lung cancer cells. In the present study, further investigation of HOTAIR and other lincRNAs in lung tumorigenesis is warranted.

We identified the top 5 down-regulated lncRNAs (FENDRR, TRHDE-AS1, LINC00842, LINC00472 and U82670.4) and mRNAs (AGER, SFTPC, GKN2, CLDN18 and CLEC3B) in lung cancer. The specific molecular mechanism by which lncRNA regulates lung cancer remains largely unknown. FENDRR is the specific gene which is expressed in mammalian embryo [25]. A recent study found that its role in the formation of body wall and myocardium in mice. Lacking this expression, the fetal cell differentiation process will be induced by transcription factor expression stemming from up-regulation of FENDRR [26]. LncRNA FENDRR binds to both the PRC2 and TrxG/MLL complexes and expression is decreased which is closely related to gastric cancer metastasis [27]. In this study, we focus on its specificity which is differentially down-expressed in lung adenocarcinoma.

Among them, LINC00472 is one of the members which expresses abnormally in breast cancer which is a novel long intergenic non-coding RNA. One report has shown that LINC00472 expression in breast tumor samples using RT-qPCR, performed a meta-analysis of over 20 microarray datasets from the Gene Expression Omnibus (GEO) database, and the effect of LINC00472 expression on cell proliferation and migration in breast cancer cells transfected with an expressing vector. High LINC00472 expression was associated with less aggressive breast tumors and more favorable disease outcomes and it had significantly reduced risk of relapse and death compared to those with low expression. High LINC00472 was also associated with favorable molecular subtypes, Luminal A or normal-like tumors [28]. Its expression has also been found more possible by promoter methylation than by the alteration of gene copy number, and that may have clinical implications in breast cancer management [29]. It has rarely been seen in lung adenocarcinoma research before.

AGER is a member of the immunoglobulin superfamily of cell surface receptors. After survival analysis from the 7 groups of patients of non-small cell lung cancer, AGER and SFTPC show a difference in the adenocarcinoma affecting prognosis. They have become lung biomarker candidates [30]. Likewise, SFTPC and SFTPB mRNA levels were further validated by real-time reverse transcription PCR-based quantification in 16 NSCLC tumors samples. Several promising molecular tumor cell markers were identified, including the new SFTPC mRNAs [31]. It is sure that SFTPC exists in lung cancer, but also difficult to test in peripheral blood. In the study, genetic aberrations have shown that tobacco-related SFTPC was deleted exclusively in tumor tissues (71%) [32].

CONCLUSION

Dysregulated lncRNAs showed more comprehensive impairment of cellular biological functions and pathways than mRNAs in lung cancer. Furthermore, these top down-regulated lncRNAs had more remarkable fold changes than top down-regulated mRNAs. Thus, we speculate that lncRNAs may have higher sensitivity and accuracy in lung cancer diagnosis and prevention than mRNAs. A limitation in this study was that the lncRNA samples were relatively small. Therefore, future studies should verify these results in a larger population.

Acknowledgement:

This work was supported by grants from the National Natural Science Foundation of China (Grant No.: 81460064, 81460325), the General Project of Science and Technology Department of Yunnan Province-Kunming Medical University Union Fund (Grant No.: 2014FB034), the Instruments Function Deployment Foundation of CAS (Grant no. 2014GK01), the CAS Key Technology Talent Program, and the Leading Talent Project of the health system in Yunnan province (L-201202). We thank all the participants of the study.

Declaration of Interest:

None of the authors claim any conflict of interest.

References:

1. Hashim D, Boffetta P, La Vecchia C, et al. The global decrease in cancer mortality: trends and disparities. Ann Oncol 2016;27:926-33 (PMID: 26802157).

2. Caleyachetty R, Tehranifar P, Genkinger JM, Echouffo-Tcheugui JB, Muennig P. Cumulative social risk exposure and risk of cancer mortality in adulthood. BMC Cancer 2015;15:945 (PMID: 26675142).

3. Chen W, Zheng R, Zuo T, Zeng H, Zhang S, He J. National cancer incidence and mortality in China, 2012. Chin J Cancer Res 2016;28:1-11 (PMID: 27041922).

4. Kachuri L, Amos CI, McKay JD, et al. Fine mapping of chromosome 5p15.33 based on a targeted deep sequencing and high density genotyping identifies novel lung cancer susceptibility loci. Carcinogenesis 2016;37:96-105 (PMID: 26590902).

5. Pal S, Amin PJ, Sainis KB, Shankar BS. Potential Role of TRAIL in Metastasis of Mutant KRAS Expressing Lung Adenocarcinoma. Cancer Microenviron 2016 Apr 23. [Epub ahead of print] (PMID: 27106232).

6. Konduri K, Gallant JN, Chae YK, et al. EGFR Fusions as Novel Therapeutic Targets in Lung Cancer. Cancer Discov 2016;6(6): 601-11 (PMID: 27102076).

7. Kohnoh T, Hashimoto N, Ando A, et al. Hypoxia-induced modulation of PTEN activity and EMT phenotypes in lung cancers. Cancer Cell Int 2016;16:33 (PMID: 27095949).

8. Gong C, Maquat LE. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. Nature 2011;470:284-8 (PMID: 21307942).

9. Clemson CM, Hutchinson JN, Sara SA, et al. An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. Mol Cell 2009;33:717-726 (PMID: 19217333).

10. Khaitan D, Dinger ME, Mazar J, et al. The melanoma-upregulated long noncoding RNA SPRY4-IT1 modulates apoptosis and invasion. Cancer Res 2011;71:3852-62 (PMID: 21558391).

11. Ginger MR, Shore AN, Contreras A, et al. A noncoding RNA is a potential marker of cell fate during mammary gland development. Proc Natl Acad Sci USA 2006;103:5781-6 (PMID: 16574773).

12. Loewer S, Cabili MN, Guttman M, et al. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. Nat Genet 2010; 42:1113-7 (PMID: 21057500).

13. Kanduri C. Kcnq1ot1: a chromatin regulatory RNA. Semin Cell Dev Biol 2011;22:343-50 (PMID: 21345374).

14. Zhu S, Mao J, Shao Y, et al. Reduced expression of the long non-coding RNA AI364715 in gastric cancer and its clinical significance. Tumour Biol 2015;36:8041-5 (PMID: 25971582).

15. Zhuang J, Lu Q, Shen B, et al. TGF β1 secreted by cancer-associated fibroblasts induces epithelial-mesenchymal transition of bladder cancer cells through lncRNA-ZEB2NAT. Sci Rep 2015; 5:11924 (PMID: 26152796).

16. Zhang XF, Liu T, Li Y, Li S. Overexpression of long non-coding RNA CCAT1 is a novel biomarker of poor prognosis in patients with breast cancer. Int J Clin Exp Pathol 2015;8:9440-5 (PMID: 26464701).

17. Pan LJ, Zhong TF, Tang RX, et al. Upregulation and Clinicopathological Significance of Long Non-coding NEAT1 RNA in NSCLC Tissues. Asian Pac J Cancer Prev 2015;16:2851-5 (PMID: 25854373).

18. Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. Bioinformatics 2010;26:2363-7 (PMID: 20688976).

19. Li WX, Dai SX, Liu JQ, Wang Q, Li GH, Huang JF. Integrated Analysis of Alzheimer's Disease and Schizophrenia Dataset Revealed Different Expression Pattern in Learning and Memory. J Alzheimers Dis 2016;51:417-25 (PMID: 26890750).

20. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47 (PMID: 25605792).

21. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009;4:44-57 (PMID: 19131956).

22. Lubin R, Zalcman G, Bouchet L, et al. Serum P53 Antibodies as Early Markers of Lung-Cancer. Nature Medicine 1995;1:701-2 (PMID: 7585154).

23. Goswami MT, Reka AK, Kurapati H, et al. Regulation of complement-dependent cytotoxicity by TGF β induced epithelial-mesenchymal transition. Oncogene 2016;35:1888-98 (PMID: 26148233).

24. Lundholm L, Haag P, Zong D, et al. Resistance to DNA-damaging treatment in non-small cell lung cancer tumor-initiating cells involves reduced DNA-PK/ATM activation and diminished cell cycle arrest. Cell Death Dis 2013;4:e478 (PMID: 23370278).

25. Grote P, Herrmann BG. The long non-coding RNA Fendrr links epigenetic control mechanisms to gene regulatory networks in mammalian embryogenesis. RNA Biol 2013;10:1579-85 (PMID: 24036695).

26. Grote P, Wittler L, Hendrix D, et al. The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. Dev Cell 2013;24:206-14 (PMID: 23369715).

27. Xu TP, Huang MD, Xia R, et al. Decreased expression of the long non-coding RNA FENDRR is associated with poor prognosis in gastric cancer and FENDRR regulates gastric cancer cell metastasis by affecting fibronectin1 expression. J Hematol Oncol 2014;7:63 (PMID: 25167886).

28. Shen Y, Katsaros D, Loo LW, et al. Prognostic and predictive values of long non-coding RNA LINC00472 in breast cancer. Oncotarget 2015;6:8579-92 (PMID: 25865225).

29. Shen Y, Wang Z, Loo LW, et al. LINC00472 expression is regulated by promoter methylation and associated with disease-free survival in patients with grade 2 breast cancer. Breast Cancer Res Treat 2015;54:473-82 (PMID: 26564482).

30. Lindskog C, Fagerberg L, Hallstrom B, et al. The lung-specific proteome defined by integration of transcriptomics and antibody-based profiling. FASEB J 2014;28:5184-96 (PMID: 25169055).

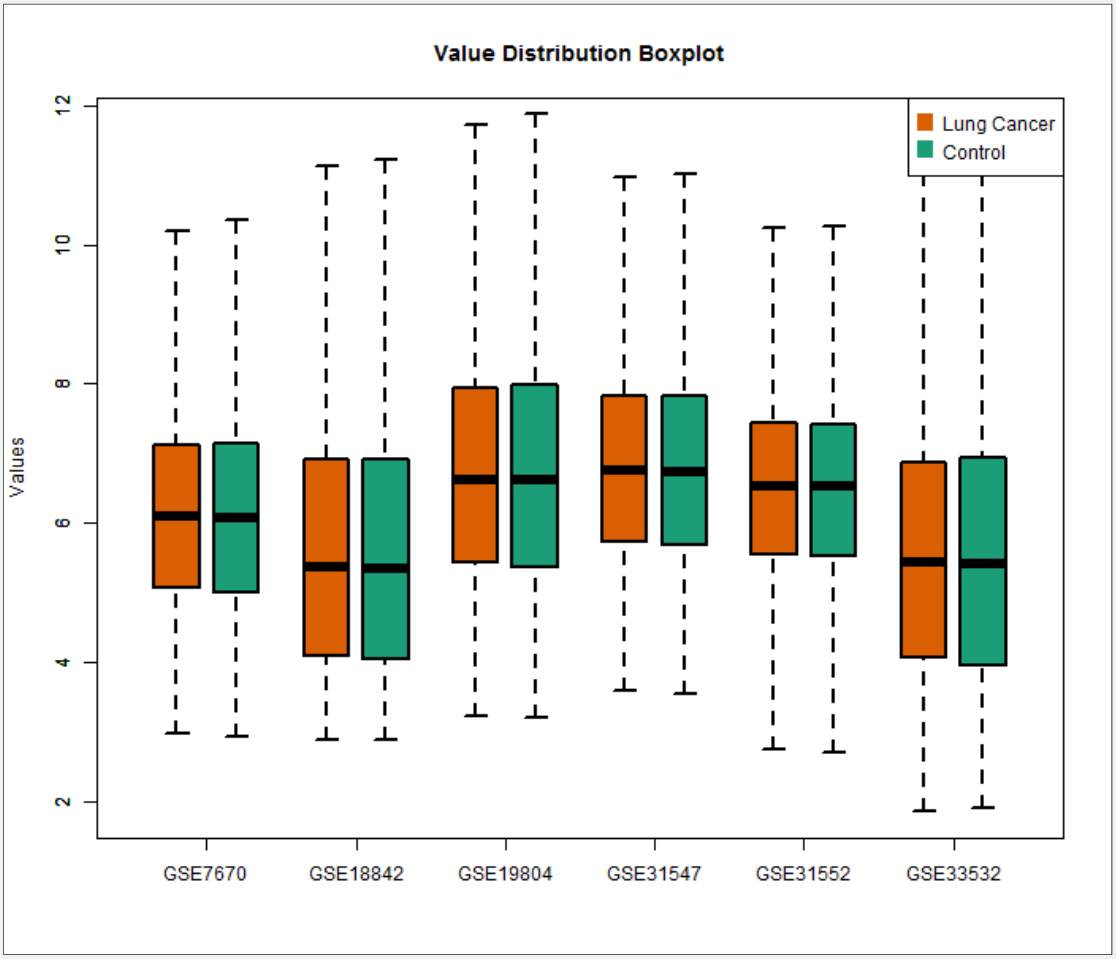
31. Nordgard O, Singh G, Solberg S, et al. Novel molecular tumor cell markers in regional lymph nodes and blood samples from patients undergoing surgery for non-small cell lung cancer. PLoS One 2013;8:e62153 (PMID: 23671585).

32. Li R, Todd NW, Qiu Q, et al. Genetic deletions in sputum as diagnostic markers for early detection of stage I non-small cell lung cancer. Clin Cancer Res 2007;13:482-7 (PMID: 17255269).

Supplementary Table 1. Summary of lung cancer datasets.

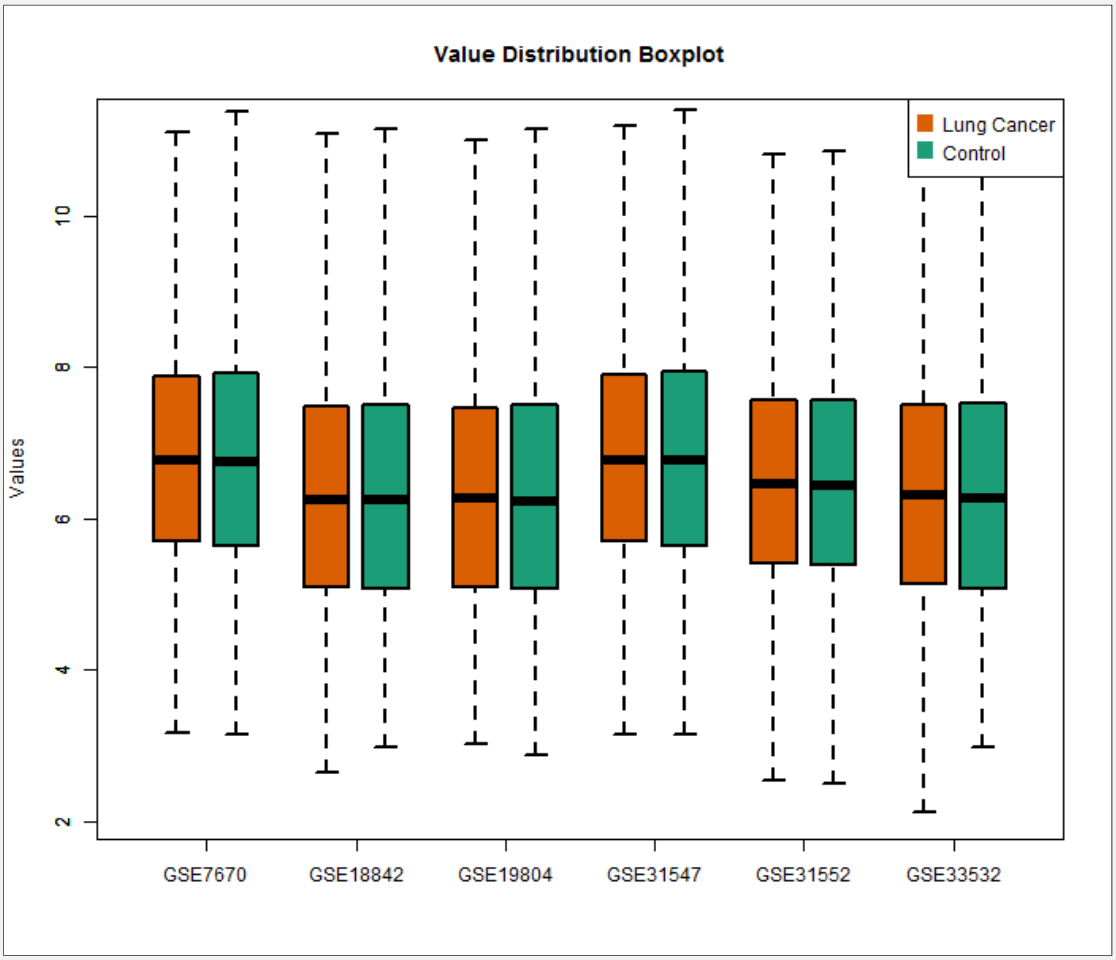
Series ID	Contributor	Samples [1]	Title
GSE7670	Su LJ, 2007	66 (54)	Expression data from lung cancer
GSE18842	Ramos AS, 2009	91 (91)	Gene expression analysis of human lung cancer and control samples
GSE19804	Lu TP, 2010	120 (120)	Genome-wide screening of transcriptional modulation in non-smoking female lung cancer in Taiwan
GSE31547	Girard L, 2011	50 (50)	MSKCC-A Primary Lung Cancer Specimens
GSE31552	Marquardt G, 2011	131 (125)	Expression Data from human lung tissue of patients with non-small cell lung cancer
GSE33532	Meister M, 2011	100 (40)	Intratumor heterogeneity of gene expression profiles in early stage non-small cell lung cancer

[1] **All samples of this dataset (samples used in this study).**



Supplementary Figure 1. The distribution of RMA processed gene expression values of lung cancer datasets.

The orange box showed the distribution of gene expression values in lung cancer patients and the green box showed the distribution of gene expression values in controls. There was a relatively large deviation in the distribution of gene expression values across these studies.



Supplementary Figure 2. The distribution of global renormalized (after RMA processed) gene expression values of lung cancer datasets.

The orange box showed the distribution of gene expression values in lung cancer patients and the green box showed the distribution of gene expression values in controls. The distribution of gene expression values across these studies had a relatively consistent range.