Tumor Biol. (2015) 36:9969–9978
DOI 10.1007/s13277-015-3714-6

RESEARCH ARTICLE

Differentially expressed protein-coding genes and long noncoding RNA in early-stage lung cancer

Ming Li[1] · Mantang Qiu[1,2] · Youtao Xu[1,3] · Qixing Mao[1,2] · Jie Wang[1,4] · Gaochao Dong[1,4] · Wenjia Xia[1,2] · Rong Yin[1] · Lin Xu[1]

Received: 30 May 2015 / Accepted: 23 June 2015 / Published online: 16 July 2015
© International Society of Oncology and BioMarkers (ISOBM) 2015

Abstract Due to the application of low-dose computed tomography screening, more and more early-stage lung cancers have been diagnosed. Thus, it is essential to characterize the gene expression profile of early-stage lung cancer to develop potential biomarkers for early diagnosis and therapeutic targets. Here, we analyzed microarray data of 181 early-stage lung cancer patients. By comparing gene expression between different tumor and lymph node metastasis stages, we identified various differentially expressed protein-coding genes and long noncoding RNA (lncRNA) in the comparisons of T2 vs. T2 and N1- vs. N0-stage lung cancer. Functional analyses revealed that these differentially expressed genes were enriched in various tumorigenesis or metastasis-related pathways. Survival analysis indicated that two protein-coding genes, C7 and SCN7A, were significantly associated survival of lung cancer. Notably, a novel lncRNA, LINC00313, was highly expressed in both T2- and N1-stage lung cancers. On the other hand, LINC00313 was also upregulated in lung cancer and metastasized lung cancer tissues, compared with adjacent lung tissues and primary lung cancer tissues. Additionally, higher expression level of LINC00313 indicated poor prognosis of lung cancer (hazard ratio=0.658). Overall, we characterized the expression profiles of protein-coding genes and lncRNA in early-stage lung cancer and found that LINC00313 could be a biomarker for lung cancer.

Keywords Lung cancer · lncRNA · Biomarker · Survival

Ming Li and Mantang Qiu contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s13277-015-3714-6) contains supplementary material, which is available to authorized users.

✉ Rong Yin
yinhero001@126.com

✉ Lin Xu
xulin83cn@gmail.com

[1] Department of Thoracic Surgery, Nanjing Medical University Affiliated Cancer Hospital, Jiangsu Key Laboratory of Molecular and Translational Cancer Research, Cancer Institute of Jiangsu Province, Baiziting 42, Nanjing 210009, China

[2] The Fourth Clinical College of Nanjing Medical University, Nanjing 210000, China

[3] The First Clinical College of Nanjing Medical University, Nanjing 210000, China

[4] Department of Scientific Research, Nanjing Medical University Affiliated Cancer Hospital, Cancer Institute of Jiangsu Province, Nanjing 210009, China

Introduction

Lung cancer is the leading cause of cancer-related death [1]. According to the randomized controlled National Lung Screening Trial (NLST), the lung cancer-related mortality and overall mortality have decreased due to the application of low-dose computed tomography (CT) instead of chest radiography [2]. With the application of low-dose CT screening, more and more lung cancer cases are diagnosed at early stage. For patients with early-stage lung cancer detected by screening, concerns have raised about the appropriate treatment for these patients [3–5]. Therefore, it is important to characterize the molecular basis and understand the altered gene expression in early-stage lung cancer. Additionally, prognostic biomarkers and makers that are predictive of metastasis and benefit from chemotherapy are needed for early-stage lung cancer [6, 7]. However, few studies have focused on these issues.

Gene expression microarray is a feasible and effective approach to characterize gene expression profile and searching

messenger RNA (mRNA)-based biomarkers. For lung cancer, microarrays have been widely used, which provide abundant resource for data mining [8–10]. Due to the advance of high-throughput technology, evidence has demonstrated that long noncoding RNA (lncRNA) is actively transcribed from human genome and plays an important role in all aspects of tumor biology [11–13]. Compared with protein-coding genes, lncRNA exhibits stronger cell-type specific expression manner [14], suggesting that lncRNA could be a potential biomarker [15]. Given that a number of probe sets were matched with lncRNA, reannotation of published microarray data and analyzed lncRNA expression profile is a feasible and widely used method [16–18].

In this study, we performed data mining of the GSE50081 dataset [9], which includes gene expression data of 181 early-stage lung cancer patients. We compared the protein-coding gene and lncRNA expression profiles between different tumor and lymph node stages and performed functional annotation of the differentially expressed protein-coding genes. SCN7A, C7, and LINC00313 were associated survival of lung cancer.

Methods and materials

Dataset and calculation of differentially expressed genes

Gene Expression Omnibus is a public online database that has various high-throughput data, including microarray. In the present study, we selected the GSE50081 dataset for further data mining [9]. The GSE50081 dataset consists of microarray data of 181 lung cancer patients, including TNM stages and survival data. The Affymetrix Human Genome U133 Plus 2.0 Array, which is widely used in various research areas, was utilized in the GSE50081 data set. For microarray data analysis, the processed series matrix file was first downloaded. Since the series matrix data has already been background subtracted and normalized by RMA method, the data was subjected to differentially expressed gene detection. The differentially expressed genes were calculated by the Limma algorithm [19], and P value<0.05 was considered as significant. RNA sequencing data of lung cancer from The Cancer Genome Atlas (TCGA) were accessed through the website lncRNAtor (http://lncrnator.ewha.ac.kr/).

Probe set annotation

Sequences of lncRNA were downloaded from the LNCipedia (http://www.lncipedia.org/) and 79,586 lncRNA larger than 200 nt were downloaded. Sequences of Affymetrix Human Genome U133 Plus 2.0 Array probe set were downloaded from the Affymetrix website. The probe sets were reannotated by Blast software, and 12,156 lncRNA completely matched with probe sets were identified.

Gene Ontology and KEGG pathway analysis

Gene Ontology (GO) analysis was applied to analyze the main function of the differential expression genes according to the Gene Ontology (www.geneontology.org), which can organize genes into hierarchical categories and uncover the gene regulatory network on the basis of biological process and molecular function [20, 21]. Specifically, two-side Fisher's exact test and X^2 test were used to classify the GO category, and the false discovery rate (FDR) [22] was calculated to correct the P value; the smaller the FDR, the smaller the error in judging the P value. The FDR was defined as $FDR = 1-\frac{N_k}{T}$ where N_k refers to the number of Fisher's test P values less than X^2 test P values. We computed P values for the GOs of all the differential genes. Enrichment provides a measure of the significance of the function: as the enrichment increases, the corresponding function is more specific, which helps us to find those GOs with more concrete function description in the experiment. Within the significant category, the enrichment Re was given by $\mathrm{Re}=(n_f/n)/(N_f/N)$ where "n_f" is the number of flagged genes within the particular category, "n" is the total number of genes within the same category, "N_f" is the number of flagged genes in the entire microarray, and "N" is the total number of genes in the microarray [23].

Pathway analysis was used to find out the significant pathway of the differential genes according to Kyoto Encyclopedia of Genes and Genomes (KEGG). Still, we turn to Fisher's exact test and X^2 test to select the significant pathway, and the threshold of significance was defined by P value and FDR. The enrichment Re was calculated like the equation above [24–26]. KEGG pathway analysis allowed us to determine the biological pathways for which a significant enrichment of differentially expressed mRNAs existed (P<0.05 was considered statistically significant).

Statistical analysis

Kaplan-Meier survival and univariate Cox proportional hazards regression analyses were conducted to explore the prognostic value of differentially expressed coding genes or lncRNA. According to the median expression value of a specific target gene, patients were classified as "high expression" or "low expression", and survival analysis was conducted between the two groups. LINC00313 expression level between lung cancer tissues and adjacent lung tissues, primary lung cancer tissues, and metastasized lung cancer tissues were calculated by Student's t test, and P<0.05 was statistically significant. All statistical analyses were performed with SPSS software (version 18.0, SPSS Inc.).

Results

Differential expression profile between T2 and T1 stages

We first analyzed the gene expression among different tumor stages. Compared with T1-stage lung cancer, there were 94 protein-coding genes upregulated and 228 genes downregulated in T2-stage lung cancer. KEGG analysis showed that the differentially expressed genes were significantly enriched in the pathways of "cell cycle", "p53 signaling pathway", "pathways in cancer", and other cancer-related pathways (Fig. 1a). For lncRNA, we found that 238 lncRNAs were upregulated and 217 were downregulated in T2-stage lung cancer (with the threshold $P<0.01$). The top differentially expressed genes were shown in Tables 1 and 2, and the full lists were provided in the supplementary materials.

Differential expression profile between N1 and N0 stages

Lymph node metastasis is an important prognostic factor for NSCLC. By comparing patients with and without lymph node metastasis, we found that 47 protein-coding genes were upregulated and 163 were downregulated in N1-stage patients compared with N0-stage patients. On the other hand, 210 lncRNA were upregulated and 81 were downregulated in N1-stage patients. Functional GO annotation analysis showed

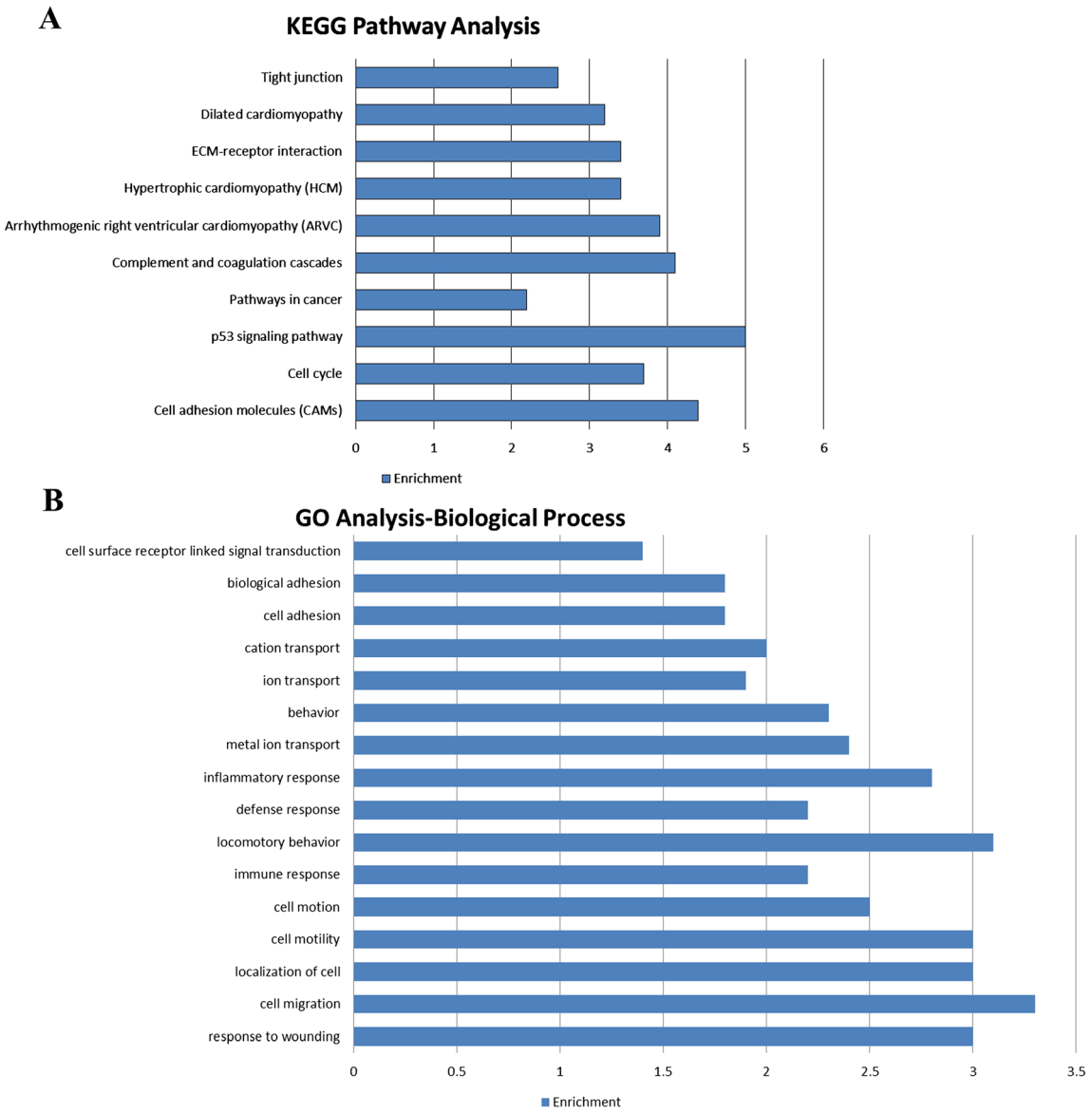

Fig. 1 Functional analysis of differentially expressed genes between different tumor and lymph node stages. KEGG analysis for differentially expressed protein-coding genes between T2- and T1-stage lung cancer (**a**). Gene Ontology (biological process) analysis of differentially expressed protein-coding genes between N1- and N0-stage lung cancer (**b**). The items with $P<0.05$ were considered as significantly enriched. The top enriched items and enrichment score were shown

Table 1 Differentially expressed protein-coding genes between different tumor stages (T2 vs. T1)

Gene symbol	*P* value	FC	Ensembl transcript ID	Transcript length	Description
SFTA3	6.84E-04	−2.969047	ENST00000521945	1116	Surfactant associated 3
SFTPC	3.98E-04	−2.907945	ENST00000318561	997	Surfactant protein C
AQP4	2.65E-05	−2.751084	ENST00000383168	5274	Aquaporin 4
CPB2	5.80E-07	−2.713209	ENST00000181383	1717	Carboxypeptidase B2 (plasma)
SFTPA2	1.82E-03	−2.713209	ENST00000372325	2189	Surfactant protein A2
SCGB3A2	4.84E-03	−2.566852	ENST00000504320	440	Secretoglobin, family 3A, member 2
C16orf89	7.16E-04	−2.549121	ENST00000586629	576	Chromosome 16 open reading frame 89
C4BPA	1.37E-03	−2.496661	ENST00000367070	2243	Complement component 4 binding protein, alpha
PEBP4	3.74E-05	−2.479415	ENST00000256404	895	Phosphatidylethanolamine-binding protein 4
TOX3	9.07E-04	−2.378414	ENST00000407228	3124	TOX high mobility group box family member 3
TRIM29	5.61E-03	2.0705298	ENST00000341846	3328	Tripartite motif containing 29
LY6D	1.48E-03	2.0994334	ENST00000301263	804	Lymphocyte antigen 6 complex, locus D
CXCL8	2.62E-04	2.1435469	ENST00000307407	1705	Chemokine (C-X-C motif) ligand 8
MMP10	3.63E-03	2.1435469	ENST00000279441	1758	Matrix metallopeptidase 10 (stromelysin 2)
KRT17	1.72E-02	2.1435469	ENST00000311208	1524	Keratin 17
MMP1	5.57E-03	2.1885874	ENST00000315274	1970	Matrix metallopeptidase 1 (interstitial collagenase)
S100P	6.01E-03	2.250117	ENST00000296370	1279	S100 calcium binding protein P
KRT6B	1.44E-02	2.4283898	ENST00000252252	2282	Keratin 6B
SERPINB5	2.70E-03	2.4452806	ENST00000382771	2783	Serpin peptidase inhibitor, clade B (ovalbumin), member 5
KRT6A	2.07E-03	3.7842306	ENST00000330722	2310	Keratin 6A

FC>0, upregulation; FC<0, downregulation

FC fold change

Table 2 Differentially expressed lncRNAs between different tumor stages (T2 vs. T1)

lncRNA	*P* value	FC	lncRNA length	chr	Strand	Start	End
lnc-SFTA3-1:1	6.84E-04	−2.96904714	1116	chr14	−	36942411	36988726
lnc-DZIP1L-1:1	4.76E-04	−2.14354693	1128	chr3	−	137749544	137750672
lnc-NDNF-1:1	2.05E-04	−2.12874036	1897	chr4	−	121954486	121956383
lnc-AC009336.1-2:13	3.64E-05	−1.86218964	3812	chr2	−	177037916	177053686
lnc-SPINK2-4:1	7.20E-03	−1.73507737	2540	chr4	−	57544923	57547463
lnc-IRS1-3:1	3.88E-04	−1.72070526	1811	chr2	−	227867429	227869240
lnc-ZDHHC11B-1:1	2.23E-04	−1.70290742	2733	chr5	−	710474	732786
lnc-SNX13-2:5	9.61E-03	−1.68179283	2978	chr7	−	17472875	17496748
lnc-C10orf31-7:13	4.12E-04	−1.64832417	597	chr10	−	10826399	10836920
lnc-CHST10-2:1	9.44E-04	−1.64832417	6696	chr2	−	100889908	100898479
lnc-PHYHIP-2:1	4.77E-04	1.383190629	2989	chr8	−	22277141	22280192
lnc-GNB2L1-3:1	4.37E-03	1.390881972	599	chr5	−	180630117	180630820
lnc-ANXA8-1:1	1.85E-03	1.398616083	1872	chr10	+	48276678	48279171
lnc-HPS4-4:1	3.98E-04	1.414213562	1233	chr22	−	26838762	26840941
lnc-HILPDA-1:2	3.99E-03	1.422077411	1370	chr7	+	128095956	128098472
lnc-CBX2-5:1	4.79E-03	1.434949535	2354	chr17	+	77759428	77761782
lnc-PIGZ-2:1	3.58E-03	1.494849249	707	chr3	−	196728612	196729319
lnc-CCDC103-1:1	3.68E-04	1.62788637	1844	chr17	+	43002078	43005641
lnc-GNB3-1:1	3.27E-04	1.669018562	1414	chr12	+	6957943	6959357
lnc-MMP3-1:1	8.93E-03	1.960198831	1873	chr11	−	102733466	102745764

FC>0, upregulation; FC<0, downregulation

FC fold change

that the altered genes were associated with "cell migration", "localization of cell", "cell motility", "cell motion", and other metastasis-related biological processes (Fig. 1b) while there were 210 lncRNAs upregulated and 81 downregulated in N1-stage lung cancer (with the threshold $P<0.01$). The top differentially expressed genes were shown in Tables 3 and 4, and the full list was provided in the supplementary materials.

Survival analysis

Given that all samples included in the GSE50081 dataset were lung cancer tissues, it is unknown whether these genes show a differential expression pattern between lung cancer tissues and normal lung tissues. Thus, we used a list of differentially expressed genes as validation set, which were differentially expressed between lung cancer tissues and corresponding adjacent tissues (these genes were identified by microarrays in our unpublished work, GSE66654). Using the differentially expressed protein-coding genes between T and N stages as two independent training sets, Venny plot revealed that 11 genes were common in the three groups (Table 5 and Supplementary Figure S5). As shown in Table 3, the 11 genes were not overlapped with the gene signature in the original study. To test the potentially prognostic role of these 11 genes, we analyzed whether they were associated with the survival of lung cancer patients. Kaplan-Meier curve and Cox regression were performed for the 11 genes, and results indicated that the expression level of two genes, SCN7A and C7, were significantly associated with the survival of lung cancer patients (Fig. 2). In addition, the predictive efficacy was improved with the combination of C7 and SCN7A (Fig. 2).

Given the specific expression nature of lncRNA, we assessed whether the differentially expressed lncRNAs could be predictive biomarkers of survival or metastasis. First, the top differentially expressed lncRNAs between different T and N stages were validated with TCGA RNA-seq data sets (tumor vs. normal and metastasis tumor vs. primary tumor by the lncRNAtor website). Notably, we found a novel lncRNA; LINC00313 was highly expressed both in lung cancer tissues and metastasized lung cancer tissues (Fig. 3). Additionally, LINC00313 expression was predictive of lung cancer survival, namely lung cancer patients with higher expression level of LINC00313 would have a shorter overall survival (hazard ratio=0.658, Fig. 3).

Table 3 Differentially expressed protein-coding genes between different lymph node stages (N1 vs. N0)

Gene symbol	*P* value	FC	Ensembl transcript ID	Transcript length	Description
PPP2R2C	0.0012214	2.0849315	ENST00000335585	4092	Protein phosphatase 2, regulatory subunit B, gamma
COL11A1	0.0058083	2.0562277	ENST00000370096	7286	Collagen, type XI, alpha 1
PKP1	0.0176348	2.027919	ENST00000367324	5384	Plakophilin 1
MMP1	0.0186381	1.9710987	ENST00000315274	1970	Matrix metallopeptidase 1 (interstitial collagenase)
FGFBP1	0.0017837	1.9480085	ENST00000382333	1359	Fibroblast growth factor binding protein 1
MMP12	0.0135682	1.9145429	ENST00000571244	1874	Matrix metallopeptidase 12 (macrophage elastase)
S100A7	0.0155603	1.9118906	ENST00000368729	4279	S100 calcium binding protein A7A
DSG3	0.0315578	1.8842625	ENST00000257189	5525	Desmoglein 3
CLCA2	0.0238188	1.8416514	ENST00000370565	4025	Chloride channel accessory 2
GJB2	0.0155691	1.7568609	ENST00000382848	2250	Gap junction protein, beta 2, 26 kDa
C4BPA	0.0001547	−2.989698	ENST00000367070	2243	Complement component 4 binding protein, alpha
SCGB3A2	0.001508	−2.928171	ENST00000504320	440	Secretoglobin, family 3A, member 2
SCGB3A1	0.0025359	−2.602684	ENST00000292641	521	Secretoglobin, family 3A, member 1
ADH1B	0.0001915	−2.584706	ENST00000305046	4072	Alcohol dehydrogenase 1B (class I), beta polypeptide
SFTPD	0.000813	−2.584706	ENST00000372292	1281	Surfactant protein D
PIGR	0.0028138	−2.531513	ENST00000356495	4279	Polymeric immunoglobulin receptor
SCGB1A1	0.0142064	−2.411616	ENST00000534397	466	Secretoglobin, family 1A, member 1 (uteroglobin)
SFTPA2	0.0075866	−2.394957	ENST00000372325	2189	Surfactant protein A2
ST6GALNAC1	0.0002184	−2.378414	ENST00000592042	2657	ST6 (alpha-*N*-acetyl-neuraminyl-2,3-beta-galactosyl-1,3)-*N*-acetylgalactosaminide alpha-2,6-sialyltransferase 1
GDF15	7.11E-06	−2.361985	ENST00000595973	890	Growth differentiation factor 15

FC>0, upregulation; FC<0, downregulation

FC fold change

Table 4 Differentially expressed lncRNAs between different lymph node stages (N1 vs. N0)

lncRNA	*P* value	FC	lncRNA length	chr	Strand	Start	End
lnc-CDHR1-1:1	0.00623008	−1.65634	486	chr10	+	85936273	85945040
lnc-GNB2L1-3:1	0.00121403	−1.46307	599	chr5	−	180630117	180630820
lnc-NEURL-3:1	0.00004936	−1.3435	2023	chr10	+	105506536	105515167
lnc-UAP1L1-2:2	0.0011911	−1.31951	625	chr9	+	139948597	139957342
lnc-NES-2:1	0.00058676	−1.31768	3222	chr1	−	156611457	156614679
lnc-GAS1-2:1	0.00485791	−1.30949	2012	chr9	−	89623369	89657041
lnc-KTN1-AS1-1:10	0.00112855	−1.29056	2007	chr14	−	56247852	56263392
lnc-ATMIN-1:1	0.00988649	−1.27368	716	chr16	+	81064374	81065090
lnc-SSTR4-2:6	0.00015019	−1.24833	2345	chr20	+	23168597	23170942
lnc-NES-2:1	0.00233602	−1.2466	3222	chr1	−	156611457	156614679
lnc-CXCL17-3:4	0.00590134	1.555092	2850	chr19	−	43011546	43014396
lnc-SIK1-4:17	0.00516173	1.574616	525	chr21	−	44881991	44884696
lnc-ITGA9-1:1	0.000874	1.582275	1138	chr3	+	37864198	37867773
lnc-SNURF-1:17	0.00106063	1.615522	21686	chr15	+	25241369	25365009
lnc-CALCRL-1:1	0.00314946	1.730273	1810	chr2	−	188328959	188330769
lnc-ADRB1-3:1	0.00057314	1.907919	3040	chr10	+	115805528	115808568
lnc-THAP6-3:1	0.00023809	1.917199	2028	chr4	+	75973295	75975323
lnc-NDNF-1:1	0.0011861	1.975202	1897	chr4	−	121954486	121956383
lnc-FGFBP2-1:1	0.00176126	2.07053	2042	chr4	−	15969852	15973568
lnc-TMC5-2:1	0.00181815	2.234574	273	chr16	+	19421817	19441962

FC>0, upregulation; FC<0, downregulation

FC fold change

Discussion

Due to the wide application of low-dose CT screening, more and more lung cancer patients are diagnosed at early stage. However, there are many debates about the primary treatment options for early-stage lung cancer [4, 27]. Nevertheless, it is paramount important to characterize the altered gene expression profile and identify biomarkers predictive of survival or chemotherapy, which will help understand molecular feature of early-stage lung cancer. To date, researchers have developed several mRNA-based biomarkers by microarray, and several gene signatures have been confirmed effective as prognostic biomarkers [7, 9, 28]. In these studies, the large amount of microarray data offer valuable source for data mining.

In the present study, we utilized a microarray series of 181 lung cancer patients and compared gene expression profiles between different tumor and lymph node stages. Comparing the expression data of patients with and without lymph node metastasis, we found 210 differentially expressed genes, and

Table 5 Genes and original probe sets of the 11 genes for survival analysis

Probe set	Gene symbol	Gene description
206311_s_at	PLA2G1B	Phospholipase A2, group IB (pancreas)
209763_at	CHRDL1	Chordin-like 1
228504_at	SCN7A	Sodium channel, voltage-gated, type VII, alpha subunit
213317_at	CLIC5	Chloride intracellular channel 5
226587_at	PWAR6	Small nuclear ribonucleoprotein polypeptide N
224061_at	INMT	Indolethylamine *N*-methyltransferase
228885_at	MAMDC2	MAM domain containing 2
204712_at	WIF1	WNT inhibitory factor 1
206742_at	FIGF	C-fos induced growth factor (vascular endothelial growth factor D)
202992_at	C7	Complement component 7
223678_s_at	SFTPA2	Surfactant protein A2///surfactant protein A1

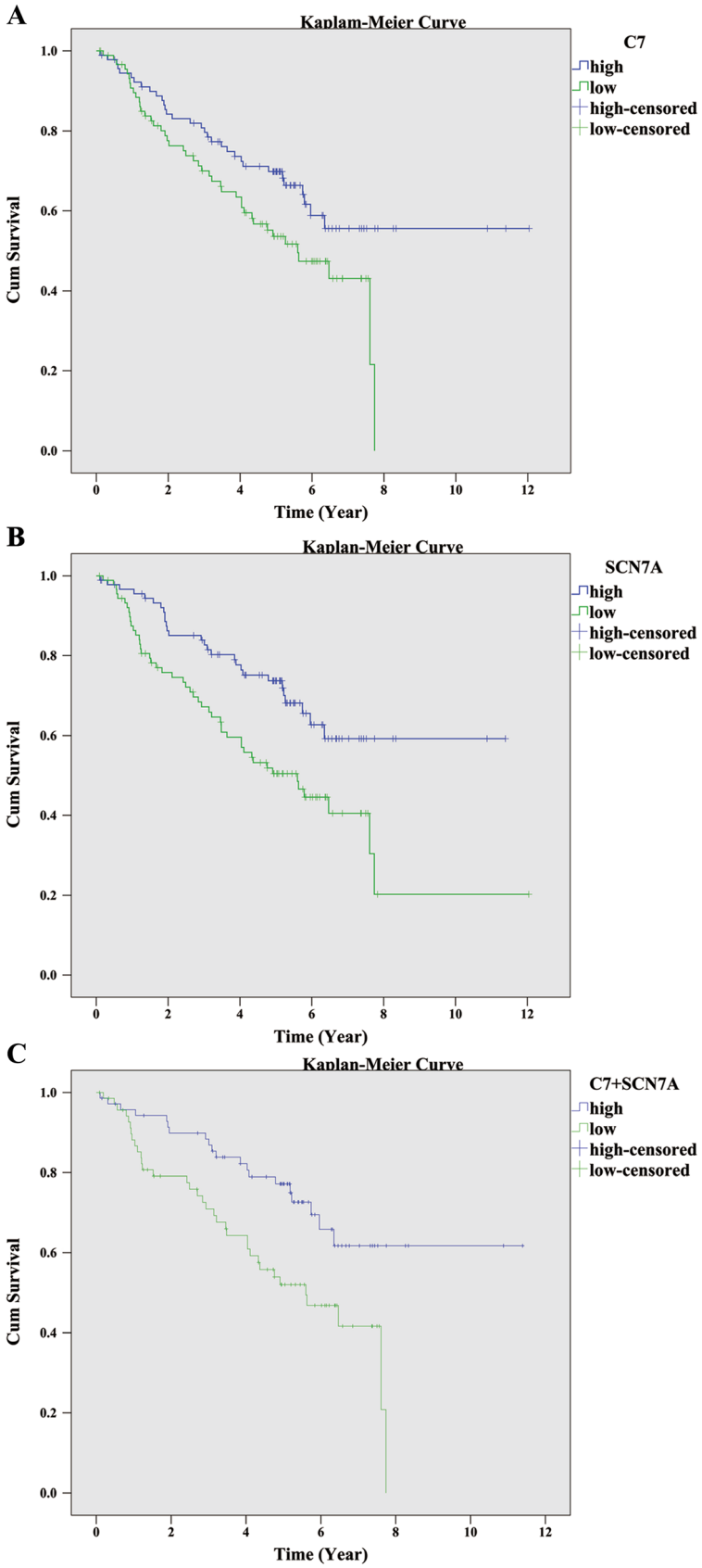

Fig. 2 Survival analysis of differentially expressed protein-coding genes and lncRNA Kaplan-Meier Curves of C7 (**a**), SCN7A (**b**), and C7 + SCN7A (**c**)

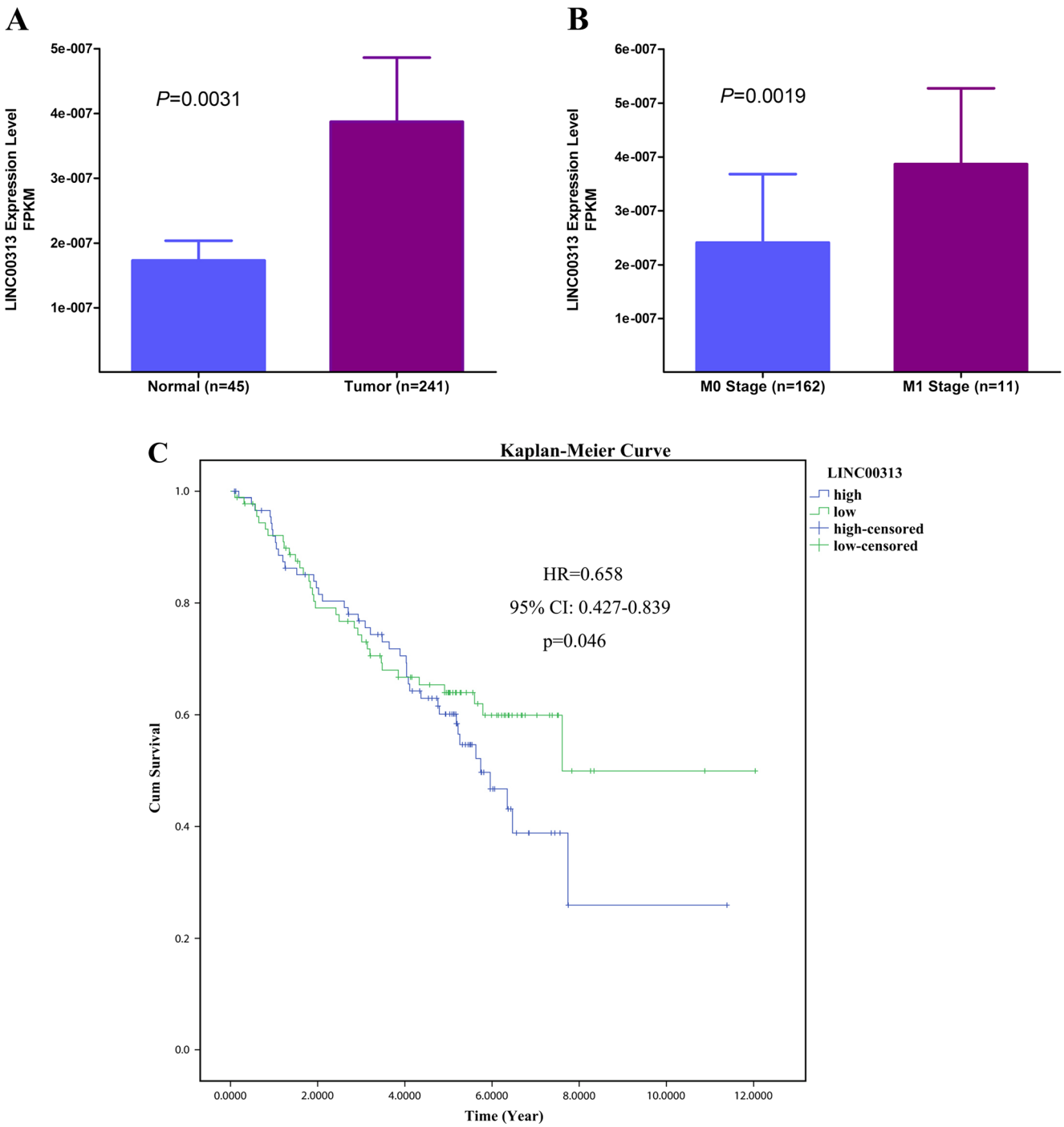

Fig. 3 Expression level of LINC00313. LINC00313 is highly expressed in lung cancer tissues compared with normal tissues (**a**) and metastasized lung cancer tissues compared with primary tumors (**b**). Higher expression level of LINC00313 indicated poor survival of lung cancer (**c**)

functional GO enrichment suggest that the differentially expressed genes were enriched in the biological processes of "cell adhererion cell motily", which were closely associated with invasion and metastasis of cancer. By comparing data of T2- and T1-stage patient, we found 322 differentially expressed genes. Functional annotation analysis revealed that many cancer-related pathways were enriched among the differentially expressed genes, such as "Cell adhesion molecules (CAMs)", "Cell cycle", "p53 signaling pathway", "Pathways in cancer", and "Tight junction". These results suggested that gene expression profiles were different among patients with different tumor and lymph node stages. KEGG pathway analysis and GO analysis are mostly used and powerful data mining tools. By KEGG and GO analyses, we found and revealed the altered pathways in different stages of lung cancer. Our work may help understand the molecular basis of lung cancer.

Since the dataset analyzed included only lung cancer patients, the expression profile of these genes between normal lung tissues and lung cancer tissues is unknown. Thus, we used a gene list of differentially expressed genes between lung cancer and normal lung tissues (the data were from our unpublished work) as a validate gene set. Using Venny plot to select genes which were common in three groups, 11 genes were identified. By cox regression, we found that the two genes (SCN7A and C7) were significantly associated with survival of lung cancer patients. The original data set was designed to validate the prognostic value and predictive efficacy of a 15-gene signature. But, C7 and SCN7A were not included in the 15-gene signature [9]. Additional literature review was performed for C7 and SCN7A, and reports about SCN7A and C7 in the paradigm of cancer research were few. This indicated that the C7 and SCN7A are potential novel prognostic biomarkers of lung cancer, and they may play an important role in lung cancer. However, further studies are warranted to validate our results.

Due to rapid development of high-throughput transcriptome, accumulating evidence suggests that at least 90% of the total mammalian genome is actively transcribed while only less than 2% of the genome sequence is protein-coding genes [29]. And numerous noncoding RNAs are transcribed from genome, of which microRNAs (miRNA) and lncRNA are mostly investigated [30, 31]. It is widely known that lncRNAs play an important role in cancer, such as the process of carcinogenesis, invasion, and metastasis of cancer [13]. Dysregulation of lncRNA has been found in many types of cancer, like breast cancer [32], prostate cancer [33], and lung cancer [34]. Although several genome-wide transcriptome studies have identified a lot of lncRNAs, only a small proportion of lncRNAs has been well characterized. The functional role and molecular mechanism of several cancer-associated lncRNAs have been well characterized. Additionally, it was also found that these cancer-associated lncRNAs could be potential biomarkers, as the dysregulated expression was associated with clinicopathological characteristics, even prognosis.

In current study, we re-annotated the probe set of Human Genome U133 Plus 2.0 microarray using the Lincpedia database. Among the differentially expressed lncRNAs, we noted that a novel lncRNA, LINC00313, which was upregulated both in T2- and N1-stage lung cancer and could be a prognostic biomarker of lung cancer. In addition, expression level of LINC00313 was also analyzed using TCGA RNA sequencing data. In consistence, LINC00313 was highly expressed in lung cancer tissues compared with normal tissues. Intriguingly, compared with primary lung cancer, the expression level of LINC00313 was higher in metastasized lung cancer tissues, which was in accordance with the high expression level in N1 stage. These findings confirmed that LINC00313 could be a potential biomarker for lung cancer while further in vitro studies are warranted to clarify the underlying molecular mechanism. Many functional lncRNAs have been characterized in lung cancer, and several of them were associated with prognosis or other clinical characteristics. By data mining of the dataset, we identified a set of differentially expressed lncRNAs between different stages of lung cancer while further studies are warranted to identify the functional roles and clinical value of these lncRNAs.

To summarize, we performed data mining of a data set of 181 microarrays and found that a set of protein-coding genes and lncRNAs was differentially expressed between different stages. Additionally, SCN7A, C7, and LINC00313 were significantly associated with the survival of lung cancer.

Acknowledgments This study is founded by the Natural Science Foundation of China (81372321 to Lin Xu; 81201830 and 81472200 to Rong Yin), Natural Science Foundation for High Education of Jiangsu Province (13KJB320010 to Rong Yin), Jiangsu Provincial Special Program of Medical Science (BL2012030 to Lin Xu), and Jiangsu Province Ordinary University Graduate Student Research Innovation Project for 2013 (CXLX13_571 to Mantang Qiu).

Conflicts of interest None

References

1. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin. 2011;61(2):69–90.
2. National Lung Screening Trial Research T, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. 2011;365(5):395–409.
3. Senthi S, Lagerwaard FJ, Haasbeek CJ, Slotman BJ, Senan S. Patterns of disease recurrence after stereotactic ablative radiotherapy for early stage non-small-cell lung cancer: a retrospective analysis. Lancet Oncol. 2012;13(8):802–9.
4. Senan S, Paul MA, Lagerwaard FJ. Treatment of early-stage lung cancer detected by screening: surgery or stereotactic ablative radiotherapy? Lancet Oncol. 2013;14(7):e270–4.
5. Chang JY, Senan S, Paul MA, Mehran RJ, Louie AV, Balter P, Groen HJ, McRae SE, Widder J, Feng L, van den Borne BE, Munsell MF, Hurkmans C, Berry DA, van Werkhoven E, Kresl JJ, Dingemans AM, Dawood O, Haasbeek CJ, Carpenter LS, De Jaeger K, Komaki R, Slotman BJ, Smit EF, Roth JA. Stereotactic ablative radiotherapy versus lobectomy for operable stage I non-small-cell lung cancer: a pooled analysis of two randomised trials. Lancet Oncol. 2015.
6. Zhu CQ, Pintilie M, John T, Strumpf D, Shepherd FA, Der SD, et al. Understanding prognostic gene expression signatures in lung cancer. Clin Lung Cancer. 2009;10(5):331–40.
7. Subramanian J, Simon R. Gene expression-based prognostic signatures in lung cancer: ready for clinical use? J Natl Cancer Inst. 2010;102(7):464–74.
8. Zhu CQ, Ding K, Strumpf D, Weir BA, Meyerson M, Pennell N, et al. Prognostic and predictive gene signature for adjuvant chemotherapy in resected non-small-cell lung cancer. J Clin Oncol. 2010;28(29):4417–24.
9. Der SD, Sykes J, Pintilie M, Zhu CQ, Strumpf D, Liu N, et al. Validation of a histology-independent prognostic gene signature for early-stage, non-small-cell lung cancer including stage IA patients. J Thorac Oncol. 2014;9(1):59–64.

10. Sandoval J, Mendez-Gonzalez J, Nadal E, Chen G, Carmona FJ, Sayols S, et al. A prognostic DNA methylation signature for stage I non-small-cell lung cancer. J Clin Oncol. 2013;31(32):4140–7.
11. Ponting CP, Oliver PL, Reik W. Evolution and functions of long noncoding RNAs. Cell. 2009;136(4):629–41.
12. Qiu MT, Hu JW, Yin R, Xu L. Long noncoding RNA: an emerging paradigm of cancer research. Tumour Biol. 2013;34(2):613–20.
13. Tsai MC, Spitale RC, Chang HY. Long intergenic noncoding RNAs: new links in cancer progression. Cancer Res. 2011;71(1): 3–7.
14. Dinger ME, Amaral PP, Mercer TR, Pang KC, Bruce SJ, Gardiner BB, et al. Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. Genome Res. 2008;18(9):1433–45.
15. Ren S, Wang F, Shen J, Sun Y, Xu W, Lu J, et al. Long non-coding RNA metastasis associated in lung adenocarcinoma transcript 1 derived miniRNA as a novel plasma-based biomarker for diagnosing prostate cancer. Eur J Cancer. 2013;49(13):2949–59.
16. Su X, Malouf GG, Chen Y, Zhang J, Yao H, Valero V, et al. Comprehensive analysis of long non-coding RNAs in human breast cancer clinical subtypes. Oncotarget. 2014;5(20):9864–76.
17. Yang J, Lin J, Liu T, Chen T, Pan S, Huang W, et al. Analysis of lncRNA expression profiles in non-small cell lung cancers (NSCLC) and their clinical subtypes. Lung Cancer. 2014;85(2): 110–5.
18. Zhang X, Sun S, Pu JK, Tsang AC, Lee D, Man VO, et al. Long non-coding RNA expression profiles predict clinical phenotypes in glioma. Neurobiol Dis. 2012;48(1):1–8.
19. Wettenhall JM, Smyth GK. limmaGUI: a graphical user interface for linear modeling of microarray data. Bioinformatics. 2004;20(18):3705–6.
20. The Gene Ontology (GO) project in 2006. Nucleic Acids Res. 2006;34(Database issue):D322-326.
21. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–9.
22. Dupuy D, Bertin N, Hidalgo CA, Venkatesan K, Tu D, Lee D, et al. Genome-scale analysis of in vivo spatiotemporal promoter activity in Caenorhabditis elegans. Nat Biotechnol. 2007;25(6):663–8.
23. Schlitt T, Palin K, Rung J, Dietmann S, Lappe M, Ukkonen E, et al. From gene networks to gene function. Genome Res. 2003;13(12): 2568–76.
24. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004;32(Database issue):D277–80.
25. Yi M, Horton JD, Cohen JC, Hobbs HH, Stephens RM. WholePathwayScope: a comprehensive pathway-based analysis tool for high-throughput data. BMC Bioinformatics. 2006;7:30.
26. Draghici S, Khatri P, Tarca AL, Amin K, Done A, Voichita C, et al. A systems biology approach for pathway level analysis. Genome Res. 2007;17(10):1537–45.
27. Louie AV, Senthi S, Palma DA. Surgery versus SABR for NSCLC. Lancet Oncol. 2013;14(12), e491.
28. Tang H, Xiao G, Behrens C, Schiller J, Allen J, Chow CW, et al. A 12-gene set predicts survival benefits from adjuvant chemotherapy in non-small cell lung cancer patients. Clin Cancer Res. 2013;19(6):1577–86.
29. Wilusz JE, Sunwoo H, Spector DL. Long noncoding RNAs: functional surprises from the RNA world. Genes Dev. 2009;23(13): 1494–504.
30. Wang KC, Chang HY. Molecular mechanisms of long noncoding RNAs. Mol Cell. 2011;43(6):904–14.
31. Carthew RW, Sontheimer EJ. Origins and mechanisms of miRNAs and siRNAs. Cell. 2009;136(4):642–55.
32. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature. 2010;464(7291):1071–6.
33. Srikantan V, Zou Z, Petrovics G, Xu L, Augustus M, Davis L, et al. PCGEM1, a prostate-specific gene, is overexpressed in prostate cancer. Proc Natl Acad Sci U S A. 2000;97(22):12216–21.
34. Schmidt LH, Spieker T, Koschmieder S, Schaffers S, Humberg J, Jungen D, et al. The long noncoding MALAT-1 RNA indicates a poor prognosis in non-small cell lung cancer and induces migration and tumor growth. J Thorac Oncol. 2011;6(12):1984–92.