Identification of a four long non-coding RNA (IncRNA) Signature for Predicting Prognosis of Patients with Non-Small Cell Lung Cancer: a Multicenter Study in China

CURRENT STATUS: UNDER REVIEW

Journal of Translational Medicine • BMC

Rui-Qi Wang
State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine

Xiao-Ran Long
State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-Sen University Cancer Center; Department of gynecology and obstetrics, Renji hospital

Chun-Lei Ge
Department of Cancer Biotherapy Center, The Third Affiliated Hospital of Kunming Medical University (Tumor Hospital of Yunnan Province)

Mei-Yin Zhang
State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine

Long Huang
Department of Oncology

Ning-Ning Zhou
Department of Medical Oncology

Yi Hu
State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Department of Thoracic Oncology

Rui-Lei Li
Department of Cancer Biotherapy Center, The Third Affiliated Hospital of Kunming Medical University

Zhen Li
Department of Cancer Biotherapy Center, The Third Affiliated Hospital of Kunming Medical
University (Tumor Hospital of Yunnan Province)

Dong-Ni Chen
State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Department of Thoracic Oncology

Lan-Jun Zhang
State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Department of Thoracic Oncology

Zhe-Sheng Wen
State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Department of Thoracic Oncology

Shi-Juan Mai
State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine

HUI-YUN WANG
State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine

✉ wanghyun@mail.sysu.edu.cn Corresponding Author
ORCiD: https://orcid.org/0000-0001-6728-4376

DOI: 10.21203/rs.3.rs-24087/v1

SUBJECT AREAS
Translational Medicine

KEYWORDS
Non-small cell lung cancer [NSCLC]; signature IncRNA; prognostic biomarker; TNM stage
Abstract

**Background:** This study aims to identify a long non-coding RNA (lncRNA) signature for predicting survival in non-small-cell lung carcinoma (NSCLC) patients and providing additional prognostic information to the tumor node metastasis (TNM) staging system.

**Methods:** NSCLC cases from a hospital were divided into a discovery cohort (n=194) and validation cohort (n=172) and analyzed using a custom lncRNA microarray. Another 73 cases obtained from another hospital were assayed using quantitative reverse transcriptase polymerase chain reaction (qRT-PCR). The differentially expressed lncRNAs were detected by significance analysis of microarrays (SAM) program and used for identifying those associated with survival in the discovery cohort, which were then employed to construct a prognostic lncRNA signature using a risk-score method. The signature was then confirmed in the validation and independent cohort as well.

**Results:** The discovery cohort was found to comprise of 305 lncRNAs, which showed differential expression between the NSCLC and the corresponding normal lung tissues, a 4-lncRNA signature was identified that was found to significantly correlate with the survival of the NSCLC patients. This signature was further validated in the validation and independent cohort. Moreover, multivariate Cox analysis demonstrates that the 4-lncRNA signature is independent of the TNM staging system as a risk-score model. The receiver operating characteristic (ROC) curve indicates that the prognostic value of the combined model is significantly higher than that of TNM staging alone in all the cohorts.

**Conclusions:** This study identified a 4-lncRNA signature, which is a powerful prognostic biomarker which related to patient survival in addition to the traditional TNM staging system.

**Background**

Lung cancer is the most common and lethal malignance in the world, about 85% of which are non-small cell lung cancer (NSCLC).[1] In clinical practice, delayed diagnosis and lack of effective prognostic biomarkers are the main reasons for poor survival of NSCLC cases.[2, 3] While in the late stages of lung cancer, only 15% of patients are known to survive for five years, 83% of patients with stage I can survive for 5 years.[4] Currently, the treatment strategy and prognosis of lung cancer are mainly determined by the TNM staging system. However, NSCLC patients with the same TNM stage
may have different prognosis.[2, 5, 6] Therefore, the development of new biomarkers that can potentially improve the accuracy of prognosis thereby enhancing the quality of life of patients and the survival rate is warranted.[7]

The development and advancement of high-throughput technology has enabled numerous studies to propose a single gene or a gene set (signature) as a biomarker for tumor diagnosis, prognosis, classification, personalized treatment, and so on. Genomic abnormalities such as DNA mutations, copy number variation, DNA methylation, and gene expression have been investigated for identification of prognostic biomarkers in NSCLC patients. Microarray and RNA-seq high-throughput technologies allow us to simultaneously analyze hundreds and thousands of genes and their relationship with clinical features including survival in cancer, leading to a large number of novel biomarkers (single genes or signatures) for diagnosis, prognosis and targeted therapy of NSCLC patients.[8, 9] However, only a few molecular biomarkers (mainly as therapeutic targets) have been applied in clinical practice[10] because most of the biomarkers show low accuracy (including low sensitivity and specificity)[11] or need further confirmation with large sample sizes in the independent validation study.[12] Therefore, biomarkers that are more reliable are still needed for diagnosis, prognosis, and personalized therapy of cancer.

Long non-coding RNAs (lncRNA) that exist in large quantities in the body have exhibited a superior potential as novel diagnostic or prognostic biomarkers as compared to protein-coding genes and raise the possibility of finding more reliable biomarkers for lung cancer.[13, 14] LncRNA is a kind of non-coding RNA larger than 200 nucleotides with no protein-coding capacity.[15, 16] A large number of studies have shown that lncRNA can participate in numerous biological processes, such as epigenetic regulation, cell cycle regulation, and cell differentiation regulation. Growing evidence shows that a large number of lncRNAs are significantly dysregulated in various types of cancers and thereby play important roles in tumorigenesis.[17-19] An increasing number of lncRNAs have been proved to be dysregulated and involved in the tumorigenesis of lung cancer and therefore, can be used as biomarkers for diagnosis and prognosis or targets for therapy. For example, lncRNA MALAT1 and NEAT1 play important roles in lung cancer cell proliferation, cell cycle, and apoptosis as well as in
tumor progression and prognosis.[20-24] Inhibitor targeting MALAT1 has been shown to significantly reduce lung cancer metastasis in a mouse model.[20] The prognostic role of IncRNA signatures in NSCLC has been investigated in many reports using the data downloaded from the Gene Expression Omnibus (GEO) database or The Cancer Genome Atlas (TCGA) database. However, an IncRNA expression profile especially for identifying prognostic signatures in a large cohort of NSCLC patients based on a multicenter study has not been reported yet. Therefore, detailed elucidation of the prognostic value and the clinical application potential of IncRNA signatures in NSCLC patients is warranted.

This study, to the best of our knowledge, is the first multicenter retrospective study assessing the prognosis of 439 NSCLC patients using a custom IncRNA microarray and qRT-PCR. NSCLC patients from South China were randomly divided into discovery cohort (194 cases) and validation cohort (172 cases) and those from Southwest China were used as an independent validation cohort (73 cases). LncRNA expression levels in NSCLC tissues were determined using a custom IncRNA microarray in the discovery and validation cohorts and a 4-IncRNA signature was established to predict overall survival (OS) and disease-free survival (DFS) for NSCLC patients in the discovery cohort. The prognostic value of the novel 4-IncRNA signature was then validated in the validation cohort and further confirmed in the independent validation cohort by qRT-PCR.

**Methods**

**Patients and Clinical information**

A total of 439 samples were collected for this study from the patients who underwent radical resection of lung cancer from the Sun Yat-Sen University Cancer Center (three hundred sixty-six cases) and Yunnan Cancer Hospital (seventy-three cases) between 2003 and 2008. Samples including cancer tissues and corresponding adjacent normal tissues were obtained from each case. The inclusion criteria for the study were: i) all cases confirmed as NSCLC by pathological diagnosis and reviewed by two experienced pathologists; ii) Cases which had not received any form of anti-tumor therapy before surgery; iii) Cases which survived more than a month after surgery; iv) Collected samples preserved at -80 °C immediately after collection. Firstly, the 366 samples collected from the
Sun Yat-Sen University Cancer Center were divided randomly into the discovery cohort (194 cases) and the validation cohort (172 cases). Seventy-three NSCLC cases from Yunnan Cancer Hospital with the same criteria as described above were assigned to an independent cohort. Overall survival (OS) was defined as the time from the date of surgery to the date of death or last follow-up and disease-free time (DFS) was defined as the time from the date of surgery to the date of first recurrence, distant metastasis, death or the last follow-up. The clinic-pathological characteristics of the patients in the three cohorts are listed in (Table 1). This study was reviewed and approved by the Ethical Committees of Sun Yat-Sen University Cancer Center and Yunnan Cancer Hospital. Written informed consent was obtained from each patient.

### Table 1

| Parameters                        | Discovery cohort (N = 194) | Validation cohort (N = 172) | Independent cohort (N = 73) |
|-----------------------------------|---------------------------|-----------------------------|-----------------------------|
| Age (X ± SD)                      | 59.2 ± 10.1               | 59.8 ± 10.2                 | 57.6 ± 9.1                  |
| Gender                            |                           |                             |                             |
| Male                              | 144 (74.2)                | 136 (79.1)                  | 52 (71.2)                   |
| Female                            | 50 (25.8)                 | 36 (20.9)                   | 21 (28.8)                   |
| TNM Stage                         |                           |                             |                             |
| I                                 | 87 (44.8)                 | 74 (43.0)                   | 23 (31.5)                   |
| II                                | 32 (16.5)                 | 34 (19.8)                   | 20 (27.4)                   |
| III                               | 75 (38.7)                 | 64 (37.2)                   | 30 (41.1)                   |
| Histological Type                 |                           |                             |                             |
| ADC                               | 95 (49.0)                 | 89 (51.7)                   | 47 (64.4)                   |
| SCC                               | 88 (45.3)                 | 76 (44.2)                   | 26 (35.6)                   |
| ADC/SCC                           | 11 (5.7)                  | 7 (4.1)                     | 0 (0.0)                     |
| Tumor Size                        |                           |                             |                             |
| < 5 cm                            | 108 (55.7)                | 97 (56.4)                   | 31 (42.5)                   |
| ≥ 5 cm                            | 86 (44.3)                 | 75 (43.6)                   | 42 (57.5)                   |
| Differentiation                   |                           |                             |                             |
| Well/Moderate                     | 126 (64.9)                | 105 (61.0)                  | 43 (58.9)                   |
| Poor                              | 68 (35.1)                 | 67 (39.0)                   | 30 (41.1)                   |
| Lymphatic metastasis              |                           |                             |                             |
| No                                | 105 (54.1)                | 81 (47.1)                   | 45 (61.6)                   |
| Yes                               | 89 (45.9)                 | 91 (52.9)                   | 28 (38.4)                   |
| Follow-up time (month)            |                           |                             |                             |
| Median (IQR)                      | 37 (24–62)                | 36 (23–54)                  | 22 (14–35)                  |
| Smoking History                   |                           |                             |                             |
| No                                | 79 (40.7)                 | 62 (36.0)                   | 33 (45.2)                   |
| Yes                               | 115 (59.3)                | 110 (64.0)                  | 40 (44.8)                   |
| Family Cancer History             |                           |                             |                             |
| No                                | 161 (83.0)                | 151 (87.8)                  | 71 (97.3)                   |
| Yes                               | 33 (17.0)                 | 21 (12.2)                   | 2 (2.7)                     |

**Abbreviation:** Int.: Internal; Ext.: External; ADC, Adenocarcinoma; SCC, Squamous cell carcinoma; IQR, Interquartile range.

### RNA extraction

RNA from tumor and normal lung tissues was extracted using TRIzol reagent (Invitrogen, Carlsbad, CA, USA) and homogenized in a bullet blender (Vortex-Genie 2) according to the manufacturer’s instructions. Briefly, 100 mg of tissues were added into 1 mL of TRIzol reagent, and homogenized in
Bullet Blender at low temperature for 15 min and then incubated at 25°C for 5 min. Chloroform was added to the mixtures and they were violently shaken for 15 s and left undisturbed at room temperature for 10 min followed by centrifugation at 4 °C and 12000 g for 15 min. The supernatant was then transferred into a new tube to which one equal volume of isopropyl alcohol was added and mixed. After standing for 10 min at room temperature, the mixture was centrifuged and the supernatant was discarded. The precipitate was washed with 75% alcohol and then the ethanol was removed by centrifugation. After evaporating the remaining residual ethanol, ddH$_2$O was added to dissolve the RNA. Finally, the concentration and quality of the extracted RNA were measured in a ND-1000 spectrophotometer (NanoDrop Technologies) to meet the requirements of microarray and qRT-PCR experiments.

**Quantitative RT-PCR**

Total RNA (1ug) was reversely transcribed using the GoScript™ Reverse Transcription System (Promega) including oligo-dT and random reverse transcription-primers. Quantitative PCR reactions were performed with GoTaq® qPCR (Promega) using SYBR Green Fluorescent dye method on a PRISM 7900 HT system (Applied Biosystems). Each sample was analyzed in triplicates and reactions without cDNA were included as negative control. The conditions of thermal cycling were as follows: 94 °C for 5 min for hot start, then 40 cycles at 94 °C for 15 s and 60 °C for 30 s. The primer sequences are listed in Table S1 in Additional File 1. The PCR data were normalized by GAPDH expression and then by the median expression value of a given IncRNA in the corresponding samples. The relative quantification of IncRNA expression was presented as $2^{-\Delta\Delta Ct}$.

**LncRNA microarray fabrication and hybridization**

Human IncRNA transcript sequences selected from the public IncRNA databases including LNCipedia, LncRNAdb, LncRNADisease, and EST database were used for designing probes in order to construct the in-house IncRNA microarray and 2,412 probes were successfully designed. The IncRNA microarray was fabricated in house and hybridized as described by previously reports[25]. RNA extracted from the 366 cases of lung cancer and normal lung tissues in the discovery cohort and validation cohort was subjected to IncRNA microarray examination. Briefly, each probe was mixed with printing buffer
to a final concentration of 40 µmol/L and printed in duplicate on the cleaned glass slides (75 × 25 mm). The total RNA (2·0 µg) was labeled with 100 nmol/L of PCp-Cy5 (Jena Bioscience, Germany) in reverse transcription. The mixture of the labeled RNA sample and 1x hybridization solution was then hybridized onto the microarray for 12–18 h at 45 °C. After hybridization, the slides were washed in 1 × saline sodium citrate/1% sodium dodecyl sulphate (1 × SSC/1% SDS) for 10 min at 45 °C, followed by sequential washing in 2 cycles of 0·5 × SSC/0·1% SDS, 2 cycles of 0·2 × SSC and 1 cycle of purified water for 1 min at room temperature and then dried in a special small centrifuge and scanned using the InnoScan 700A Scanner (Innopsys Inc, France).

Microarray Data Procession

The raw microarray data were first subtracted with background and then normalized using the quantile method and log transformation. This log-transformed data was submitted and deposited to GEO database in National Center for Biotechnology Information website (GEO accession number: GSE143018) (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE143018).

The IncRNAs differentially expressed between lung cancer and paired normal tissues were identified using the significance analysis of microarrays (SAM) program with the threshold of fold change > 1·25, P-value < 0·01 and false discovery rate < 0·01 (t test). Hierarchical clustering analysis for classification was applied to the samples of the discovery cohort using average linkage method and uncentered Pearson’s correlation coefficients in MEV 4·2 version.

Statistical analysis

The correlation of the 4-IncRNA prognostic signature with clinical characteristics was assessed by Fisher's exact test and χ² test. All these statistical analyses were done using the SPSS Version 23·0 software. The prognostic accuracy of the 4-IncRNA signature, TNM staging system, and combined risk model were compared with Receiver Operating Characteristics (ROC) curve in MedCalc Version 11·4·2. The OS and DFS of patients were assessed by the Kaplan-Meier method and drawn by GraphPad Prism 8·0.

The impact of IncRNA expression level and clinical characteristics on DFS and OS was determined by the univariate and multivariate Cox regression model. According to the risk-score method reported
previously, (25, 26) 15 lncRNAs were incorporated into different combinations to construct a signature and tested by survival analysis and the lncRNAs were gradually subtracted from the combinations to obtain a final 4-lncRNA signature with the greatest prognostic value.

Results

LncRNA expression profile of NSCLC tissues detected by a custom microarray in a discovery cohort

The 366 NSCLC patients from Sun Yat-Sen University Cancer center in South China were randomly divided into a discovery cohort and a validation cohort. The clinical characteristics of these patients are listed in Table 1. The lncRNA expression profile was first determined in lung tissues of 194 NSCLC patients and in 97 normal tissues in the discovery cohort using a custom lncRNA microarray containing 2412 human lncRNA probes. After subtracting background, normalization and log transformation of microarray data, the lncRNA expression profile was analyzed using the SAM program and student t test and 305 lncRNAs differentially expressed between NSCLC and adjacent normal lung tissues (FDR = 0 and fold change > 1.25) were identified, out of which 138 lncRNAs were unregulated and 167 were down-regulated in the NSCLC tissues.

The potential role of the differentially expressed lncRNAs in distinguishing lung cancer tissues from normal lung tissues was established by performing a hierarchical clustering analysis of the 194 lung cancer samples and the corresponding normal lung tissues with 305 differentially expressed lncRNAs in the discovery cohort. The result showed that a 41-lncRNA signature could distinguish NSCLC tissues from normal tissues with an accuracy of 96.44% (Supplementary Fig S1), and only 10 samples (7 tumor samples and 3 normal samples) were misclassified by the 41-lncRNA signature, implying that these differentially expressed lncRNAs may play an important role in the development and progression of lung cancer.

Identification of a 4-lncRNA prognostic signature for NSCLC patients in the discovery cohort

The elucidation of the prognostic significance of lncRNAs in NSCLC involved univariate Cox regression analysis on all 305 differentially expressed lncRNAs in the discovery cohort. Based on the threshold of the \( P\)-value < 0.05, 15 lncRNAs were found to be significantly associated with OS of the lung cancer patients (Table 2), of which 6 lncRNAs were risky and 9 were protective.
Table 2  
Summary of 15 lncRNAs associated with overall survival of NSCLC patients in the discovery cohort

| NO | LncRNA         | Weight | P value | HR (95%CI)     | Putative Function |
|----|----------------|--------|---------|----------------|------------------|
| 1  | BF768381       | 0.168  | 0.048   | 1.183 (1.001–1.390) | High-risk         |
| 2  | DD3            | 0.212  | 0.035   | 1.236 (1.015–1.500) | High-risk         |
| 3  | BF944729       | 0.228  | 0.045   | 1.255 (1.005–1.560) | High-risk         |
| 4  | SRG1           | 0.439  | 0.006   | 1.552 (1.136–2.120) | High-risk         |
| 5  | NEAT1          | 0.412  | 0.003   | 1.510 (1.154–1.970) | High-risk         |
| 6  | Zeb2NAT        | 0.344  | 0.019   | 1.411 (1.057–1.880) | High-risk         |
| 7  | ASLNC03555     | -0.574 | 0.025   | 0.563 (0.342-0.920) | Protective        |
| 8  | ASLNC09137     | -0.488 | 0.025   | 0.614 (0.401-0.940) | Protective        |
| 9  | GSO_1539211_377| -0.578 | 0.039   | 0.561 (0.324-0.970) | Protective        |
| 10 | GSO_1539832_035| -0.486 | 0.041   | 0.615 (0.386-0.980) | Protective        |
| 11 | Lnc-GAN1       | -0.349 | 0.048   | 0.705 (0.499-0.990) | Protective        |
| 12 | GSO_1539211_480| -0.446 | 0.007   | 0.640 (0.463-0.880) | Protective        |
| 13 | ASLNC11245     | -1.269 | 0.000   | 0.281 (0.143-0.550) | Protective        |
| 14 | BF375442       | -0.348 | 0.026   | 0.706 (0.520-0.950) | Protective        |
| 15 | GSO_1539832_023| -0.503 | 0.010   | 0.605 (0.412-0.880) | Protective        |

The reliability and repeatability of the microarray results were confirmed by evaluating 5 out of the 15 selected prognostic lncRNAs by qRT-PCR in 30 pairs of samples randomly selected from the discovery cohort. Of the 5 lncRNAs, two (NEAT1 and XLOC_009261) were found to be up-regulated and three (XLOC_005302, XLOC_001306, and lnc-GAN1) were found to be down-regulated in the lung cancer tissues as compared to the normal lung tissues. The expression level ratios of the 5 lncRNAs in the cancer tissues to the normal adjacent tissues detected by qRT-PCR were consistent with the results obtained by microarray analysis (Fig. 1a) and significant correlations were found between qRT-PCR and microarray data of the five lncRNAs (Fig. 1b-1f). These results reveal that the lncRNA expression levels detected by lncRNA microarray are reliable and reproducible which can be used for further analysis.

An optimal lncRNA combination (signature) for predicting the survival outcome in NSCLC patients was identified by employing the 15 lncRNAs associated with survival to establish a prognostic signature with a risk-score method as previously reported.[26, 27] Using this method, a 4-lncRNA signature was established with the highest prognostic power, consisting of NEAT1, Inc-GAN1, ASLNC11245, and...
Based on the expression levels of the 4 IncRNAs measured by microarray and weighted by their corresponding regression coefficient derived from univariate Cox regression analysis, the risk score formula is as follow:

\[
\text{Risk score} = (0.412 \times \text{NEAT1 level}) + (-0.349 \times \text{Inc-GAN1 level}) + (-1.269 \times \text{ASLNC11245 level}) + (-0.503 \times \text{GSO_1539832_023 level}).
\]

A risk score was calculated for each patient using the risk-score formula and the scores were divided into high- and low-risk groups according to the median risk score. Kaplan-Meier survival analysis displays that patients with high-risk have remarkable poor OS and DFS than those with low-risk (Fig. 2a), implying that this IncRNA signature could prove to be a highly effective potential prognostic signature for NSCLC patients.

Validation of the 4-IncRNA prognostic signature in NSCLC patients selected from a multicenter registry

The prognostic value of the 4-IncRNA signature identified in the discovery cohort was verified by validating it in NSCLC patients from two different geographical areas, one used as an internal validation cohort and the other as an independent validation cohort. The 4-IncRNA signature was first tested in the validation cohort (172 NSCLC samples) acquired from the same center as the discovery cohort in South China. These NSCLC samples were also detected with the same IncRNA microarray as the discovery cohort and the risk scores were computed for each patient in the validation cohort using the same risk-score formula as used in the discovery cohort. Based on the risk scores, patients were classified into high-risk and low-risk groups. Survival analysis showed that patients with high-risk have much worse OS and DFS than those with low-risk (Fig. 2b), which is consistent with the results obtained in the discovery cohort.

The 4-IncRNAs prognostic signature was then tested in 73 more NSCLC samples (as an independent cohort) obtained from another medical center in Southwest China and the expression of the 4 IncRNAs was detected using qRT-PCR. Univariate Cox regression analysis was then performed on the 4 IncRNAs formulating a risk-score formula using the same method as in the discovery cohort:

\[
\text{Risk score} = (0.297 \times \text{NEAT1 level}) + (-0.259 \times \text{Lnc-GAN1 level}) + (-0.706 \times \text{ASLNC11245 level}) + \]
The risk score for each of the patients in the independent cohort was calculated using the formula.
The median risk score was applied as the cutoff point and patients were categorized into high- and low-risk groups. As shown in Fig. 2c, OS and DFS of NSCLC patients in the high-risk group were found to be significantly worse than those in the low-risk group, which is in concordance with the results obtained in the discovery and validation cohorts. These results demonstrate that the 4-IncRNA signature is significantly correlated with the prognosis of the NSCLC patients from the multicenters in different geographical areas, suggesting that it is a new and powerful prognostic biomarker in NSCLC patients from different areas of China.

The 4-IncRNA prognostic signature is independent of TNM staging system
The clinical significance of the 4-IncRNA signature was established by first conducting a correlation analysis on the clinical characteristics of the signature. The result shows that the 4-IncRNA signature is not correlated with any of the clinical characteristics in the three cohorts (Table 3), implying that the signature is independent of the clinical characteristics. A univariate Cox regression analysis was then carried out on the signature and the clinical characteristics. The results indicate that only the 4-IncRNA signature and the TNM stage are associated with OS (Table 4) and DFS (Table 5) of NSCLC patients in all three cohorts, providing yet evidence that the 4-IncRNA signature is a prognostic factor. Finally, a multivariate Cox regression analysis was performed on the signature and all the clinical characteristics. Various other clinic-pathological variables were considered and both the 4-LncRNA signature and TNM stage were found to be significantly correlated with OS and DFS of patients in all the three cohorts, while other factors were not (Table 6). The independence of the signature as a predictive factor for survival was further confirmed by a stratified analysis on the three different clinical stages with the 4-IncRNA prognostic signature. Based on the risk score of the 4-IncRNA prognostic signature, patients in the same TNM stage (stage I, II, or III) were divided into high- or low-risk subgroups. The results indicated that NSCLC patients with high-risk scores generally had significantly worse OS and DFS than those with low-risk scores (Fig. 3) in stages I, II and III, indicating that the prognostic signature is independent of the TNM staging system. These results, therefore,
indicate that 4-lncRNA molecular signature is a powerful and independent prognostic factor for NSCLC patients.

Table 3
Clinical characteristics of NSCLC patients with high and low signature risk scores:

| Characteristic                  | Discovery cohort (N = 194) | Validation cohort (N = 172) | Independent cohort (N = 73) |
|--------------------------------|---------------------------|----------------------------|----------------------------|
|                                | Low-risk                  | High-risk                  | Low-risk                  | High-risk | P value |
| Age (years)                    | n (%)                     | n (%)                      | n (%)                     | n (%)     |         |
| ≥ 60                           | 52 (53.6)                 | 49 (50.5)                  | 46 (53.5)                 | 46 (53.5) | 0.706   |
| < 60                           | 45 (46.4)                 | 48 (49.5)                  | 40 (46.5)                 | 18 (48.6) | 15 (41.7) |
| Gender                         |                           |                            |                           |           |         |
| Male                           | 82 (84.5)                 | 72 (74.2)                  | 67 (77.9)                 | 69 (80.2) | 0.269   |
| Female                         | 15 (15.5)                 | 25 (25.8)                  | 19 (22.1)                 | 17 (19.8) | 0.287   |
| TNM Stage                      |                           |                            |                           |           |         |
| I                              | 46 (47.4)                 | 41 (42.3)                  | 39 (45.3)                 | 35 (40.7) | 0.637   |
| II                             | 11 (11.3)                 | 21 (21.6)                  | 18 (21.0)                 | 16 (18.6) | 0.269   |
| III                            | 40 (41.2)                 | 35 (36.1)                  | 29 (33.7)                 | 35 (40.7) | 0.493   |
| Histological Type              |                           |                            |                           |           |         |
| ADC                            | 55 (56.7)                 | 40 (41.2)                  | 39 (45.3)                 | 50 (58.1) | 0.297   |
| SCC                            | 39 (40.2)                 | 49 (50.5)                  | 40 (46.5)                 | 36 (41.9) | 0.291   |
| ADC/SCC                        | 3 (3.1)                   | 8 (8.3)                    | 7 (8.2)                   | 0 (0)     | 0.518   |
| Tumor Size                     |                           |                            |                           |           |         |
| < 5 cm                         | 59 (60.8)                 | 49 (50.5)                  | 46 (53.5)                 | 51 (59.3) | 0.332   |
| ≥ 5 cm                         | 38 (39.2)                 | 48 (49.5)                  | 40 (46.5)                 | 35 (40.7) | 0.737   |
| Differentiation                |                           |                            |                           |           |         |
| Well/Moderate                  | 58 (59.8)                 | 68 (70.1)                  | 44 (51.2)                 | 61 (70.9) | 0.203   |
| Poor                           | 39 (40.2)                 | 29 (29.9)                  | 42 (48.8)                 | 25 (29.1) | 0.197   |
| Lymphatic metastasis           |                           |                            |                           |           |         |
| No                             | 47 (48.5)                 | 58 (59.8)                  | 39 (45.3)                 | 42 (48.8) | 0.257   |
| Yes                            | 50 (51.5)                 | 39 (40.2)                  | 47 (54.7)                 | 44 (51.2) | 0.574   |
| Smoking History                |                           |                            |                           |           |         |
| No                             | 36 (37.1)                 | 43 (44.3)                  | 29 (33.7)                 | 33 (38.4) | 0.503   |
| Yes                            | 61 (62.9)                 | 54 (55.7)                  | 57 (66.3)                 | 53 (61.6) | 0.307   |
| Family Cancer History          |                           |                            |                           |           |         |
| No                             | 84 (86.6)                 | 77 (79.3)                  | 73 (75.3)                 | 78 (90.7) | 0.296   |
| Yes                            | 13 (13.4)                 | 20 (20.7)                  | 13 (13.4)                 | 8 (9.3)   | 1 (2.7)  |

13
Table 4
Univariate Cox regression analysis of the impact of the IncRNA signature and other clinicopathological features on OS in three NSCLC patient cohorts:

| Parameters                  | Training cohort | Validation cohort | Independent cohort |
|-----------------------------|-----------------|-------------------|--------------------|
|                             | Hazard Ratio (95% CI) | P Value | Hazard Ratio (95% CI) | P Value | Hazard Ratio (95% CI) | P Value |
| Signature (high vs low)     | 3.20 (0.58 - 1.65) | < 0.001 | 2.84 (1.59 - 5.07) | < 0.001 | 2.84 (1.59 - 5.07) | 0.009 |
| Age (≥ 60 vs < 60)          | 1.24 (0.73 - 2.09) | 0.417 | 1.13 (0.67 - 1.91) | 0.33  | 0.88 (0.36 - 2.13) | 0.782 |
| Gender (male vs female)     | 0.86 (0.48 - 1.52) | 0.619 | 1.24 (0.63 - 2.41) | 0.05  | 0.98 (0.37 - 2.57) | 0.978 |
| TNM Stages (I vs II vs III) | 1.67 (1.28 - 2.19) | < 0.001 | 1.70 (1.30 - 2.23) | 0.001 | 1.74 (1.04 - 2.89) | 0.031 |
| Histological Type (ADC vs SCC) | 1.30 (0.74 - 2.29) | 0.346 | 1.39 (0.82 - 2.36) | 0.589 | 0.53 (0.15 - 1.83) | 0.32  |
| Tumor Size (≥ 5 cm vs < 5 cm) | 1.17 (0.69 - 1.98) | 0.545 | 1.15 (0.68 - 1.95) | 0.017 | 3.00 (1.00 - 9.00) | 0.048 |
| Differentiation (Poor vs Well/Moderate) | 0.98 (0.58 - 1.65) | 0.951 | 1.43 (0.85 - 2.42) | 0.079 | 1.80 (0.74 - 4.32) | 0.188 |
| Lymphatic metastasis (Yes vs No) | 1.44 (0.86 - 2.43) | 0.163 | 0.73 (0.43 - 1.24) | 0.025 | 1.38 (0.46 - 4.14) | 0.561 |
| Smoking History (Yes vs No) | 0.91 (0.54 - 1.54) | 0.736 | 1.63 (0.92 - 2.91) | 0.024 | 1.11 (0.46 - 2.68) | 0.812 |
| Family Cancer History (Yes vs No) | 1.04 (0.52 - 2.06) | 0.899 | 1.18 (0.58 - 2.42) | 0.58  | 0.47 (2.97 - 7.64) | 0.618 |
Table 5
· Univariate Cox regression analysis of the impact of lncRNA signature and other clinicopathological features on DFS in three NSCLC patient cohorts.

| Parameters               | Training group | Validation group | Independent group |
|--------------------------|----------------|------------------|-------------------|
|                          | Hazard Ratio (95% CI) | P Value | Hazard Ratio (95% CI) | P Value | Hazard Ratio (95% CI) | P Value |
| Signature (high vs low)  | 2.61 (1.50–4.56) | <0.001 | 3.21 (1.80–5.71) | <0.001 | 2.18 (1.10–4.34) | 0.025  |
| Age (≥ 60 vs < 60)       | 1.30 (0.76–2.21) | 0.33  | 1.30 (0.76–2.21) | 0.51   | 1.19 (0.60–2.35) | 0.599  |
| Gender (male vs female)  | 0.57 (0.33–1.00) | 0.05  | 0.97 (0.53–1.78) | 0.945  | 0.77 (0.37–1.60) | 0.496  |
| TNM Stages (I vs II vs III) | 1.55 (1.18–2.04) | 0.001 | 1.70 (1.29–2.25) | <0.001 | 1.46 (1.00–2.12) | 0.045  |
| Histological Type (ADC vs SCC) | 0.83 (0.44–1.59) | 0.589 | 1.48 (0.88–2.49) | 0.133  | 1.02 (0.45–2.29) | 0.954  |
| Tumor Size (≥ 5 cm vs < 5 cm) | 1.89 (1.12–3.22) | 0.017 | 1.57 (0.94–2.61) | 0.082  | 1.92 (0.92–4.03) | 0.081  |
| Differentiation (Poor vs Well/Moderate) | 0.62 (0.36–1.05) | 0.079 | 1.22 (0.73–2.04) | 0.427  | 1.29 (0.65–2.57) | 0.453  |
| Lymphatic metastasis (Yes vs No) | 1.82 (1.07–3.10) | 0.025 | 1.72 (1.01–2.91) | 0.042  | 2.13 (0.87–5.23) | 0.095  |
| Smoking History (Yes vs No) | 0.54 (0.32–0.92) | 0.024 | 1.15 (0.68–1.95) | 0.586  | 1.00 (0.51–1.97) | 0.989  |
| Family Cancer History (Yes vs No) | 0.80 (0.38–1.71) | 0.58  | 0.69 (0.31–1.53) | 0.369  | 1.23 (0.16–9.12) | 0.834  |
Table 6
Multivariate Cox regression analysis of the impact of lncRNA signature and clinicopathological features on OS and DFS in three NSCLC patient cohorts.

| Dataset       | Parameters          | Overall Survival |                | Disease-free Survival |                |
|---------------|---------------------|------------------|-----------------|-----------------------|-----------------|
|               |                     | Hazard Ratio     | P Value         | Hazard Ratio          | P Value         |
|               |                     | (95% CI)         |                 | (95% CI)              |                 |
| Training      | Signature           | 3.18 (1.62 – 6.23) | 0.001          | 2.17 (1.35 – 3.47)   | 0.001          |
|               | Age                 | 1.07 (0.78 – 1.45) | 0.401          | 1.03 (0.76 – 1.40)   | 0.845          |
|               | Gender              | 1.61 (0.57 – 4.51) | 0.365          | 1.42 (0.54 – 3.73)   | 0.474          |
|               | TNM Stages          | 1.61 (1.09 – 2.15) | 0.008          | 1.47 (1.06 – 2.05)   | 0.022          |
|               | Histological Types  | 0.99 (0.61 – 1.61) | 0.965          | 0.94 (0.58 – 1.51)   | 0.783          |
|               | Tumor Sizes         | 0.91 (0.47 – 1.73) | 0.767          | 0.86 (0.45 – 1.64)   | 0.657          |
|               | Differentiation     | 1.18 (0.78 – 1.78) | 0.43           | 1.12 (0.75 – 1.68)   | 0.587          |
|               | Pleural Invasion    | 1.60 (0.85 – 3.01) | 0.147          | 1.79 (0.96 – 3.35)   | 0.068          |
|               | Vascular Invasion   | 2.17 (0.75 – 6.28) | 0.154          | 1.86 (0.65 – 5.34)   | 0.248          |
|               | Smoking History     | 2.94 (1.15 – 7.50) | 0.024          | 2.60 (1.09 – 6.24)   | 0.032          |
|               | Family Cancer History | 0.58 (0.25 – 1.37) | 0.218         | 0.53 (0.23 – 1.25)   | 0.146          |
| Validation    | Signature           | 2.41 (1.47 – 3.97) | 0.001          | 2.49 (1.53 – 4.05)   | < 0.001        |
|               | Age                 | 1.14 (0.87 – 1.49) | 0.359          | 1.13 (0.87 – 1.48)   | 0.349          |
|               | Gender              | 0.64 (0.32 – 1.27) | 0.201          | 0.79 (0.41 – 1.55)   | 0.498          |
|               | TNM Stages          | 1.40 (1.03 – 1.91) | 0.031          | 1.40 (1.04 – 1.88)   | 0.026          |
|               | Histological Types  | 0.92 (0.62 – 1.38) | 0.697          | 0.94 (0.63 – 1.4)    | 0.763          |
|               | Tumor Sizes         | 1.41 (0.80 – 2.49) | 0.24           | 1.27 (0.73 – 2.23)   | 0.402          |
|               | Differentiation     | 0.89 (0.59 – 1.32) | 0.552          | 0.88 (0.60 – 1.31)   | 0.537          |
|               | Pleural Invasion    | 1.42 (0.85 – 2.39) | 0.185          | 1.51 (0.90 – 2.52)   | 0.116          |
|               | Vascular Invasion   | 5.40 (1.73 – 16.8) | 0.004          | 4.91 (1.59 – 15.17)  | 0.006          |
|               | Smoking History     | 0.54 (0.28 – 1.04) | 0.064          | 0.52 (0.27 – 1.00)   | 0.05           |
|               | Family Cancer History | 0.92 (0.66 – 2.28) | 0.521         | 0.90 (0.81 – 2.70)   | 0.206          |
| Independent   | Signature           | 1.88 (1.15 – 3.08) | 0.012          | 1.80 (1.14 – 2.84)   | 0.012          |
|               | Age                 | 1.00 (0.70 – 1.42) | 0.988          | 1.19 (0.86 – 1.66)   | 0.294          |
|               | Gender              | 0.80 (0.46 – 1.40) | 0.43           | 0.70 (0.42 – 1.18)   | 0.183          |
|               | TNM Stages          | 1.80 (1.28 – 2.54) | 0.001          | 1.69 (1.24 – 2.30)   | 0.001          |
|               | Histological Types  | 1.26 (0.74 – 2.13) | 0.395          | 1.12 (0.69 – 1.84)   | 0.64           |
|               | Tumor Sizes         | 1.78 (1.06 – 2.97) | 0.028          | 1.92 (1.19 – 3.09)   | 0.008          |
|               | Differentiation     | 1.66 (1.03 – 2.66) | 0.037          | 1.81 (1.16 – 2.83)   | 0.009          |
|               | Pleural Invasion    | 1.26 (0.68 – 2.33) | 0.466          | 1.59 (0.87 – 2.91)   | 0.13           |
|               | Vascular Invasion   | 1.75 (0.39 – 7.92) | 0.468          | 2.81 (0.79 – 9.98)   | 0.11           |

The 4-lncRNA signature provides additional prognostic information to the TNM staging system in NSCLC patients.
In clinical practice, the traditional TNM staging system is the main approach for predicting the survival of patients with NSCLC and determining the treatment strategy. However, TNM staging system is mainly based on anatomic information and does not include the tumor biology factors. Therefore, this system is insufficient to predict survival outcome in NSCLC patients.[28] For example, Kaplan-Meier survival analysis on the three cohorts in this study showed that TNM stage system cannot effectively predict the prognosis of NSCLC patients in different stages, especially in stage I and II (Fig. 4). In order to improve the survival prediction of the TNM staging system, a new risk score model was established by combining the risk scores of the signature and the TNM staging systems. The low and high-risk cases were scored as 0 and 1, respectively while the stage I, II, and III were scored as 1, 2, and 3, respectively. Patients with the combined score of 1, 2–3, and 4 were classified as low-, medium- and high-risk, respectively. The Kaplan-Meier survival analysis was then performed on the patients with different combined risk scores in the three cohorts. The results showed that there was a significant difference in OS and DFS between patients with low-, medium-, and high-risk scores in the discovery cohort (Fig. 5a) and these results were confirmed in the validation and independent cohorts (Fig. 5b-5c).

The ROC analysis was then performed to compare the accuracy of the TNM staging system and the combined risk model. In the ROC curve analysis, the combined risk model achieved a significantly higher predictive accuracy for OS (AUC = 0.726 vs 0.644) and DFS (AUC = 0.723 vs 0.641) than the TNM staging system in the discovery cohort (Fig. 6a), and the same results were observed in the validation and the independent cohorts, respectively (Fig. 6b-6c). All these results proved that the 4-lncRNA signature could provide additional prognostic information and enhance the prognostic power of the TNM staging system.

Discussion
LncRNAs, a novel class of non-coding RNA, have been widely observed to be dysregulated and involved in a diverse range of biological functions in a variety of cancers. Numerous aberrant lncRNAs were detected as hallmarks in cancers and have the potential for diagnosis, prognosis and targeted therapy in cancer. Some studies have described lncRNA profiles and identified lncRNA signatures in
NSCLC patients by data mining from the Gene Expression Omnibus (GEO) and Cancer Genome Atlas (TCGA). For example, Meng Zhou et al.[29] obtained the lncRNA expression profiles of 603 patients derived from three independent NSCLC cohorts in GEO database and developed a risk score model based on the expression of 8 lncRNAs, which was significantly associated with overall survival in NSCLC patients. Ting Lin et al.[9] identified a seven-lncRNA signature to predict OS of NSCLC patients using the combination of four GEO datasets and validated the signature in two independent datasets (TCGA and GSE31210). A recent study by He R et al.[30] proposed a novel 8-gene signature for prognostic prediction in early-stage NSCLC patients by analyzing data from GEO and TCGA projects. However, these prognostic signatures from data mining have neither been confirmed in actual clinical NSCLC cases nor in prospective multicenter studies. Therefore, the clinical application of prognostic lncRNA biomarkers in NSCLC is still very limited. This study, to the best of our knowledge is the first to report lncRNA expression profiling by microarray in a large cohort of NSCLC patients and established the identification of an effective prognostic 4-lncRNA signature.

In this study, 305 aberrantly expressed lncRNAs were identified in 194 NSCLC tissues compared with corresponding normal tissues in the discovery cohort using a custom lncRNA microarray containing 2520 probes. Hierarchical clustering using a set of 41 differentially expressed lncRNAs clearly separated NSCLC tissues from normal tissues with an accuracy of 96.44% in the discovery cohort, highlighting the important roles of the abnormal lncRNAs in lung cancer carcinogenesis. Notably, a novel 4-lncRNA prognostic signature was identified for NSCLC patients in the discovery cohort. Kaplan-Meier survival analysis demonstrated the significant prognostic performance of the signature in all the three cohorts. Multivariate Cox regression analysis identified the 4-lncRNA signature as an independent prognostic factor for NSCLC patients in all cohorts.

Although the TNM staging system is widely accepted for the prediction of prognosis and to guide treatment decisions for most solid cancers including NSCLC at present, there are critical limitations and insufficiencies in the application of the TNM staging system in clinical practice due to the intra-tumor molecular and genetic heterogeneity among patients with lung cancer. The clinical outcomes of lung cancer patients with similar clinical and pathological features are often quite different after
receiving similar treatment. Therefore, more personalized molecular markers are urgently needed to ensure better prognosis and treatment outcomes in clinical practice. In the stratified analysis, the 4-lncRNA signature showed prognostic value in patients within the same stage. Moreover, a risk score model combining the 4-lncRNA signature and the TNM staging was developed and was demonstrated to have a superior power in predicting OS and DFS in all three cohorts as compared to the TNM staging system using Kaplan-Meier survival analysis and ROC analysis. These findings demonstrated that the 4-lncRNA signature can significantly improve the prognostic accuracy of TNM staging and could be considered as a marker for risk assessment in NSCLC patients. Combination of the 4-lncRNA signature with the traditional TNM staging parameters might be a powerful predictor of prognosis in NSCLC patients with the potential to facilitate selection of more aggressive patients who would benefit from adjuvant therapy.

Among the four lncRNAs consisting of the signature, only NEAT1 has been reported to be linked with cancer. NEAT1 is aberrantly expressed in many human malignancies including lung cancer and functions as an oncogene. Higher NEAT1 expression is correlated with advanced TNM stages and lymphatic metastasis in NSCLC patients.[31] Previous studies have revealed that NEAT1 promotes epithelial mesenchymal transition (EMT) and metastasis in NSCLC via the Wnt/b-catenin pathway.[24, 32] However, the association of NEAT1 with the survival of lung cancer patients has not been reported until now. Consistent with published reports, the results obtained in this study found NEAT1 expression to be significantly higher in NSCLC tissues compared with adjacent normal tissues (Fold change = 1.7). Moreover, this study reported for the first time that NEAT1 is an independent prognostic predictor for NSCLC patients (data not published). There is no available functional annotation for the remaining three lncRNAs (Inc-GAN1, ASLNC11245, and GSO_1539832_023) included in the prognostic signature, to the best of our knowledge. In the present study, these three lncRNAs were significantly down-regulated in lung cancer tissues compared with adjacent normal tissues (Fold change = 0.189, 0.749, and 0.785, respectively) and these higher levels could serve as indicators for good prognosis in patients with NSCLC.

Current treatment strategies for lung cancer include a comprehensive treatment plan including
surgery, radiotherapy, chemotherapy, targeted therapy, gene therapy and immunotherapy.\[33, 34\] With advances in molecular knowledge in the past 10 years, commonly mutated genes such as EGFR-TKIs (EGFR tyrosine kinase inhibitors), Programmed cell death protein 1 (PD-1) and Epidermal growth factor receptor (EGFR) super-family have been regarded as therapeutic targets in NSCLC[8, 35]. Despite improved survival and the quality of life in NSCLC patients due to these therapies, the effect is far from satisfactory in many patients. Most of the patients experience drug-resistance or disease progression after receiving treatment for a certain period.\[36, 37\] Therefore, specific biomarkers to monitor therapeutic response in NSCLC are warranted. With the application of micro-array and RNA-seq technology in cancer research, numerous molecular biomarkers have been identified with the ability to predict response to a specific treatment regimen.\[38\] Of the 4-lncRNA signature identified in this study, NEAT1 has been reported to be up-regulated significantly in Paclitaxel-resistant NSCLC cells and is known to contribute to Paclitaxel-resistance through activation of Akt/mTOR signaling pathway. A recent study showed that NEAT1 inhibits apoptosis in multiple myeloma by regulating genes involved in DNA repair processes including the homologous recombination pathway, suggesting its association with drug resistance.\[39\] Therefore, NEAT1, one of the components of the 4-lncRNA signature plays an important role in NSCLC.

Although the 4-lncRNA prognostic signature is a novel and potentially powerful predictor for survival in NSCLC patients, further prospective validation studies in larger cohorts and clinical trials are warranted. There are also other limitations of the present study. Firstly, although the 4-lncRNA signature was identified in a large number of NSCLC samples from two different regions of China, this signature still needs to be validated in a prospective multicenter study, in which patients should be from more institutions belonging to different countries. Secondly, the efficacy of multi-markers-based models was regarded to provide a better prognostic value than a single marker. Thus, further studies are required to identify a multi-gene panel integrating lncRNA, miRNA, and mRNA, with the aim to obtain a more accurate prognostic assessment of NSCLC. Lastly, further experiments are required to elucidate the characteristics and functions of the identified prognostic lncRNAs.

Conclusions
The findings in this study revealed tumor specific lncRNA expression profile in NSCLC tissues and identified a novel prognostic signature based on 4 lncRNAs, which was proved a powerful and independent predictor for OS and DFS of NSCLC patients. Moreover, a prognostic model combining the 4-lncRNA signature and the TNM stage was developed to refine the current staging system and improve the prediction power. This study suggests that the 4-lncRNA classifier system might be a potential predictive biomarker with high precision for selection of high-risk patients who might benefit from adjuvant therapy and thus can guide personalized management of NSCLC patients.

Declarations

Acknowledgements

None.

Author`s contributions

Rui-Qi Wang: Searched Literature, drew figures, analyzed and interpreted data, wrote manuscript; Xiao-Ran Long: designed study, searched literature, performed experiments, analyzed and interpreted data; Chun-Lei Ge: collected samples, interpreted data; Mei-Yin Zhang: performed experiment, collected and analyzed data; Long Huang: designed the microarray and analyzed data; Ning-Ning Zhou: collected samples and interpreted data; Yi Hu: collected samples and analyzed data; Rui-Lei Li: collected samples and interpreted data; Zhen Li: collected clinical information and analyzed data; Dong-Ni Chen: collected samples and performed experiments; Lan-Jun Zhang: collected samples and interpreted data; Zhe-Sheng Weng: collected clinical information and interpreted data; Shi-Juan Mai1: analyzed and interpreted data and wrote manuscript; Hui-Yun Wang: designed study, analyzed and interpreted data, supervised experiments, and wrote manuscript.

Funding

This study was supported by National Natural Science Foundation of China (Grant No: 81772991, 81572466 and 81372564 to HYW) and National Natural Science Foundation of China (Grant No.81772884 to MSJ).

Availability of data and materials

All data in our study are available upon request.
Ethics approval and consent to participate

The study was approved by the Research Ethics Committee of Sun Yat-Sen University Cancer Center. Research was conducted according to all ethical standards, and written informed consent was obtained from all patients.

Consent for publication

Consent to publish has been obtained from all authors.

Competing interests

The authors declare that they have no competing interests.

References

1. Siegel RL, Miller KD, Jemal A: Cancer statistics, 2020. CA Cancer J Clin 2020, 70(1):7-30.
2. Rusch VW, Chansky K, Kindler HL, Nowak AK, Pass HI, Rice DC, Shemanski L, Galateau-Salle F, McCaughan BC, Nakano T et al: The IASLC Mesothelioma Staging Project: Proposals for the M Descriptors and for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Mesothelioma. J Thorac Oncol 2016, 11(12):2112-2119.
3. Herbst RS, Morgensztern D, Boshoff C: The biology and management of non-small cell lung cancer. Nature 2018, 553(7689):446-454.
4. Allemani C, Matsuda T, Di Carlo V, Harewood R, Matz M, Niksic M, Bonaventure A, Valkov M, Johnson CJ, Esteve J et al: Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. Lancet 2018, 391(10125):1023-1075.
5. Carter BW, Lichtenberger JP, 3rd, Benveniste MK, de Groot PM, Wu CC, Erasmus JJ, Truong MT: Revisions to the TNM Staging of Lung Cancer: Rationale, Significance, and Clinical Application. Radiographics 2018, 38(2):374-391.
6. Reck M, Rabe KF: Precision Diagnosis and Treatment for Advanced Non-Small-Cell Lung Cancer. *N Engl J Med* 2017, **377**(9):849-861.

7. Wei MM, Zhou GB: Long Non-coding RNAs and Their Roles in Non-small-cell Lung Cancer. *Genomics Proteomics Bioinformatics* 2016, **14**(5):280-288.

8. Shukla S, Evans JR, Malik R, Feng FY, Dhanasekaran SM, Cao X, Chen G, Beer DG, Jiang H, Chinnaiyan AM: Development of a RNA-Seq Based Prognostic Signature in Lung Adenocarcinoma. *J Natl Cancer Inst* 2017, **109**(1).

9. Lin T, Fu Y, Zhang X, Gu J, Ma X, Miao R, Xiang X, Niu W, Qu K, Liu C et al: A seven-long noncoding RNA signature predicts overall survival for patients with early stage non-small cell lung cancer. *Aging (Albany NY)* 2018, **10**(9):2356-2366.

10. Hiley CT, Le Quesne J, Santis G, Sharpe R, de Castro DG, Middleton G, Swanton C: Challenges in molecular testing in non-small-cell lung cancer patients with advanced disease. *Lancet* 2016, **388**(10048):1002-1011.

11. Lochowska BA, Nowak D, Bialasiewicz P: Cell-free tumour DNA as a diagnostic and prognostic biomarker in non-small cell lung carcinoma. *Adv Respir Med* 2019, **87**(2):118-122.

12. Chen HY, Yu SL, Chen CH, Chang GC, Chen CY, Yuan A, Cheng CL, Wang CH, Terng HJ, Kao SF et al: A five-gene signature and clinical outcome in non-small-cell lung cancer. *N Engl J Med* 2007, **356**(1):11-20.

13. Yu H, Xu Q, Liu F, Ye X, Wang J, Meng X: Identification and validation of long noncoding RNA biomarkers in human non-small-cell lung carcinomas. *J Thorac Oncol* 2015, **10**(4):645-654.

14. Spizzo R, Almeida MI, Colombatti A, Calin GA: Long non-coding RNAs and cancer: a new frontier of translational research? *Oncogene* 2012, **31**(43):4577-4587.
15. Khurana E, Fu Y, Chakravarty D, Demichelis F, Rubin MA, Gerstein M: **Role of non-coding sequence variants in cancer.** *Nat Rev Genet* 2016, **17**(2):93-108.

16. Mercer TR, Dinger ME, Mattick JS: **Long non-coding RNAs: insights into functions.** *Nat Rev Genet* 2009, **10**(3):155-159.

17. Kitagawa M, Kitagawa K, Kotake Y, Niida H, Ohhata T: **Cell cycle regulation by long non-coding RNAs.** *Cell Mol Life Sci* 2013, **70**(24):4785-4794.

18. Lee JT: **Epigenetic regulation by long noncoding RNAs.** *Science* 2012, **338**(6113):1435-1439.

19. Schmitt AM, Chang HY: **Long Noncoding RNAs in Cancer Pathways.** *Cancer Cell* 2016, **29**(4):452-463.

20. Gutschner T, Hammerle M, Eissmann M, Hsu J, Kim Y, Hung G, Revenko A, Arun G, Stentrup M, Gross M et al: **The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells.** *Cancer Res* 2013, **73**(3):1180-1189.

21. Schmidt LH, Spieker T, Koschmieder S, Schaffers S, Humberg J, Jungen D, Bulk E, Hascher A, Wittmer D, Marra A 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):1984-1992.

22. Schmidt LH, Gorlich D, Spieker T, Rohde C, Schuler M, Mohr M, Humberg J, Sauer T, Thoenissen NH, Huge A et al: **Prognostic impact of Bcl-2 depends on tumor histology and expression of MALAT-1 IncRNA in non-small-cell lung cancer.** *J Thorac Oncol* 2014, **9**(9):1294-1304.

23. Li S, Yang J, Xia Y, Fan Q, Yang KP: **Long Noncoding RNA NEAT1 Promotes Proliferation and Invasion via Targeting miR-181a-5p in Non-Small Cell Lung Cancer.** *Oncol Res* 2018, **26**(2):289-296.
24. Qi L, Liu F, Zhang F, Zhang S, Lv L, Bi Y, Yu Y: **IncRNA NEAT1 competes against let-7a to contribute to non-small cell lung cancer proliferation and metastasis.** *Biomed Pharmacother* 2018, **103**:1507-1515.

25. Weng NQ, Chi J, Wen J, Mai SJ, Zhang MY, Huang L, Liu J, Yang XZ, Xu GL, Fu JH et al: **The prognostic value of a seven-IncRNA signature in patients with esophageal squamous cell carcinoma: a IncRNA expression analysis.** *J Transl Med* 2020, **18**(1):47.

26. Yu SL, Chen HY, Chang GC, Chen CY, Chen HW, Singh S, Cheng CL, Yu CJ, Lee YC, Chen HS et al: **MicroRNA signature predicts survival and relapse in lung cancer.** *Cancer Cell* 2008, **13**(1):48-57.

27. Liu N, Chen NY, Cui RX, Li WF, Li Y, Wei RR, Zhang MY, Sun Y, Huang BJ, Chen M et al: **Prognostic value of a microRNA signature in nasopharyngeal carcinoma: a microRNA expression analysis.** *Lancet Oncol* 2012, **13**(6):633-641.

28. Hu Z, Chen X, Zhao Y, Tian T, Jin G, Shu Y, Chen Y, Xu L, Zen K, Zhang C et al: **Serum microRNA signatures identified in a genome-wide serum microRNA expression profiling predict survival of non-small-cell lung cancer.** *J Clin Oncol* 2010, **28**(10):1721-1726.

29. Zhou M, Guo M, He D, Wang X, Cui Y, Yang H, Hao D, Sun J: **A potential signature of eight long non-coding RNAs predicts survival in patients with non-small cell lung cancer.** *J Transl Med* 2015, **13**:231.

30. He R, Zuo S: **A Robust 8-Gene Prognostic Signature for Early-Stage Non-small Cell Lung Cancer.** *Front Oncol* 2019, **9**:693.

31. Pan LJ, Zhong TF, Tang RX, Li P, Dang YW, Huang SN, Chen G: **Upregulation and clinicopathological significance of long non-coding NEAT1 RNA in NSCLC tissues.** *Asian Pac J Cancer Prev* 2015, **16**(7):2851-2855.
32. Kong X, Zhao Y, Li X, Tao Z, Hou M, Ma H: Overexpression of HIF-2alpha-Dependent NEAT1 Promotes the Progression of Non-Small Cell Lung Cancer through miR-101-3p/SOX9/Wnt/beta-Catenin Signal Pathway. Cell Physiol Biochem 2019, 52(3):368-381.

33. Kris MG, Gaspar LE, Chaft JE, Kennedy EB, Azzoli CG, Ellis PM, Lin SH, Pass HI, Seth R, Shepherd FA et al: Adjuvant Systemic Therapy and Adjuvant Radiation Therapy for Stage I to IIIA Completely Resected Non-Small-Cell Lung Cancers: American Society of Clinical Oncology/Cancer Care Ontario Clinical Practice Guideline Update. J Clin Oncol 2017, 35(25):2960-2974.

34. Camidge DR, Doebele RC, Kerr KM: Comparing and contrasting predictive biomarkers for immunotherapy and targeted therapy of NSCLC. Nat Rev Clin Oncol 2019, 16(6):341-355.

35. Wu YL, Cheng Y, Zhou X, Lee KH, Nakagawa K, Niho S, Tsuji F, Linke R, Rosell R, Corral J et al: Dacomitinib versus gefitinib as first-line treatment for patients with EGFR-mutation-positive non-small-cell lung cancer (ARCHER 1050): a randomised, open-label, phase 3 trial. Lancet Oncol 2017, 18(11):1454-1466.

36. Hensing TA, Schell MJ, Lee JH, Socinski MA: Factors associated with the likelihood of receiving second line therapy for advanced non-small cell lung cancer. Lung Cancer 2005, 47(2):253-259.

37. Zhang L, Li S, Choi YL, Lee J, Gong Z, Liu X, Pei Y, Jiang A, Ye M, Mao M et al: Systematic identification of cancer-related long noncoding RNAs and aberrant alternative splicing of quintuple-negative lung adenocarcinoma through RNA-Seq. Lung Cancer 2017, 109:21-27.

38. Best MG, Sol N, Kooi I, Tannous J, Westerman BA, Rustenburg F, Schellen P, Verschueren H, Post E, Koster J et al: RNA-Seq of Tumor-Educated Platelets
Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics. *Cancer Cell* 2015, **28**(5):666-676.

39. Li B, Gu W, Zhu X: **NEAT1 mediates paclitaxel-resistance of non-small cell of lung cancer through activation of Akt/mTOR signalling pathway.** *J Drug Target* 2019, **27**(10):1061-1067.

Figures
Comparison of microarray data with qRT-PCR data. The microarray data are reliable and reproducible, five lncRNAs were measured by real-time quantitative RT-PCR in the 30 pairs of lung cancer and corresponding normal tissues. (a) The expression levels of 5 lncRNAs detected by microarray were consistent with those measured by qRT-PCR. (b-f) There is significant correlation between the expression of five lncRNAs detected by real-time quantitative PCR and those by microarray (Pearson correlation, P<0.001).

**Figure 1**

**a. Discovery Cohort**

**Overall survival**

|                | Low Risk | High Risk |
|----------------|----------|-----------|
| Number at Risk | 97       | 96        |
|                | 79       | 61        |
|                | 41       | 33        |
|                | 6        | 8         |
|                | 1        | 1         |
|                | 0        | 0         |

**Disease-free survival**

|                | Low Risk | High Risk |
|----------------|----------|-----------|
| Number at Risk | 97       | 96        |
|                | 77       | 52        |
|                | 38       | 27        |
|                | 11       | 7         |
|                | 2        | 2         |
|                | 0        | 0         |

HR 3.563 (95% CI 2.118-5.995)  
P<0.001

**b. Validation Cohort**

**Overall survival**

|                | Low Risk | High Risk |
|----------------|----------|-----------|
| Number at Risk | 86       | 86        |
|                | 70       | 61        |
|                | 40       | 31        |
|                | 8        | 6         |
|                | 2        | 1         |
|                | 1        | 0         |

**Disease-free survival**

|                | Low Risk | High Risk |
|----------------|----------|-----------|
| Number at Risk | 86       | 86        |
|                | 70       | 60        |
|                | 38       | 30        |
|                | 7        | 4         |
|                | 1        | 1         |
|                | 1        | 0         |

HR 3.236 (95% CI 2.116-4.547)  
P<0.001

**c. Independent Cohort**

**Overall survival**

|                | Low Risk | High Risk |
|----------------|----------|-----------|
| Number at Risk | 86       | 86        |
|                | 70       | 61        |
|                | 40       | 31        |
|                | 8        | 6         |
|                | 2        | 1         |
|                | 1        | 0         |

**Disease-free survival**

|                | Low Risk | High Risk |
|----------------|----------|-----------|
| Number at Risk | 86       | 86        |
|                | 70       | 60        |
|                | 38       | 30        |
|                | 7        | 4         |
|                | 1        | 1         |
|                | 1        | 0         |

HR 2.967 (95% CI 1.949-4.517)  
P<0.001
The 4-lncRNA signature is a predictor for OS and DFS of patients in three cohorts NSCLC patients were divided into high- and low-risk groups based on the 4-lncRNA signature risk, and analyzed with Kaplan-Meier survival curves. Patients with high-risk have significantly worse OS (left panel) and DFS (right panel) in (a) the discovery cohort (n=194), (b) validation cohort (n=172) and (c) independent cohort (n=73).
The 4-lncRNA signature predicts different survivals in NSCLC patients with the same stage.

Based on the 4-lncRNA signature risk score, NSCLC patients in the same stage were classified to high- or low-risk group. Kaplan-Meier survival analysis was used to estimate patients’ survival in the discovery cohort. Patients with high signature risk have significantly poorer OS (left panel) and DFS (right panel) than those with low signature risk in (a) stage I (n=87), (b) stage II (n=32) and (c) stage III (n=84).

**a. Discovery Cohort**

**b. Validation Cohort**
TNM staging system does not predict survival well in three NSCLC cohorts. TNM staging system is the main tool for predicting survival and determining the treatment strategy, but does not predict survival well in NSCLC patients. The Kaplan-Meier survival curves for OS and DFS of patients with stage I, II, and III in (a) the discovery cohort (n=194), (b) validation cohort (n=172), and (c) independent cohort (n=73).
The prognostic value of the combination of 4-lncRNA signature and TNM staging in the three NSCLC cohorts. In order to improve the TNM staging system, the 4-lncRNA signature is combined with it to construct a new risk model for predicting survival in NSCLC patients. Based on the new risk scores, patients were categorized into low-, medium-, and high-risk groups. Then Kaplan-Meier survival analysis was used to compare OS and DFS of patients with low-, medium-, and high-risk in (a) the discovery cohort, (b) validation cohort, and (c) independent cohort.

**Fig. 5**

**Figure 5**

The prognostic value of the combination of 4-lncRNA signature and TNM staging in the three NSCLC cohorts. In order to improve the TNM staging system, the 4-lncRNA signature is combined with it to construct a new risk model for predicting survival in NSCLC patients. Based on the new risk scores, patients were categorized into low-, medium-, and high-risk groups. Then Kaplan-Meier survival analysis was used to compare OS and DFS of patients with low-, medium-, and high-risk in (a) the discovery cohort, (b) validation cohort, and (c) independent cohort.
Figure 6

The combination model is significantly better than TNM staging in predicting survivals of NSCLC patients. ROC analysis was employed to compare the predictive accuracy of three survival predictors including the 4-IncRNA signature, TNM stage and the combination model. Comparison of the three survival predictors in predicting OS (left panel) and DFS (right panel) in (a) the discovery cohort, (b) validation cohort and (c) independent cohort.
Supplementary Files
This is a list of supplementary files associated with this preprint. Click to download.

4lncRNANSCLCSupplementaryMaterial.pdf