

Award Number : W81XWH-17-1-0254

TITLE: Sputum Biomarkers to Improve CT Screening for the Early Detection of Lung Cancer in Veterans

PRINCIPAL INVESTIGATOR: Feng Jiang

CONTRACTING ORGANIZATION: University of Maryland, Baltimore
Baltimore, MD 21201-1116

REPORT DATE: Oct 2020

TYPE OF REPORT: Final Report

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE Oct 2020		2. REPORT TYPE Final		3. DATES COVERED 07/01/2017 - 06/30/2020	
4. TITLE AND SUBTITLE Sputum Biomarkers to Improve CT Screening for the Early Detection of Lung Cancer in Veterans				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER W81XWH-17-1-0254	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Feng Jiang E-Mail: fjiang@som.umaryland.edu				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) University of Maryland, Baltimore Department of Pathology (Research) Medical School Teaching Facility, 734C 10 South Pine Street Baltimore, MD 21201-1116				8. PERFORMING ORGANIZATION REPORT NUMBER U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION / AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited					
13. SUPPLEMENTARY NOTES no					
14. ABSTRACT We propose to develop a model for identifying lung cancer in indeterminate pulmonary nodules (PNs). Specific aims: (1) identify new sputum small noncoding RNA (ncRNA) biomarkers for lung cancer, (2) validate the previously and newly identified biomarkers and develop a prediction model, and (3) blindly validate the prediction model for distinguishing benign from malignant growths in a different cohort. Until 08/2020, we produced 13 peer-reviewed research articles published. We have the important findings as shown below: Demonstrating that aberrant ncRNAs detected in the bronchial epitheliums of sputum could reflects those in malignant PNs (lung tumors). Developing and validating panels of sputum miRNA, snoRNA, and lncRNA biomarkers for lung cancer. Defining a landscape of various types of ncRNAs of bronchial epitheliums of lung cancer patients. Combined analysis of ncRNAs and DNA methylation has a synergistic effect for lung cancer diagnosis. We have developed a prediction model that has more than 90% sensitivity and 95% specificity for lung cancer detection. The diagnostic value of the prediction model is validated in a different cohort.					
15. SUBJECT TERMS Lung cancer, biomarkers, early detection					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
	Unclassified	Unclassified	Unclassified	1	19b. TELEPHONE NUMBER (include area code)

Table of contents

Front cover	Page 1
Standard Form (SF 298)	Page 2
Table of Contents	Page 3
Introduction	Page 3
Keywords	Page 3
Accomplishments	Pages 4-9
Training and professional development	Page 10
Impact	Page 11
Changes/Problems	Page 11
Products	Page 11-12
Participants and Other Collaborating Organizations	Page 12
Special Reporting Requirements	Page 12
Appendices	Page 12 and the attachment

Introduction,

To complete the objective, we have three specific aims:

Aim 1. Identify new sputum small ncRNA biomarkers for lung cancer from our previously NGS-defined small ncRNA profile of lung tumors.

Aim 2. Validate the previously and newly identified biomarkers and develop a prediction model by integrating the biomarkers with clinical and imaging variables for identifying NSCLC.

Aim 3. Blindly validate the prediction model for distinguishing benign from malignant growths in a different cohort of patients with LDCT-discovered PNs.

We have achieved the goal of the three specific aims. The importance of the findings is described below:

- Demonstrating that aberrant ncRNAs detected in the bronchial epitheliums of sputum could reflect those in malignant PNs (lung tumors)
- Developing and validating panels of sputum miRNA and lncRNA biomarkers for lung cancer by using ddPCR
- Defining a landscape of various types of ncRNAs of bronchial epitheliums of lung cancer patients, providing novel biomarker candidates for malignant PNs
- Demonstrating that ddPCR can sensitively and robustly quantify DNA methylation in clinical specimens
- Combined analysis of ncRNAs and DNA methylation has a synergistic effect for lung cancer diagnosis
- We produced 13 peer-reviewed research articles published.
- We have presented these results at national meetings.

Keywords,

Lung cancer, Veterans, small non-coding RNA, biomarkers, diagnosis.

Accomplishments

1. The major goals of the project

Aim #1 is to identify new sputum small ncRNA biomarkers for lung cancer from our previously NGS-defined small ncRNA profile of lung tumors.

1), we have local IRB protocol approved to collected sputum and blood samples and associated clinical information.

2), we have analyzed the changes of the NGS-defined lung tumor tissue-based ncRNAs (68 miRNAs and 29 snoRNAs) in sputum of patients with either malignant or benign PNs using ddPCR. We developed panels of sputum biomarker for NSCLC.

Particularly, we determined expressions of the lncRNAs in sputum of a discovery cohort of 67 lung cancer patients and 65 smokers with benign diseases. Four of the lncRNAs displayed a different level in sputum of the NSCLC patients vs. cancer-free smokers (all $P < 0.001$). From the lncRNAs, 3 lncRNAs (SNHG1, H19, and HOTAIR) are identified as a sputum biomarker panel with an AUC of 0.9 (Figure 1A), producing 82.1% sensitivity and 89.2% specificity for diagnosis of lung cancer. Furthermore, the biomarker panel has a higher sensitivity (82.1% vs. 52.2%, $P=0.02$) compared with sputum cytology (Figure 1B).

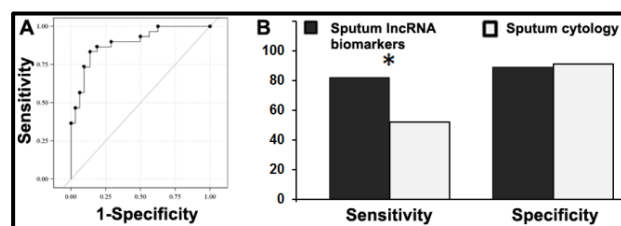


Fig. 3. Three lncRNA were selected as the best sputum biomarkers and incorporated into an algorithm: Probability $P = \exp(U) / [1 + \exp(U)]$, where $U = 2.93 + 2.12 * \log(\text{SNHG1}) - 1.26 * \log(\text{H19}) - 1.1795 * \log(\text{HOTAIR})$. The 3 sputum lncRNAs used in combination by this algorithm produced 0.9 AUC (A). The sputum lncRNA biomarker panel has a higher sensitivity compared with sputum cytology for lung cancer detection (B).

Part of the study was recently published (Sputum long non-coding RNA biomarkers for diagnosis of lung cancer. C Gupta, J Su, M Zhan, SA Stass, F Jiang. 2019. Cancer Biomarkers, 1-9).

3), we have for the first time shown that ddPCR can sensitively and robustly quantify DNA methylation, and further develop a sputum epigenetic classifier for early stage lung cancer.

Conventional qPCR-based platforms, such as methylation-specific PCR (qMSP), have been used for detecting DNA methylation. However, qMSP is an indirect approach, which requires internal controls for data normalization. Furthermore, the sensitivity is poor. We have demonstrated that ddPCR is a straightforward means for precise, direct, and absolute quantification of nucleic acids (e.g., miRNAs, snoRNAs, and lncRNAs), without requiring an internal control gene.

To investigate if ddPCR could sensitively and robustly quantify DNA methylation, we performed methylation-specific ddPCR (ddMSP) and qMSP in specimens of methylated and unmethylated DNA and compared their performance for quantification of DNA methylation. ddMSP had a higher sensitivity for the quantification of DNA methylation compared with qMSP ($p = 0.01$) (Figure 2).

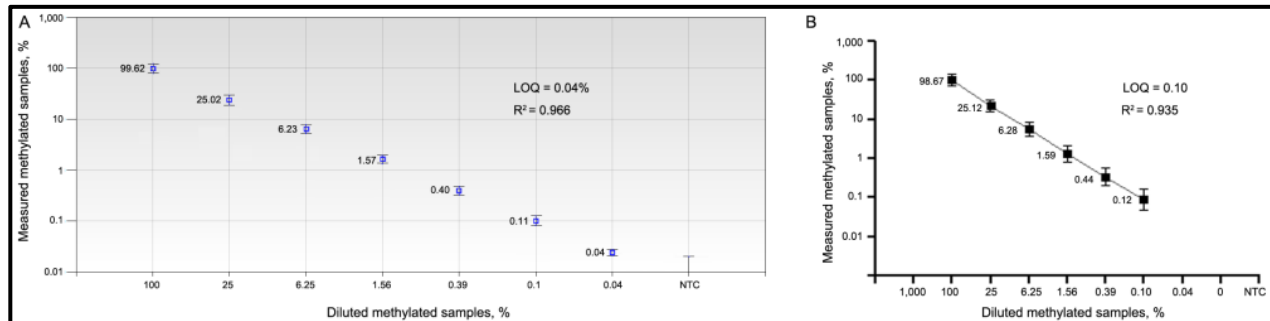


Fig. 2. Comparing the dynamic ranges and sensitivities of ddMSP and qMSP for quantification of DNA methylation. (A), In methylated DNA serially diluted in sputum DNA of a healthy nonsmoker, ddMSP can detect levels of methylated 3OST2 as low as 0.04% (Limit of detection (LOD) = 0.04%). (B), qMSP can detect methylated 3OST2 at 0.10% (LOD = 0.10%) in the same diluted samples ($p = 0.01$).

Furthermore, ddMSP had higher precision and reproducibility than did qMSP (all $p < 0.05$).

To evaluate analytic performance of ddMSP and qMSP in clinical sputum samples, sputum of 20 lung cancer patients and 20 cancer-free controls was tested for *RASSF1A*, whose aberrant methylation level was shown to be elevated in sputum of lung cancer patients. Each well of the samples contained at least 10,000 droplets (Fig. 3A). Therefore, ddMSP analysis of DNA methylation could successfully be performed in clinical sputum specimens. *RASSF1A* analyzed by both the techniques displayed a high methylation level in lung cancer patients vs. controls (all $p < 0.05$). In the ddMSP assay, a specimen with $\geq$ one copy of DNA methylation of *RASSF1A* per microliter was considered to be positive. When the criteria was used, of 20 sputum specimens of lung cancer patients, 11 (55%) had positive methylation of the gene detected by ddMSP. In the qMSP assay, a PMR $\geq 1\%$ was classified as positive for *RASSF1A* in a given sample. When the criteria was used, 9 (45%) were positive for *RASSF1A* by qMSP. The same 4 sputum specimens of control subjects had positive methylation of the gene detected by both ddMSP and qMSP. Therefore, ddMSP analysis of DNA methylation of *RASSF1A* in sputum had a higher sensitivity (55%) than did qMSP (45%) ($p = 0.01$) for distinguishing lung cancer patients from control subjects, while maintaining the same specificity (80%). Furthermore, the CVs of repeated measures by ddMSP on different days by different researchers were approximately twofold lower compared with those generated by qMSP. Altogether, in clinical sputum specimens, ddMSP also exhibited higher sensitivity, accuracy, and reproducibility than did qMSP for quantification of DNA methylation.

We then used ddMSP to analyze 29 potential DNA methylation biomarkers that were previously proposed as potential sputum biomarkers for lung cancer in a set of 127 lung cancer patients and 159 control individuals. All the 29 genes displayed a higher level of methylation in patients vs. controls (all $p < 0.05$). ROC curve and AUC analysis showed that the genes had 29–88% sensitivities and 26–92% specificities in differentiating lung

cancer patients from healthy controls. Four genes (*HOXA9*, *RASSF1A*, *SOX17*, and *TAC1*) were identified as the best biomarkers (all $p < 0.001$) and incorporated into a logistic classifier: Probability of lung cancer = $e^x / (1 + e^x)$, where $x = 1.69 + 1.48 \times \log(\text{HOXA9}) - 1.25 \times \log(\text{RASSF1A}) + 0.27 \times \log(\text{SOX17}) + 0.16 \times \log(\text{TAC1})$. The logistic classifier produced 0.92 AUC for lung cancer detection (Fig. 3B). The logistic classifier had higher accuracy (88.8 vs. 70.63%, $p = 0.004$) and sensitivity (86.6 vs. 44.8%, $p < 0.001$) than did sputum cytology, while maintaining a similar specificity (90.6 vs. 91.2%, $p = 0.34$) (Fig. 3C).

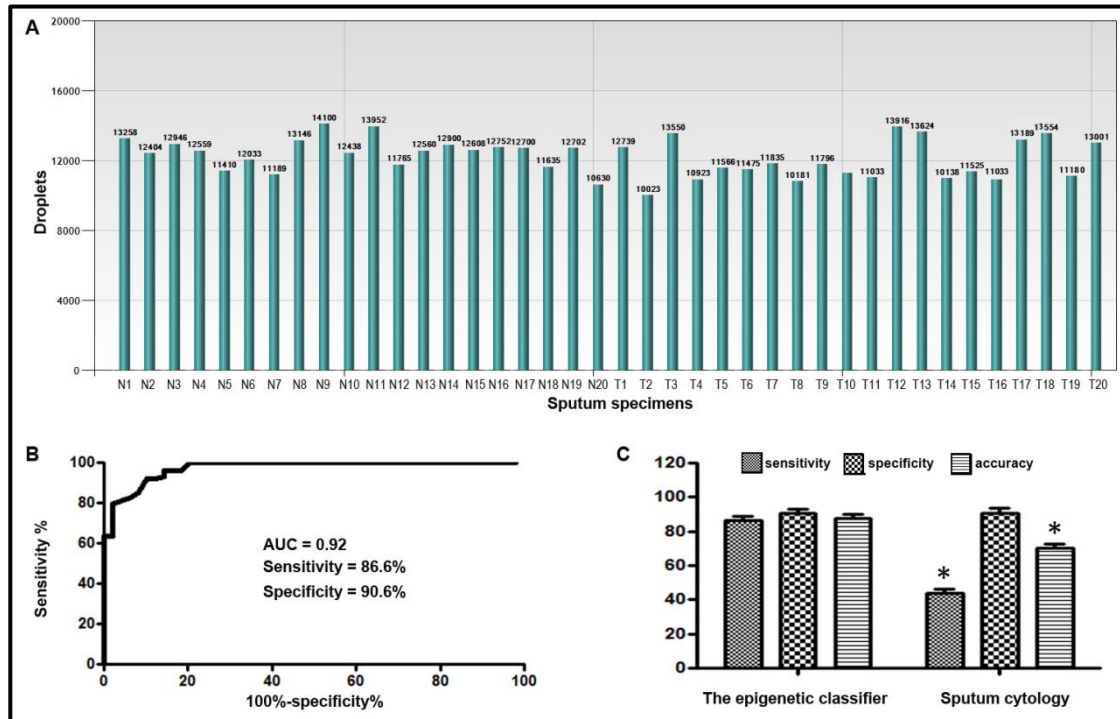


Fig. 3. ddMSP and qMSP analyses of DNA methylation in clinical sputum samples. **A**, ddMSP analysis of sputum samples of 20 cancer-free controls (normal subjects, N) and 20 patients diagnosed with lung tumor (T) for DNA methylation of *RASSF1A*. **B**, ddMSP analysis of 29 genes in a training set of 127 lung cancer patients and 159 controls developed an epigenetic classifier consisting of four DNA methylation biomarkers. The epigenetic classifier produced 0.92 AUC with 86.6% sensitivity and 90.6% specificity for diagnosis of lung cancer. **C**, the epigenetic classifier had higher accuracy (88.8 vs. 70.63%, $p = 0.004$) and sensitivity (86.6 vs. 44.8%, $p < 0.001$) compared with sputum cytology, whereas keeping a similar specificity (90.6 vs. 91.2%, $p = 0.34$).

Part of the study was recently published (An epigenetic classifier for early stage lung cancer. Y Su, HB Fang, F Jiang. 2018, Clinical epigenetics 10 (1), 68).

4), establishing reference samples as surrogates for clinical specimens to develop plasma biomarkers in CLIA-laboratory.

We worked with a Biomarker Laboratory (BDL) and a Biomarker Reference Laboratory (BRL) that is a CLIA laboratory of the NCI-Early Detection Research Network (EDRN) to establish reference samples that could be used as surrogates for clinical specimens. In this study, both BDL and BRL isolated RNA from the cell lines, split and shared the samples with the other site, and profiled the expression profiles of miRNAs. As shown in Fig.4,

miRNA abundance was consistently observed among all possible measurement processes resulting from all combinations of isolation.

Altogether, our developed reference samples have the following properties: 1) express the genes of interest; 2) be re-generable for use over time as a process control; 3) be amenable to identical processing as clinical samples; 4) contain well-established relative differences in the genes; and 5) be commutable between measurement systems.

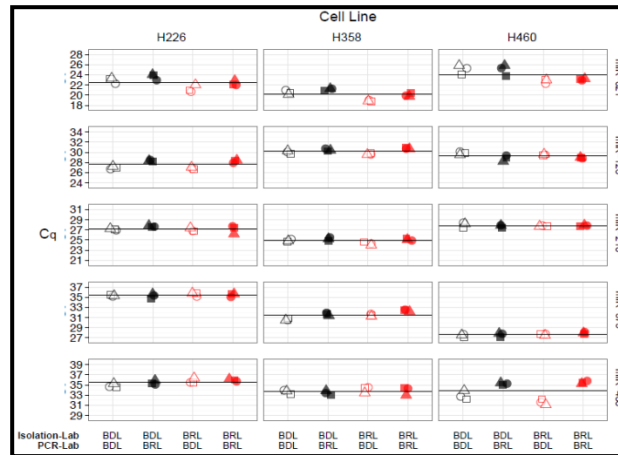


Fig. 4. Profiling of miRNAs in cancer cells. RNAs are isolated in the BDL (black symbols) and BRL (red symbols) on 3 separate occasions (circles, squares, and triangles). qPCR results are from BDL (open symbols) and BRL (filled symbols), where the black line corresponds to the mean Cq for all isolation and PCR combinations^{70, 71}.

Part of the study was recently published (P Scott Pine, Steven P Lund, Sanford A Stass, Debra Kukuruga, Feng Jiang, Lynn Sorbara, Sudhir Srivastava, Marc Salit. BMC biotechnology. 2018).

5), integrating sputum and circulating immunological biomarkers for the early detection of lung cancer

We have demonstrated that assessments of microRNA (miRNA) expressions in circulating peripheral blood mononucleated cell (PBMC) and sputum specimens, respectively, may help diagnose lung cancer. To assess the individual and combined analysis of the miRNAs across the different body fluids for lung cancer early detection, we analyzed a panel of 3 sputum miRNAs (miRs-21, 31, and 210) and a panel of 2 PBMC miRNAs (miRs-19b-3p and 29b-3p) in a discovery cohort of 68 patients with lung cancer and 66 cancer-free smokers.

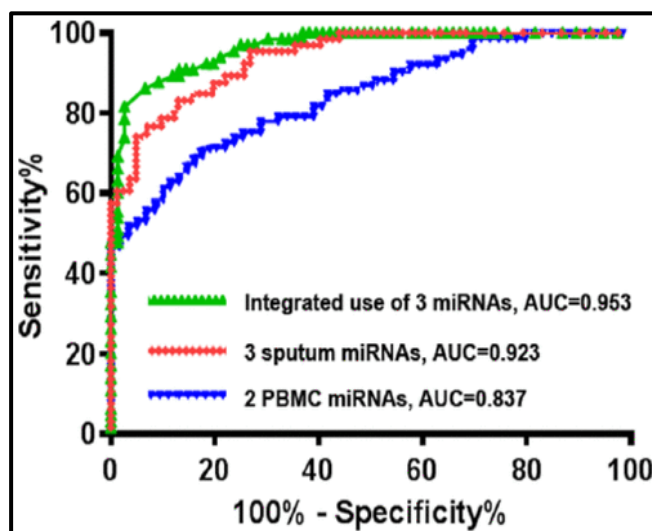


Fig. 5. The comparison of integrated miRNA biomarkers with panels of sputum and PBMC miRNAs in a discovery cohort. A prediction model based on integrated use of 3 miRNA biomarkers (miRs-31, 210, and 19b-3p) across sputum and PBMC specimens was developed for distinguishing lung cancer patients from cancer-free smokers. The ROC curve of the integrated miRNA biomarkers produced a higher AUC (0.953), as compared with the panel of sputum miRNAs (0.923) and the panel of PBMC miRNAs (0.837) (all P's < .05). AUC indicates area under the ROC; miRNAs, microRNAs; PBMC, peripheral blood mononucleated cell.

We find that integrating 2 sputum miRNAs (miRs-31 and 210) and 1 PBMC miRNA (miR-19b-3p) has higher sensitivity (86.8%) and specificity (92.4%) compared with the individual panels (Fig. 4). The synergistic value of the integrated panel of 3 biomarkers is confirmed in a validation cohort, independent of stage and histological type of lung cancer, and patients' age, sex, and ethnicity. Integrating circulating immunological and sputum biomarkers could improve the early detection of lung cancer.

Part of the study was recently published (Jian Su, Qixin Leng, Yanli Lin, Jie Ma, Fangran Jiang, Cheng-Ju Lee, HongBin Fang, Feng Jiang. Biomarkers in cancer. 2018. 10, 1179299X18759297).

Aim 2 is to validate the newly identified sputum biomarkers, previously identified sputum DNA methylation and protein biomarkers, and PBMC biomarkers, and develop a prediction model by integrating the biomarkers with clinical and imaging variables for identifying NSCLC.

1), validating the previously developed biomarkers for lung cancer.

We have analyzed the new panel of small ncRNA biomarkers, ENO1 and three DNA methylation biomarkers in sputum, and two PBMC miRNA biomarkers in patients with either malignant or benign PNs.

Parts of the results are shown in the table below.

Table 1. Successfully validating the biomarkers in the new specimens			
Names	Sample size	Sensitivity	Specificity
3 sputum miRNAs (miR-21-5p, -31-3p, and -210-3p)	117 patients with NSCLC and 116 cancer-free smokers	82.3%	87.9%
2 sputum snoRNAs (snoRD66 and snoRD78)	67 patients with NSCLC and 69 cancer-free smokers	74.6%	84.1%
3 sputum lncRNAs (SNHG1, H19, and HOTAIR)	59 patients with NSCLC and 60 cancer-free smokers	81.4%	88.3%
4 sputum methylation markers (HOXA9, RASSF1A, SOX17, and TAC1)	89 patients with NSCLC and 107 cancer-free smokers	85.4%	88.9%
ENO1	65 patients with NSCLC and 78 cancer-free smokers	82.4%	78.9%
PBMC biomarkers	117 patients with NSCLC and 116 cancer-free smokers	72.1%	81.8%

We validated the miRNAs by using ddPCR in bronchial epithelial cells of the paired sputum samples of lung cancer patients and controls. The miRNAs in sputum displayed changes consistent with those in the matched lung tumor tissues. Therefore, the altered miRNAs detected in the bronchial epitheliums of sputum could reflect those in malignant PNs (lung tumors).

Furthermore, from the miRNAs, we optimized a panel of sputum miRNA biomarkers (miR-21, -31, and -210) for distinguishing malignant from benign PNs with 82% sensitivity and 88% specificity in two internal sets of patients with either malignant or benign PNs.

In addition, we validated the three sputum miRNA biomarkers in a different cohort show that the biomarkers have consistent performance as showed in the discovery phase.

Our recent validation of the sputum miRNA biomarkers using an external testing cohort of patients with either malignant (n=117) or benign PNs (n=116) confirmed its diagnostic value (Figure 6).

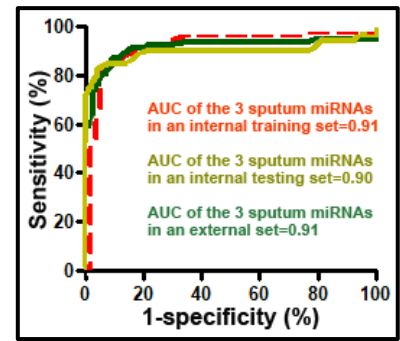


Fig. 6. Comparison of the 3 sputum miRNA biomarkers in internal and external sets of cases and controls shows that the 3 sputum miRNA biomarkers has consistent performance for distinguishing malignant from benign PNs.

2), by working with bio-statistician, we have developed a prediction model that consists of 2 sputum miRNAs (miRs-31 and 210) and 1 PBMC miRNA (miR-19b-3p), smoking pack years, and size of pulmonary nodules. The prediction model has more than 91% sensitivity and 94% specificity for detection of NSCLC in LDCT-found PNs.

Aim 3 is to validate the prediction model for distinguishing benign from malignant growths in a different cohort (Table 2) of patients with LDCT-discovered PNs.

To validate the prediction model for distinguishing benign from malignant growths, we obtained sputum and PBMC samples from an additional cohort (Table 2) of 300 smokers with benign pulmonary growths and 100 smokers with malignant growths (lung cancer).

Please note that the cohort has not been used in previous studies for the development of biomarkers.

The prediction model is validated as judged by the reproducible ability to distinguish between benign and malignant PNs on the different data, and the reproducible excellent calibration throughout a range of the probability of malignancy in this cohort. The prediction model has 90% sensitivity and 95% specificity for distinguishing benign from malignant growths.

Table 2. A validation cohort		
	100 smokers with malignant	300 smokers with benign
Age	Mean, 70	Mean, 68
Sex Male	57	171
Female	43	129
Race		
African	25	75
White	74	223
Others	1	2
Smoking	32 ± 1	31 ± 2
FEV1	0.45-0.76	0.37-0.78
Histology	55 AC and 45	
TNM Stage I	33	
II	33	
III-IV	34	

In sum, **we have achieved the goals of the Specific Aims.**

Training and professional development opportunities or there is nothing significant to report during this reporting period.

"Nothing to Report."

Describe opportunities for training and professional development provided to anyone who worked on the project or anyone who was involved in the activities supported by the project.

During the funded period, with the support, a research-postdoctoral fellow has been trained in the field of development of molecular biomarker for lung cancer.

Workshops, conferences, seminars, study groups, and individual study. Include participation in conferences, workshops, and seminars not listed under major activities.

We have presented our studies in March 2019 EDRN Steering Committee Meeting, Nashville, TN and 11th EDRN Scientific Workshop, September 10-12, 2019 Bethesda, Maryland and

How were the results disseminated to communities of interest?

We have published seven papers in peer-reviewed Journals.

If there is nothing significant to report during this reporting period

"Nothing to Report."

Describe how the results were disseminated to communities of interest. Include any outreach activities that were undertaken to reach members of communities who are not usually aware of these project activities, for the purpose of enhancing public understanding and increasing interest in learning and careers in science, technology, and the humanities.

"Nothing to Report."

If this is the final report

Yes

Describe briefly what you plan to do during the next reporting period to accomplish the goals and objectives.

We will prospectively validate this prediction model in a large lung cancer screening program and bring it into clinical settings for patient care.

Impact

1. Future use of the prediction model for excluding lung cancer in a CT screening positive setting will spare Veterans with benign growths from harmful diagnostic and therapeutic procedures, and enable effective treatments to be immediately initiated for lung cancer. The prediction model will complement LDCT for the early detection of lung cancer, and offer the best opportunity for treatment. The use of the prediction model in clinical settings will reduce the tremendous burden of NSCLC on patients, families and societies, which is the number one cause of cancer deaths among U.S. military Veterans.
2. The project will advance scientific knowledge or clinical practice in the early detection of NSCLC, because the noninvasive approach could overcome obstacles of conventional techniques. Particularly, the success of the project will increase the molecular diagnostic value and practical usefulness of sputum and blood. Furthermore, the research project will translate the strengths of small ncRNA profiling data of next generation sequencing analysis into clinical practice by cost-effective assays. Therefore, the proposed project will bridge the gap between basic research and patient care for Veterans.
3. The study will have important factors on the concepts, methods, technologies, treatments, services, or preventative interventions that can drive the research for lung cancer early detection. Particularly, future use of the prediction model as the first screening step and use of CT scan or bronchoscopy as the second line test in early detection and/or prevention programs of lung cancer will minimize unnecessary cost, workload, pain and suffering for the Veterans patients. Furthermore, the use of the simple prediction model would lead to more personalized therapy by enabling effective treatments to be initiated at an early stage of the malignancy in Veterans.

Changes/problems

No.

Products,

13 Peer-reviewed articles

1. An epigenetic classifier for lung cancer. Y Su, HB Fang, F Jiang. Clinical epigenetics 10 (1), 68. 2018.
2. Cell-based reference samples designed with specific differences in microRNA biomarkers. PS Pine, SP Lund, SA Stass, D Kukuruga, ***F Jiang***, L Sorbara, et al. BMC biotechnology 18 (1), 17. 2018.
3. Fucosylation genes as circulating biomarkers for lung cancer. Q Leng, JH Tsou, M Zhan, ***F Jiang***. Journal of cancer research and clinical oncology 144 (11), 2109-2115, 2018.
4. A plasma long noncoding RNA signature for early detection of lung cancer. Y Lin, Q Leng, M Zhan, ***F Jiang***. Translational oncology 11 (5), 1225-1231. 2018.
5. A direct plasma miRNA assay for early detection and histological classification of lung cancer. Q Leng, Y Wang, ***F Jiang***. Translational oncology 11 (4), 883-889. 2018.
6. An integromic signature for lung cancer early detection. Q Leng, ***F Jiang***. Oncotarget 9 (37), 2018.

7. Radiomics analysis of pulmonary nodules in low-dose CT for early detection of lung cancer. W Choi, JH Oh, S Riyahi, CJ Liu, **F Jiang**, C White, A Rimner, et al. Medical physics 45 (4), 1537-1549. 2018.
8. Integrating circulating immunological and sputum biomarkers for the early detection of lung cancer. J Su, Q Leng, F Jiang, CJ Lee, HB Fang, **F Jiang**. Biomarkers in cancer 10, 117-123. 2018.
9. A non-coding RNA landscape of bronchial epitheliums of lung cancer patients. J Ma, MA Guarnera, **F Jiang**. Biomedicines 8 (4), 88. 2020.
10. Sputum long non-coding RNA biomarkers for diagnosis of lung cancer. C Gupta, **F Jiang**. Cancer biomarkers. 6, 23-9. 2019.
11. MicroRNA-based biomarkers for diagnosis of non-small cell lung cancer. J Liao, C Gupta, **F Jiang**. Thoracic Cancer. 8, 26-31. 2020.
12. Integrated analysis of miRNAs and DNA methylation identifies miR-132-3p as a tumor suppressor in lung adenocarcinoma. Y Su, A Shetty, **F Jiang**. Thoracic Cancer.
13. A CRISPR Test for Detection of Circulating Nuclei Acids. JH Tsou, Q Leng, **F Jiang**. Translational oncology 12 (12), 1566-1573.

Participants and other collaborating organizations,

The Veterans Affairs Maryland Health Care System

The University of Maryland School of Medicine

Special reporting requirements

No.

Appendices.

Please see the attached file with the 13 published papers.



Article

A Non-Coding RNA Landscape of Bronchial Epitheliums of Lung Cancer Patients

Yanli Lin ¹, Van Holden ² , Pushpawallie Dhilipkannah ¹, Janaki Deepak ² , Nevins W. Todd ² and Feng Jiang ^{2,*}

¹ Departments of Pathology, University of Maryland School of Medicine, 10 S. Pine St. Baltimore, MD 21201, USA; liny11089@163.com (Y.L.); PDhilipkannah@som.umaryland.edu (P.D.)

² Department of Medicine, University of Maryland School of Medicine, 22 S. Greene St. Baltimore, MD 21201, USA; VHolden@som.umaryland.edu (V.H.); jadeepak@som.umaryland.edu (J.D.); ntodd@som.umaryland.edu (N.W.T.)

* Correspondence: fjiang@som.umaryland.edu

Received: 4 March 2020; Accepted: 9 April 2020; Published: 13 April 2020



Abstract: We propose to systematically identify a non-coding RNA (ncRNA) profile of exfoliated bronchial epitheliums of sputum from lung cancer patients. Bronchial epithelial cells enriched from sputum of 32 lung cancer patients and 33 cancer-free smokers were analyzed by next-generation sequencing to comprehensively characterize the ncRNA profiles. In addition, 108 miRNAs, 88 small nucleolar RNAs, 13 piwi-interacting RNAs, 6 transfer RNAs, 4 ribosomal RNAs, 19 small nuclear RNAs, and 25 long-noncoding (lnc) RNAs displayed a significantly different level in bronchial epitheliums of sputum of lung cancer patients versus cancer-free smokers (all <0.001). PCR analysis confirmed their different expression levels in the sputum specimens. A high expression of SNHG9, an lncRNA, was validated in 78 lung tumor tissues, and the expression was inversely associated with overall survival of lung cancer patients ($p = 0.002$). Knockdown of SNHG9 in cancer cells reduced the cell growth, proliferation, and invasion in vitro and tumorigenesis in vivo. The multiple differentially expressed ncRNAs in bronchial epitheliums may contribute to the development and progression of lung cancer and provide potential biomarkers and therapeutic targets for the disease.

Keywords: lung cancer; non-coding RNAs; sputum; function; and biomarkers

1. Introduction

Lung cancer is the number one cancer killer in both men and women [1]. Smoking causes 90% lung cancer cases, of which 85% are non-small cell lung cancers (NSCLC) [1]. NSCLC mainly consists of adenocarcinoma (AC) and squamous cell carcinoma (SCC). Lung tumors develop from a field defect characterized by molecular abnormalities resulted from repeated exposure of the entire airway to the tobacco carcinogens [2]. Airway field changes, such as loss-of-heterozygosity, methylation, and mutation of DNA, and gene expression profiles, were found between primary lung tumors and normal appearing airway epitheliums. Furthermore, prior investigations in field carcinogenesis of the lung airway have proven that the molecular alterations in the large bronchial airways reflected those in primary lung tumors in the distal lung, regardless of the anatomic location relative to the tumors. Therefore, profiling the molecular aberrations in bronchial epitheliums could provide important biologic insights into lung tumorigenesis, and potential biomarkers and therapeutic targets for the disease.

Non-coding RNA molecules (ncRNAs) are a large and diverse class of transcribed RNA molecules and have participated in many biological processes. ncRNAs are divided into two categories based on their length. ncRNAs <200 nucleotides-long are referred to as small ncRNAs, which include transfer (t)

RNAs, ribosomal (r) RNAs, micro (mi) RNA, small nucleolar (sno) RNAs, small nuclear (sn) RNAs, piwi-interacting RNA (pi-RNA), etc. [3]. Long noncoding RNAs (ncRNAs) are transcripts longer than 200 nucleotides. ncRNAs play critical roles in the development and progression of cancer [4,5]. For instance, numerous studies including our own have found that dysregulations of miRNAs can drive multiple processes of lung tumorigenesis by regulating cell cycle, apoptosis, and migration [6,7]. Furthermore, we have demonstrated that aberrant miRNAs detected in sputum can reflect those in primary lung cancer, thus providing potential biomarkers for lung cancer [8–14].

Aside from miRNAs whose functions in tumorigenesis have been extensively investigated, other ncRNAs (or non-miRNA ncRNAs) are emerging as important players in the development and progression of NSCLC. However, aberrations of other types of ncRNAs in bronchial epitheliums of lung cancer patients remain largely unexplored. Using global next generation sequencing (NGS), in this study, we systematically characterized changes of various types of ncRNAs in exfoliated bronchial epithelial cells enriched from sputum of lung cancer patients. The identified multiple differentially expressed ncRNAs of airway epitheliums may contribute actively to lung cancer development and progression, making them potential selective targets for lung cancer diagnosis and therapeutics.

2. Experimental Section

2.1. Collecting Sputum and Enriching Exfoliated Bronchial Epithelial Cells

This study was approved by the University of Maryland Baltimore Ethics Committee under ethic codes HP-00040666. Written informed consent was obtained from all enrolled individuals. Before receiving any treatment, participants were instructed to spontaneously cough sputum as described in our previous studies [8–11,15–21]. When the participants (mainly former smokers) were not able to spontaneously cough sputum, they underwent sputum induction using a Lung Flute (Medical Acoustics, Buffalo, NY, USA)-based technique as described in our previous work [17]. Sputum was centrifuged at 1000× *g* for 15 min. Cytospin slides were prepared and underwent Papanicolaou staining for evaluating whether the specimens were representative of deep bronchial cells. Sputum has remarkable cell heterogeneity with a large number of macrophages and neutrophils and a limited number of bronchial epithelia [20]. The cell heterogeneity presents a major difficulty in analysis of bronchial epithelial cells of sputum. We have developed a protocol to enrich exfoliated bronchial epithelial cells for analysis of molecular changes in sputum [20]. Using the sample protocol, we purified bronchial epithelial cells from the sputum samples. The bronchial epithelial cell pellet from each sample was washed in phosphate buffered saline (Sigma-Aldrich, St. Louis, MO, USA) and stored at −80 °C until being tested.

2.2. NGS Analysis

RNA was extracted from the cells as described in our previous reports [22–25]. A NanoDrop spectrophotometer (Thermo Fisher Scientific Inc. Waltham, MA, USA) was used to determine the purity and concentration of RNA from OD260/280 readings as described in our published articles [26,27]. Agilent's RNA 6000 Nano lab-on-a-chip kit and Bioanalyzer (Agilent Technologies) were used to determine the RNA integrity number (RIN) [26,27]. Only RNA with a 260/280 ratio of 1.8–2.0 and a RIN of ≥7 underwent NGS analysis to ensure accurate determination of expression levels of ncRNAs in each sample [26,27]. Illumina HiSeq 2000 system (Illumina, Inc., San Diego, CA, USA) was used to define ncRNA profiles as described in our published articles [26,27]. Briefly, we started with 100 ng RNA, which was diluted in 1 µL nuclease-free dH₂O as a ligation reaction. RT and random primers were used to create cDNA. Illumina PE adapters were used to create cDNA libraries. Bioinformatics analysis of the NGS data was performed as previously described [26,27]. In addition, 100 nucleotides of genomic sequences flanking each side of the sequences were extracted, and the RNA secondary structures were predicted using RNAfold. The raw reads were mapped and aligned to sequenced databases (miRBase, piRBase, Genomic tRNA database, Ensembl annotations, Repeat Masker annotation, and

the human genes UCSC). The raw reads were also mapped and aligned with lncRNAs by using reference annotation LNCipedia version 3.1. Read counts of identified ncRNAs were normalized to the total number of reads. After normalization, DESeq2 was used to identify ncRNAs that were differentially expressed at 2.0-fold change cutoff in sputum samples of lung cancer patients versus cancer-free smokers.

2.3. Reverse Transcription-PCR (RT-PCR) Analysis of NcRNAs in Sputum

To validate the results generated by NGS analysis, we used a different platform, RT-PCR, to test 10 genes in the samples set of specimens. The criteria used the validation was that RT-PCR results of the genes should have ≥ 2 FC difference between specimens of lung cancer patients compared with controls in the same direction as by NGS, although the scale of changes differed between the two methods. The expression levels of ncRNAs were determined in sputum by using RT-PCR with Taqman assays (Applied Biosystems, Foster City, CA, USA) as previously described [28,29]. The cycle threshold (Ct) was defined as the number of cycles required for the fluorescent signal to cross the threshold. Relative expression of a targeted lncRNA in a given sample was computed using the equation $2^{-\Delta Ct}$, where $\Delta Ct = Ct(\text{targeted lncRNA}) - Ct(U6)$ and U6 was used as an internal control. All experiments were repeated three times, each sample in triplicate. Data represented the mean ($\pm$ SD) of three independent experiments.

2.4. RNA Interference

Specific siRNA targeting RNA sequence of SNHG9 (NCBI accession number NR_003142.2) and corresponding scrambled sequence were designed and synthesized by Integrated DNA Technologies, Inc. (IDT). Transfection was performed by using Opti-MEM medium and LipofectamineTM RNAiMAX (Invitrogen, California, CA, USA) as previously described [30,31]. All experiments were repeated three times, each sample in triplicate. Data represent the mean ($\pm$ SD) of three independent experiments.

2.5. Cell Viability, Proliferation, and Colony Formation Assays

Cell viability assay was carried out by using Cell Counting Kit 8 (WST-8/CCK8) (Abcam) as previously described [32]. Briefly, cells with a quantity of 5×10^3 cells/well were seeded in a 96-well plate and grown to 80% confluence. CCK-8 reagent was added into each well. Cellular viability was determined by measuring the absorbance of the converted dye at 450 nm at 0, 24, 48, and 72 h. Methylthiazol tetrazolium (MTT) assay was performed to determine cell proliferation as previously described [7,30,31]. Colony Formation Assays were performed as previously described [30,31]. Briefly, the cells were counted and seeded in 6-well plates at a density of 1000 cells per well after transfection. Culture medium was replaced every 3 days and cells allowed to grow for 2, 4, 6, 8, 10, 12, and 14 days. The cells were stained using crystal violet, and each colony with a minimum of 50 cells was counted at each time point. All experiments were repeated three times, each sample in triplicate. Data represented the mean ($\pm$ SD) of three independent experiments.

2.6. Transmigration and Wound Healing Assays

To determine migration and invasion of cells, cells were plated in medium without serum in the top chamber of a transwell (Corning, Moneta, NY, USA). The bottom chamber contained standard medium with 10% FBS. After incubation for 24 h or 48 h, the cells that had migrated to the lower surface of the membrane were fixed with formalin and stained with crystal violet. The migrating cells were examined microscopically and determined by counting the migrating/invasive cells in 5 randomly selected fields using an Olympus BX41 microscope. Photomicrographs were taken using a Qcolor5 digital camera system fitted to this microscope. To determine whether the inhibition of cell migration by SNHG9 knockdown was due to the inhibition of cell proliferation, the cells transfected with siRNA-SNHG9 were treated with aphidicolin (1 mg/mL) (Sigma-Aldrich), a proliferation inhibitor, for 24 h or 48 h. A wound healing assay was performed in 12 well plates (1×10^5 per well). When

the cells were grown to 90 to 95% confluences, transfection of cancer cells with siRNA-SNHG9 and scrambled siRNA as well as mock transfection were performed. Wound lines were created manually by scratching the monolayer with a sterile 200 mL pipette tip and migration of the cells was assessed after 24, 48, and 72 h. Pictures were taken using a Nikon inverted phase-contrast microscope (Nikon, Melville, NY, USA). The distance between the parallel lines was measured using ImageJ software. All experiments were carried out at least three times. All experiments were repeated three times, each sample in triplicate. Data represented the mean ($\pm$ SD) of three independent experiments.

2.7. Cell Cycle Analysis by Flow Cytometry

DNA content was analyzed by using flow cytometry as previously described [30,31]. Briefly, harvested cells were washed twice with 5 mM EDTA/PBS and fixed with 95–100% cold ethanol and kept at 4 °C overnight. Cells were incubated with RNase A (50 μ g/mL), and then stained with Propidium iodide (PI, 10 mg/mL) (Sigma) for DNA content analysis on a FACScan flow cytometer (Becton Dickinson). Statistical analysis of variance (ANOVA) and least significant difference *t*-tests were used to compare cell cycle data of cells with and without specific siRNA treatments. All experiments were repeated three times, each sample in triplicate. Data represented the mean ($\pm$ SD) of three independent experiments.

2.8. Tumorigenicity Assays in Nude Mice

The animal study was performed with the approval of the University of Maryland Baltimore under code IACUC# 0516007. Seven athymic Balb/c, Nu/Nu mice per group were subcutaneously inoculated with 1×10^6 A549 cells transfected with SHGN9-siRNA and 1×10^6 A549 cells with scrambled siRNA, respectively. All the animals were monitored regularly, and tumor growth was measured at regular intervals. The mice were observed for four weeks and then euthanized under deep anesthesia with pentobarbital (Sigma). We calculated volume of the tumors by using formula (length (mm)) $\times$ (width (mm)) $\times 0.52$. The tumor size was represented by mean $\pm$ SD mm³.

2.9. Statistical Analysis

Based on our previous studies [26,27,33–36], the standard deviation of gene expressions was estimated to be 0.757, which was used to guide our power analysis. Since the experiments involved a large number of genes, we controlled the false discovery rate at a 0.025 level. Assuming the proportion of non-expressed ncRNAs was 99%, with sample size of 28 per group, we had 82% power to detect ncRNAs with a fold change (FC) ≥ 2 . If the proportion of non-expressed ncRNAs was lower than 99%, the power was even higher with the same sample size [37]. A specific gene was deemed to be significantly differentially expressed if the *p*-value is ≤ 0.001 with a $2 \geq$ FC. Both univariate and multivariate Cox proportional hazard models were applied to assess the effect of clinical variants and SNHG9 expression on survival data of patients. Association of expression of SNHG9 with survival rate was analyzed by using the Kaplan–Meier method. We used ANOVA and least significant difference *t*-tests to compare cell cycle data of the cells with and without specific siRNA treatments. All experiments, except animal study, were repeated three times, each sample in triplicate. For animal study, with seven mice per group, we could have at least 80% statistical power to determine a 2-fold difference in mean tumor volume between groups, tumor volume variation of up to 50% within groups, and a statistical significance level of $\alpha = 0.05$ [37]. Results of the animal study were analyzed and compared by using Fisher's exact test between the two groups. A value of $p < 0.05$ was considered statistically significant.

3. Results

3.1. Study Population

For sputum collection, the participants were recruited at the point of their referral for suspected lung cancer between the ages of 55–80. Exclusion criteria included pregnancy, current pulmonary

infection, surgery within six months, radiotherapy within one year, and life expectancy of <1 year. Furthermore, 78 frozen NSCLC tumor tissues and the matched noncancerous lung tissues were obtained from a tissue bank of the University of Maryland Marlene and Stewart Greenebaum Comprehensive Cancer Center. The frozen tissues specimens were collected from consecutive patients with stage I NSCLC who underwent either lobectomy or pneumonectomy from 1999 to 2013. We were able to obtain complete medical records and follow-up data for the patients. Clinical diagnosis of lung cancer was made using histopathologic examinations of specimens obtained by CT-guided transthoracic needle biopsy, transbronchial biopsy, video-assisted thoracoscopic surgery, or surgical resection. The surgical pathologic staging was determined according to the TNM classification of the International Union Against Cancer with the 8th American Joint Committee on Cancer and the International Staging System for Lung Cancer. Histopathological classification was determined according to the World Health Organization classification. Overall survival of the lung cancer patients was obtained by reviewing documentation that the patient was deceased in the medical record or using the Social Security Death Index (SSDI). Demographic and clinical characteristics of the cases and controls from whom sputum were collected are shown in Table 1. Clinical characteristics and histopathological data of the surgically resected tissues specimens are shown in Table 2.

Table 1. Demographic and clinical data of 32 lung cancer patients and 33 cancer-free smokers.

	Lung Cancer Patients	%	Cancer-Free Controls	%
Parameter				
Mean of age (years)	68.2		65.6	
Gender				
Men	21	65.6	22	66.7
Women	11	34.4	11	33.3
Race				
White American	23	71.9	23	69.7
African American	9	28.1	10	30.3
Mean of smoking pack-years	49.3		37.8	
Stage	All are stage I NSCLC			
Histology				
Adenocarcinoma	16	50		
Squamous cell carcinoma	16	50		

Abbreviations: NSCLC, non-small cell lung cancer.

Table 2. Demographic and clinical characteristics of 78 NSCLC patients.

Characteristics	Number of Patients in Each Category (%)
Mean of age (years) at diagnosis	66.9
Sex	
Male	52 (66.7)
Female	26 (33.3)
Race	
White	49 (62.8)
Black	29 (37.2)
Smoker	
Yes	70 (89.7)
No	8 (10.3)
Tumor histology	
Squamous cell carcinoma	39 (50.0)
Adenocarcinoma	39 (50.0)
T stage	
I	30 (38.5)
II	29 (37.2)
III–IV	19 (24.3)

NSCLC, non-small cell lung cancer.

3.2. Enriched Bronchial Epithelial Cells of Sputum

Sputum was successfully collected from 65 of 70 subjects, including 32 stage I lung cancer patients and 33 heavy smokers. All the selected sputum specimens were mucoid and of lower respiratory origin as indicated by the presence of bronchial epithelial cells. After being enriched by magnetic-assisted cell sorting, sputum yielded the average yield of bronchial epithelial cells of $386,953 \pm 20,425$ (mean $\pm$ standard deviation, SD). Given that the molecular alterations in bronchial epithelial cells of airway reflect the changes in lung tumors, the analysis of the enriched bronchial epitheliums, rather than mixed sputum cells, for the assessment of the molecular changes of lung tumors could provide an efficient tool for molecular analysis of NSCLC in sputum.

3.3. Differentially Expressed ncRNAs in Bronchial Epithelial Cells from Sputum of Stage I NSCLC Patients versus Cancer-Free Smokers

Twenty-eight sputum samples of 32 stage I NSCLC patients and 29 of 33 sputum samples cancer-free smokers (Table 1) were successfully sequenced by using Illumina HiSeq 2500 system (Illumina, San Diego, CA, USA). The rest of the sputum samples had degraded RNA and did not meet our defined minimum requirements, and thus did not proceed with NGS. Each of the enriched bronchial epithelia samples was deep sequenced twice. The number of raw reads obtained per specimen ranged from 17,858,276 to 28,834,167 (average = 23,346,222). Raw reads in the replicates were highly correlated (all $p < 0.0001$), demonstrating that the deep sequencing approach could produce robust results. From the raw reads, an average of 20,896,621 reads (ranging from 15,445,619 to 20,896,621) were filtered. The raw reads were mapped and aligned to the ncRNA sequenced databases. In addition, 2422 miRNAs, 1946 snRNAs, 1522 snoRNAs, 226 piRNAs, 128 tRNAs, 532 rRNAs, and 5061 lncRNAs were annotated in the sputum samples. From the annotated ncRNAs, by using DESeq2, we identified 108 miRNAs, 88 snoRNAs, 13 piRNAs, 6 tRNAs, 4 rRNAs, 19 snRNAs, and 25 lncRNAs, which were differentially expressed at 2.0-fold change cutoff in sputum of cancer patients versus cancer free controls (all $p \leq 0.001$) (Table S1) (Figure S1). The top 65 altered ncRNAs in bronchial epitheliums of sputum of lung cancer patients versus controls are listed in Table 3. Furthermore, the change levels of the ncRNAs were higher in sputum of lung cancer patients with SCC versus patients with AC (all $p < 0.05$), suggesting that the aberrant ncRNAs presenting in sputum could be more closely related to SCC.

Table 3. Fold-change (FC) of top 65 altered ncRNAs in bronchial epitheliums of sputum of lung cancer patients versus controls *.

Genes	FC	Genes	FC
miRs		piRNAs	
MIR-9-1	28.146	piR-004987	5.637
MIR-577	21.745	piR-020809	5.058
MIR-410	21.007	piR-023338	−3.489
MIR-487B	18.242	piR-011186	−4.161
MIR-409	14.424	tRNAs	
MIR-338	−7.723	TRNAV33P	16.688
MIR-486	−8.086	TRNAE27P	15.772
MIR-135A1	−19.011	TRNAG34P	2.159
MIR-184	−28.380	TRNAG32P	16.774
snRNAs		TRNAE40P	2.107
RNU5E-1	32.515	TRNAK42P	−2.006
U4	11.005	rRNAs	
RNU7-1	6.874	RN5-8S5	4.936
RNU4ATAC	6.400	RN5-8S3	2.008
RNU5A-1	5.736	RN5-8S2	5.151
snoRNAs		RN5248	−2.188
SNORD114-20	43.008	lncRNAs	
SNORD113-5	36.658	SNHG9	8.474
SNORD114-25	31.656	SNHG2	7.568

Table 3. Cont.

Genes	FC	Genes	FC
SNORD114-28	30.173	MEG8	7.078
SNORD114-26	22.244	LINC00461	7.786
SNORD113-7	19.579	SNHG11	6.635
SNORD114-21	17.826	GAS5	−3.082
SNORD33	16.235	TUG1	−4.866
SNORD114-23	16.090	PANDAR	−5.099

* All $p > 0.05$.

3.4. Validation of the ncRNAs in Sputum of Stage I NSCLC Patients versus Cancer-Free Smokers

To confirm the deep sequencing results, we used qRT-PCR to assess expression of four miRNAs, three snoRNAs, and three lncRNAs in the same sputum samples of 28 stage I NSCLC patients and 29 cancer-free smokers (Table 4). One of the 10 ncRNAs was miR-486. Previously published papers from our and others' laboratories have proven that miR-486 is tumor a suppressor miRNA in NSCLC and is downregulated in specimens of lung cancer patients [6,7,38]. Of the ncRNAs tested, all had >35 Ct value in 90% of the sputum specimens. The results suggested that the ncRNAs were reliably measured in sputum by using a different technique. Furthermore, all the 10 ncRNAs displayed changes by qRT-PCR in the same direction as by NGS, even though the magnitude of changes differed between the two methods (Table 4). Therefore, the aberrant ncRNAs in bronchial epitheliums of lung cancer patients were confirmed by using a different technique. Furthermore, of the 10 tested ncRNAs, there were five genes whose changes were associated with smoking Status (Table S2).

Table 4. Expression of 10 ncRNAs detected by NGS and RT-PCR in sputum of stage I non-small cell lung cancer patients vs. cancer-free smokers.

Expression Levels of the ncRNAs Detected by NGS				
ncRNAs	Median of RPM in NSCLC Patients	Median of RPM in Controls	Log2 Fold Change (FC) of Patients/Controls	p-Value
miRNAs				
miR-21	39789.25	578.96	6.10	<0.001
miR-31	4800.34	973.20	2.30	<0.001
miR-210	5935.71	226.72	4.71	<0.001
MIR-486	140.87	37632.26	−8.06	
snoRNAs				
snoRD66	4822.28	151.60	4.99	<0.001
snoRD78	3983.48	161.80	4.62	<0.001
snoRA42	4518.27	86.72	5.70	<0.001
lncRNAs				
SNHG9	70896.26	198.99	8.48	<0.001
H19	6073.51	160.89	5.24	<0.001
HOTAIR	2616.69	309.18	3.08	0.0060
Expression levels of the ncRNAs detected by RT-PCR				
ncRNAs	Mean of Level in NSCLC Patients	Mean of Level in Controls		p-Value
miRNAs				
miR-21	297.26	7.16	5.38	<0.001
miR-31	4.89	0.75	2.70	<0.001
miR-210	178.94	12.86	3.80	<0.001
MIR-486	19	378	−4.31	
snoRNAs				
snoRD66	1.71	0.04	5.42	<0.001
snoRD78	1.86	0.12	3.95	0.001
snoRA42	2.65	0.1	4.73	<0.001
lncRNAs				
SNHG9	16.78	0.33	5.67	<0.001
H19	9.34	0.86	3.44	<0.001
HOTAIR	3.11	0.38	3.03	<0.001

Abbreviations: NSCLC, non-small cell lung cancers; NGS, next generation sequencing; RT-PCR, reverse transcriptase PCR; RPM, reads per million.

3.5. Dysregulation of SNHG9 Contributes to Tumorigenesis of NSCLC

Since SNHG9 was one of the lncRNAs that displayed the highest expression level in bronchial epitheliums of sputum from lung cancer patients, we performed in vitro and in vivo analyses to investigate its possible role in tumorigenesis of NSCLC. SNHG9 was highly expressed in ten (76.9%) of the 13 tested cancer cell lines compared with their normal counterpart (Figure S2). Small interfering RNAs specifically targeting SNHG9 (SNHG9-siRNA) were designed that could dramatically reduce SNHG9 expression level in the cancer cells with the treatment (Figure S3). Cell viability was decreased by approximately 47.6% in cancer cells transfected with SNHG9-siRNA compared to cells with scrambled siRNA ($p = 0.0017$) (Figure 1A) (Figure S4A). Furthermore, cell proliferation was reduced in NSCLC cells treated with SNHG9-siRNA compared to cancer cells treated with scrambled siRNA (Figure 1B) (Figure S4B). In addition, suppression of SNHG9 decreased the colony formation, migration, and invasion of the lung cancer cells (Figure 1C–E) (Figure S4C,D). To evaluate whether the inhibition of cell migration of the cancer cells was due to the inhibition of cell proliferation, a proliferation inhibitor, aphidicolin, was added in the Migration assays. Aphidicolin did not affect cell proliferation and migration states of the cancers treated with SNHG9 knockdown (Figures S5 and S6), suggesting that the phenotype was independent of cell proliferation and migration. Moreover, the knockdown of SNHG9 in lung cancer cells increased the percentage of cells in the G1/G0 phase by more than 10%, whereas it reduced the percentage of S phase at least 10%, as compared with cancer cells without knockdown of SNHG9 (all $p < 0.05$) (Figure 1F).

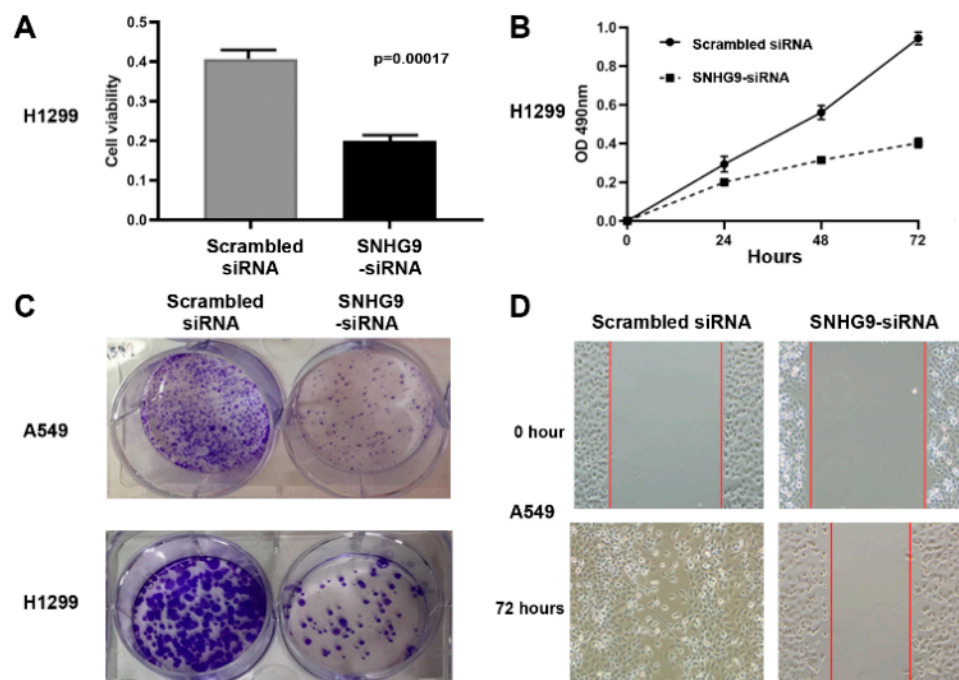


Figure 1. Cont.

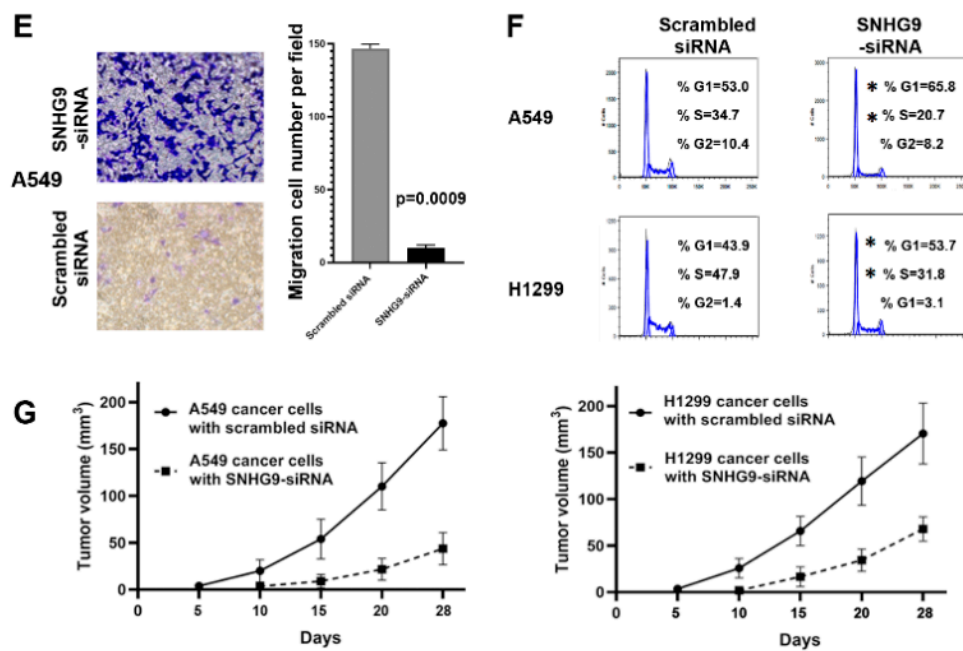


Figure 1. SNHG9 knockdown inhibits the tumorigenicity of A549 and H1299 lung cancer cells. (A) SNHG9 knockdown can significantly reduce cell viability by 72 h. The cell viability is determined by a cell viability assay using Cell Counting Kit 8 (Abcam); (B) SNHG9 knockdown can suppress cell proliferation in the cancer cells treated with SNHG9-siRNA; (C) SNHG9 knockdown significantly inhibits colony formation of cancer cells. The figure shows the results one week after the cells are seeded in the plates (34.8 mm diameter). (D) In the wound-healing assays, cancer cells transfected with SNHG9-siRNA show a slower gap closure compared with cells transfected with scrambled siRNA. The figure only shows the results of A549 cells from the time points 0 and 72 h, respectively. (E) Transwell migration assays show that SNHG9 knockdown can constrain migration and invasion of A549 cancer cells. The migratory cells are counted, and the results are expressed as the mean number of migratory cells $\pm$ SD/selected microscopic field ($n = 5$). The number of migrated cells was evaluated by counting 5 random fields at $\times 50$ magnification. The figure shows the results from the time point 24 h of A549 cells. (F) Knockdown of SNHG9 in lung cancer cells (A549 and H1299) elevates the percentage of the cancer cells in the G1/G0 phase and reduces the percentage of S phase (*, $p < 0.05$). (G) Quantification of the tumor size over the time in vivo experiment shows that tumor growth in the mice injected with SNHG9-siRNA cancer cells (A549 and H1299) was significantly lower than in the mice injected with the cells transfected with scrambled siRNA. Tumor sizes (mean $\pm$ SD³) were measured at the indicated intervals and plotted.

We subcutaneously inoculated A549 or H1299 cancer cells with SNHG9-siRNA, and A549 or H1299 cancer cells with scrambled siRNA into flanks of nude mice. Tumors were visible as early as five days in the mice that received scrambled siRNA-cancer cells. On 14 days post-injection, tumors appeared in all the seven mice injected with the scrambled siRNA-cancer cells. However, on 14 days post-injection, only five tumors appeared in the seven mice injected with SNHG9-siRNA-cancer cells. On 17 days post-injection, tumor appeared in another two mice injected with SNHG9-siRNA-cancer cells. Overall, tumor growth in the mice injected with SNHG9-siRNA cancer cells was significantly lower than in the mice injected with the cells transfected with scrambled siRNA (Figure 1G). The tumors generated from scrambled siRNA-cancer cells were significantly larger compared to those produced from the cancer cells with SNHG9-siRNA at the end of observation (28 days) ($174.72 \pm 96.69 \text{ mm}^3$ vs. $24.78 \pm 12.98 \text{ mm}^3$, $p = 0.03$). These in vivo findings in ectopic xenograft mouse models are consistent with the in vitro observations, and hence support that suppression of SNHG9 could inhibit in vivo tumorigenicity of lung cancer cells.

3.6. Upregulation of SNHG9 Is Associated with Advanced Stage and Survival of Lung Cancer Patients

To investigate clinical significance of SNHG9 dysregulation, we used qRT-PCR to determine expression of the gene in frozen surgically resected lung tumor tissues and matched noncancerous lung tissues of 78 patients with NSCLC (Table 2). U6 was used as an internal control gene to determine relative expression of SNHG9 in the tissue specimens. SNHG9 displayed a significantly higher expression level in lung tumor tissues compared with the corresponding noncancerous lung specimens ($p=0.013$) (Figure 2A). Both univariate and multivariate analyses showed that expression of the SNHG9, age, and stage were significantly associated with survival of the patients (all $p < 0.05$) (Tables S3 and S4). No statistically significant correlation was observed between SNHG9 expression in the tissue specimens and patient age, sex, and tumor histological type (all $p < 0.05$), except smoking history of the patients ($p < 0.05$). SNHG9 expression was positively associated with stage of lung cancer ($p < 0.05$). The 78 NSCLC patients were classified into two groups according to a median SNHG9 expression value (0.00034 ± 0.0001) in lung tumor tissues. A high SNHG9 expression in tumor tissues was a predictive of shorter overall survival time ($p < 0.01$). In addition, the Kaplan–Meier Curve indicated that the lung cancer patients with a high SNHG9 expression level had a poor survival compared with the patients who had a lower level of the SNHG9 in the tumor tissues ($p = 0.001$) (Figure 2B). Therefore, the assessment of SNHG9 might provide a potential approach for predicting outcome of NSCLC.

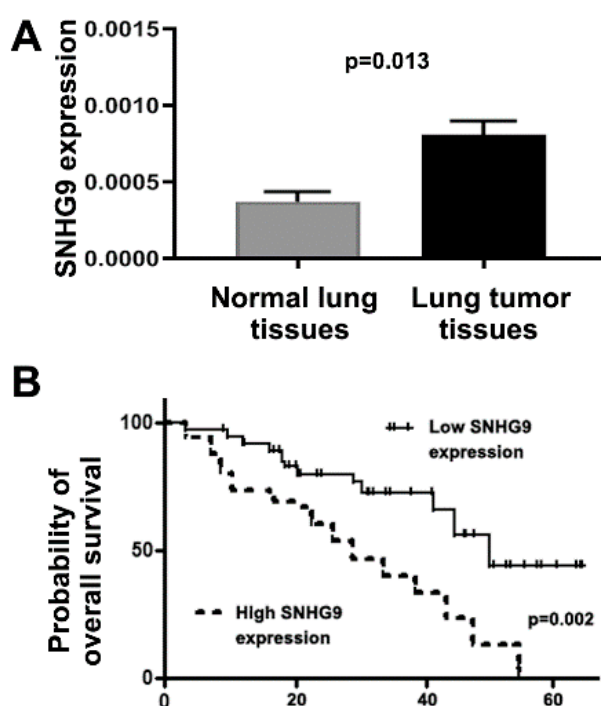


Figure 2. Expression of SNHG9 determined by using RT-PCR in surgically resected lung tumor tissues and the matched noncancerous lung tissues of 78 lung cancer patients. (A) SNHG9 has a higher expression level in lung tumor tissues compared with the corresponding noncancerous lung specimens ($p = 0.013$). (B) The 78 NSCLC patients were classified into two groups according to a median SNHG9 expression value in the lung tumor tissues (0.00034 ± 0.0001). The patients with high expression of SNHG9 in the lung tumor tissues had significantly shorter survival time than those with low expression level of the gene ($p = 0.001$).

4. Discussion

In this study, we identified a broad spectrum of ncRNA features in bronchial epitheliums of sputum from NSCLC patients, which include seven major types of ncRNAs. Of the various types of ncRNAs, miRNAs have been extensively investigated for the functions in lung cancer development

and progression. For instance, analyze epithelial brushings from the main stem-bronchus, Pavel et al. identified 42 miRNAs that showed abnormal expressions in lung cancer patients [39]. However, collecting bronchial epithelial cells from the mainstem bronchus through bronchoscopy is an invasive procedure. Furthermore, using NGS to analyze surgically resected lung tumor tissues, we identified miRNA and snoRNA changes in the tissue specimens of lung cancer patients [26,27,40]. This present study is the first to systematically and comprehensively characterize changes of ncRNAs, including miRNAs, in exfoliated bronchial epithelial cells enriched from sputum of lung cancer patients by using global NGS. The differentially expressed miRNAs in bronchial epitheliums enriched from sputum could contribute to the development of lung tumor.

Besides these miRNAs, other ncRNAs (non-miRNA ncRNAs), including 23 snRNAs, 93 snoRNAs, 13 piRNAs, 7 tRNAs, 15 rRNAs, and 25 lncRNAs, also differentially expressed in bronchial epitheliums of lung cancer patients compared with those of cancer-free smokers. The non-miRNA ncRNAs are gaining prominence and more actively involved in carcinogenesis than previously though. For example, snoRNAs serve as guides for the 2'-O-ribose methylation of rRNAs or snRNAs and their isomerization of uridine residues into pseudo uridine [41]. Our previous studies have shown that snoRA42 has oncogenic function in the development and progression of NSCLC by regulating features of lung tumor-initiating cells [30,31]. Such pleiotropy of snoRA42 dysregulation could be achieved partially through increased apoptosis of NSCLC cells in a p53-dependent manner [30]. piRNAs are Dicer-independent and can interact with the PIWI subfamily of Argonaute proteins involved in the regulation of genome stability. A high expression level of piR-651 in lung tumor tissues is associated with cancer progression in the patients with NSCLC [42]. Upregulation of piR-651 could induce NSCLC progression via the cyclin D1 and CDK4 pathway [42–44]. However, inhibition of piR-651 reduced cell proliferation and significantly increased the apoptotic rate of lung cancer cells [44]. rRNA has essential functions for protein synthesis in all living organisms [45]. Recent studies have shown that rRNA expression and modifications have an important function in cancer progression. Increased rRNA expression was associated with cancer development in prostate and cervical cancer [46]. In colorectal cancer, high expression of the pre-45S rRNA promoted G1/S cell-cycle transition and was associated with poor prognosis [47]. tRNAs have the function of transporting the amino acids to the ribosome. Some tRNAs and tRNA derivatives have been suggested to involve in proliferation, metastasis, and invasiveness of cancer cells [48]. An elevated expression level of tRNAs-Leu and tRNAs-Val was found in lung tumor tissues [49]. Methionyl-tRNA synthetase overexpression was associated with poor clinical outcomes in lung cancer [50]. snRNAs have profound effects on the cellular transcriptome and are complexed with small nuclear ribonucleoproteins [51]. Furthermore, snRNAs could recognize 5' and 3' intron/exon boundaries during splicing of introns from pre-messenger RNA transcripts. snRNAs might have important functions in cancer development and progression, as well as in drug resistance [3]. For instance, U1 snRNP could regulate cancer cell migration and invasion in vitro [52]. Nonetheless, further investigation of the multiple ncRNAs whose changes exist in bronchial epitheliums would be important to understand the mechanism of field cancerization in lung carcinogenesis.

lncRNAs become one of the focuses in the research field, since they play a central role in numerous physiological and pathological processes of a wide variety of diseases. Particularly, lncRNAs can regulate different molecular signaling pathways via changing gene expression, and, therefore, are implicated in numerous mechanisms of lung carcinogenesis. We have found that expression levels of SNHG1 and RMRP are reliably measured in plasma, and the lncRNAs may provide cell-free circulating biomarkers for lung cancer [28]. Interestingly, our identified altered lncRNAs in bronchial epitheliums of lung cancer patients include the previously characterized lung cancer-associated lncRNAs. Dysregulation of the lncRNAs in bronchial epitheliums may contribute to the field cancerization in lung carcinogenesis. The SNHG family includes numerous members, whose dysregulations are associated with a variety of malignancies [53]. Each SNHG might contribute to tumorigenesis via multiple molecular regulatory mechanisms [53]. For example, some SNHGs could act as sponges of microRNAs to inhibit the roles of miRNAs. They can also bind proteins

to influence target genes or impact tumorigenesis via different signaling pathways, including the EMT, Wnt, PIK3CA, NF- κ B, and TP53 signaling pathways [54]. Herein, we found that SNHG9 may have an important function in pathogenesis of NSCLC, since its transient knockdown contributes to NSCLC cell growth, proliferation, and invasion. In addition, the restriction of cancer cell growth and proliferation by the SNHG9 downregulation were accompanied by an accumulation of cells in the G1 phase and a decreased rate of S-phase. Moreover, the in vitro tumorigenicity of SNHG9 was confirmed by the observation in ectopic xenograft mouse models. The elevated SNHG9 expression was frequently observed in lung tumor tissues. Importantly, SNHG9 expression in tumor tissues is positively related to advanced stages of lung cancer and inversely correlated with survival of the patients. The results obtained from the clinical specimens provide further evidence that the dysregulation of SNHG9 contributes to lung cancer development and progression. The detection of SNHG9 aberrations might be a useful approach to identify NSCLC patients who have poor prognoses. Furthermore, the in vitro and in vivo data of siRNA-mediated silencing of the lncRNA would form the basis developing novel therapeutic targets for the malignancy. However, the molecular mechanisms underlying the biologic effects of SNHG9 in lung tumorigenesis remain unknown. We are performing a new study to investigate possible mechanisms or pathways, by which SNHG9 might participate in the development and progression of NSCLC. Furthermore, of the ncRNAs that displayed the highest expression level in bronchial epitheliums of sputum from lung cancer patients, only SNHG9 was further analyzed by in vitro and in vivo approaches in this present study. Our ongoing studies are investigating possible roles of other differentially expressed ncRNAs in tumorigenesis of lung cancer.

Since the molecular abnormalities in the large bronchial airways reflect those in primary lung tumors in the distal lung, molecular changes, such as aberrant ncRNAs in epithelial cells collected from the normal-appearing mainstem bronchus of smokers could be developed as biomarkers for NSCLC. As a mirror to lung diseases, sputum contains bronchial epithelial cells from the lungs and lower respiratory tract. Examination of the exfoliated bronchial epitheliums of airway in sputum might detect the lung tumor-related molecular alterations, and hence provide a non-invasive and specific means for diagnosis of NSCLC. We have demonstrated that, due to their relative resistance to nucleases, miRNAs are highly stable in sputum, providing biomarkers for lung cancer [8,9,11,12,22,23]. Given that the various types of abnormal ncRNAs in bronchial epitheliums that have highly different and active roles in lung tumorigenesis via numerous cellular pathways, developing the non-miRNA ncRNAs as sputum biomarkers and integrating them with the existing miRNAs biomarkers might have a synergistic effect for NSCLC detection.

5. Conclusions

By transcriptomic sequencing airway epithelial cells of sputum, we identified multiple differentially expressed ncRNAs in bronchial epitheliums of lung cancer patients. Further functional studies of the broad spectrum of ncRNA features would deepen our understanding of the molecular mechanisms driving lung carcinogenesis and provide potential biomarkers and therapeutic targets for the disease.

Supplementary Materials: Supplementary materials can be found at <http://www.mdpi.com/2227-9059/8/4/88/s1>.

Author Contributions: Y.L. and F.J. conducted the experiments and participated in data interpretation. V.H., P.D., J.D., and N.W.T. collected samples; F.J. participated in study design, coordination, and data analysis. Y.L. and F.J. prepared the manuscript. All authors read and approved the final manuscript.

Funding: This work was supported in part by VA Merit Award I01 CX000512, NCI 1R21CA240556, the Geaton and JoAnn DeCesaris Family Foundation, University of Maryland Marlene & Stewart Greenebaum Comprehensive Cancer Center Pilot Grant Program, and DoD-Idea Development Award (F.J.).

Acknowledgments: We thank the Institute for Genome Sciences of the University of Maryland School of Medicine to provide sequencing and analysis services.

Conflicts of Interest: The authors declare no competing interests

References

1. American Cancer Society. *Cancer Facts & Figures*; American Cancer Society: Atlanta, GA, USA, 2020.
2. Tang, X.; Shigematsu, H.; Bekele, B.N.; Roth, J.A.; Minna, J.D.; Hong, W.K.; Gazdar, A.F.; Wistuba, I.I. EGFR tyrosine kinase domain mutations are detected in histologically normal respiratory epithelium in lung cancer patients. *Cancer Res.* **2005**, *65*, 7568–7572. [\[CrossRef\]](#)
3. Romano, G.; Veneziano, D.; Acunzo, M.; Croce, C.M. Small non-coding RNA and cancer. *Carcinogenesis* **2017**, *38*, 485–491. [\[CrossRef\]](#)
4. Sozzi, G.; Pastorino, U.; Croce, C.M. MicroRNAs and lung cancer: From markers to targets. *Cell Cycle* **2011**, *10*, 2045–2046. [\[CrossRef\]](#)
5. Irimie, A.I.; Zimta, A.A.; Ciocan, C.; Mehterov, N.; Dudea, D.; Braicu, C.; Berindan-Neagoe, I. The Unforeseen Non-Coding RNAs in Head and Neck Cancer. *Genes* **2018**, *9*, 134. [\[CrossRef\]](#)
6. Pang, W.; Tian, X.; Bai, F.; Han, R.; Wang, J.; Shen, H.; Zhang, X.; Liu, Y.; Yan, X.; Jiang, F.; et al. Pim-1 kinase is a target of miR-486-5p and eukaryotic translation initiation factor 4E, and plays a critical role in lung cancer. *Mol. Cancer* **2014**, *13*, 240. [\[CrossRef\]](#)
7. Wang, J.; Tian, X.; Han, R.; Zhang, X.; Wang, X.; Shen, H.; Xue, L.; Liu, Y.; Yan, X.; Shen, J.; et al. Downregulation of miR-486-5p contributes to tumor progression and metastasis by targeting protumorigenic ARHGAP5 in lung cancer. *Oncogene* **2014**, *33*, 1181–1189. [\[CrossRef\]](#)
8. Shen, J.; Liao, J.; Guarnera, M.A.; Fang, H.; Cai, L.; Stass, S.A.; Jiang, F. Analysis of MicroRNAs in sputum to improve computed tomography for lung cancer diagnosis. *J. Thorac. Oncol.* **2014**, *9*, 33–40. [\[CrossRef\]](#)
9. Xie, Y.; Todd, N.W.; Liu, Z.; Zhan, M.; Fang, H.; Peng, H.; Alattar, M.; Deepak, J.; Stass, S.A.; Jiang, F. Altered miRNA expression in sputum for diagnosis of non-small cell lung cancer. *Lung Cancer* **2010**, *67*, 170–176. [\[CrossRef\]](#)
10. Xing, L.; Todd, N.W.; Yu, L.; Fang, H.; Jiang, F. Early detection of squamous cell lung cancer in sputum by a panel of microRNA markers. *Mod. Pathol.* **2010**, *23*, 1157–1164. [\[CrossRef\]](#)
11. Yu, L.; Todd, N.W.; Xing, L.; Xie, Y.; Zhang, H.; Liu, Z.; Fang, H.; Zhang, J.; Katz, R.L.; Jiang, F. Early detection of lung adenocarcinoma in sputum by a panel of microRNA markers. *Int. J. Cancer* **2010**, *127*, 2870–2878. [\[CrossRef\]](#)
12. Xing, L.; Su, J.; Guarnera, M.A.; Zhang, H.; Cai, L.; Zhou, R.; Stass, S.A.; Jiang, F. Sputum microRNA biomarkers for identifying lung cancer in indeterminate solitary pulmonary nodules. *Clin. Cancer Res.* **2015**, *21*, 484–489. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Su, J.; Liao, J.; Gao, L.; Shen, J.; Guarnera, M.A.; Zhan, M.; Fang, H.; Stass, S.A.; Jiang, F. Analysis of small nucleolar RNAs in sputum for lung cancer diagnosis. *Oncotarget* **2016**, *7*, 5131–5142. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Su, Y.; Guarnera, M.A.; Fang, H.; Jiang, F. Small non-coding RNA biomarkers in sputum for lung cancer diagnosis. *Mol. Cancer* **2016**, *15*, 36. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Yu, L.; Shen, J.; Mannoor, K.; Guarnera, M.; Jiang, F. Identification of ENO1 As a Potential Sputum Biomarker for Early-Stage Lung Cancer by Shotgun Proteomics. *Clin. Lung Cancer* **2014**, *15*, 372–378. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Li, N.; Ma, J.; Guarnera, M.A.; Fang, H.; Cai, L.; Jiang, F. Digital PCR quantification of miRNAs in sputum for diagnosis of lung cancer. *J. Cancer Res. Clin. Oncol.* **2014**, *140*, 145–150. [\[CrossRef\]](#)
17. Anjuman, N.; Li, N.; Guarnera, M.; Stass, S.A.; Jiang, F. Evaluation of lung flut in sputum samples for molecular analysis of lung cancer. *Clin. Transl. Med.* **2013**, *2*, 15. [\[CrossRef\]](#)
18. Jiang, F.; Todd, N.W.; Li, R.; Zhang, H.; Fang, H.; Stass, S.A. A panel of sputum-based genomic marker for early detection of lung cancer. *Cancer Prev. Res. (Phila)* **2010**, *3*, 1571–1578. [\[CrossRef\]](#)
19. Jiang, F.; Todd, N.W.; Qiu, Q.; Liu, Z.; Katz, R.L.; Stass, S.A. Combined genetic analysis of sputum and computed tomography for noninvasive diagnosis of non-small-cell lung cancer. *Lung Cancer* **2009**, *66*, 58–63. [\[CrossRef\]](#)
20. Qiu, Q.; Todd, N.W.; Li, R.; Peng, H.; Liu, Z.; Yfantis, H.G.; Katz, R.L.; Stass, S.A.; Jiang, F. Magnetic enrichment of bronchial epithelial cells from sputum for lung cancer diagnosis. *Cancer* **2008**, *114*, 275–283. [\[CrossRef\]](#)
21. Li, R.; Todd, N.W.; Qiu, Q.; Fan, T.; Zhao, R.Y.; Rodgers, W.H.; Fang, H.-B.; Katz, R.L.; Stass, S.A.; Jiang, F. Genetic deletions in sputum as diagnostic markers for early detection of stage I non-small cell lung cancer. *Clin. Cancer Res.* **2007**, *13 Pt 1*, 482–487. [\[CrossRef\]](#)

22. Shen, J.; Stass, S.A.; Jiang, F. MicroRNAs as potential biomarkers in human solid tumors. *Cancer Lett.* **2013**, *329*, 125–136. [[CrossRef](#)] [[PubMed](#)]
23. Shen, J.; Jiang, F. Applications of MicroRNAs in the Diagnosis and Prognosis of Lung Cancer. *Expert Opin. Med. Diagn.* **2012**, *6*, 197–207. [[CrossRef](#)] [[PubMed](#)]
24. Shen, J.; Liu, Z.; Todd, N.W.; Zhang, H.; Liao, J.; Yu, L.; Guarnera, M.; Li, R.; Cai, L.; Zhan, M.; et al. Diagnosis of lung cancer in individuals with solitary pulmonary nodules by plasma microRNA biomarkers. *BMC Cancer* **2011**, *11*, 374. [[CrossRef](#)] [[PubMed](#)]
25. Shen, J.; Todd, N.W.; Zhang, H.; Yu, L.; Lingxiao, X.; Mei, Y.; Guarnera, M.; Liao, J.; Chou, A.; Lu, C.L.; et al. Plasma microRNAs as potential biomarkers for non-small-cell lung cancer. *Lab. Investig.* **2011**, *91*, 579–587. [[CrossRef](#)] [[PubMed](#)]
26. Ma, J.; Mannoor, K.; Gao, L.; Tan, A.; Guarnera, M.A.; Zhan, M.; Shetty, A.; Stass, S.A.; Xing, L.; Jiang, F. Characterization of microRNA transcriptome in lung cancer by next-generation deep sequencing. *Mol. Oncol.* **2014**, *8*, 1208–1219. [[CrossRef](#)] [[PubMed](#)]
27. Gao, L.; Ma, J.; Mannoor, K.; Guarnera, M.A.; Shetty, A.C.; Zhan, M.; Xing, L.; Stass, S.A.; Jiang, F. Genome-wide small nucleolar RNA expression analysis of lung cancer by next-generation deep sequencing. *Int. J. Cancer* **2015**, *136*, E623–E629. [[CrossRef](#)]
28. Lin, Y.; Leng, Q.; Zhan, M.; Jiang, F. A Plasma Long Noncoding RNA Signature for Early Detection of Lung Cancer. *Transl. Oncol.* **2018**, *11*, 1225–1231. [[CrossRef](#)]
29. Leng, Q.; Lin, Y.; Zhan, M.; Jiang, F. An integromic signature for lung cancer early detection. *Oncotarget* **2018**, *9*, 24684–24692. [[CrossRef](#)]
30. Mei, Y.-P.; Liao, J.-P.; Shen, J.; Yu, L.; Liu, B.; Liu, L.; Li, R.-Y.; Ji, L.; Dorsey, S.G.; Jiang, Z.-R.; et al. Small nucleolar RNA 42 acts as an oncogene in lung tumorigenesis. *Oncogene* **2012**, *31*, 2794–2804. [[CrossRef](#)]
31. Mannoor, K.; Shen, J.; Liao, J.; Liu, Z.; Jiang, F. Small nucleolar RNA signatures of lung tumor-initiating cells. *Mol. Cancer* **2014**, *13*, 104. [[CrossRef](#)]
32. Riss, T.L.; Moravec, R.A.; Niles, A.L.; Duellman, S.; Benink, H.A.; Worzella, T.J.; Minor, L. *Cell Viability Assays*; Eli Lilly & Company and the National Center for Advancing Translational Sciences: San Antonio, TX, USA, 2004.
33. Liu, Z.; Jiang, F. Gene identification and survival prediction with Lp Cox regression and novel similarity measure. *Int. J. Data Min. Bioinform.* **2009**, *3*, 398–408. [[CrossRef](#)] [[PubMed](#)]
34. Liu, Z.; Tan, M.; Jiang, F. Regularized F-measure maximization for feature selection and classification. *J. Biomed. Biotechnol.* **2009**, *2009*, 617946. [[CrossRef](#)] [[PubMed](#)]
35. Liu, Z.; Gartenhaus, R.B.; Tan, M.; Jiang, F.; Jiao, X. Gene and pathway identification with Lp penalized Bayesian logistic regression. *BMC Bioinform.* **2008**, *9*, 412. [[CrossRef](#)] [[PubMed](#)]
36. Liu, Z.; Jiang, F.; Tian, G.; Wang, S.; Sato, F.; Meltzer, S.J.; Tan, M. Sparse logistic regression with Lp penalty for biomarker identification. *Stat. Appl. Genet. Mol. Biol.* **2007**, *6*, 6. [[CrossRef](#)]
37. Liu, P.; Hwang, J.T. Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. *Bioinformatics* **2007**, *23*, 739–746. [[CrossRef](#)] [[PubMed](#)]
38. Tian, F.; Wang, J.; Ouyang, T.; Lu, N.; Lu, J.; Shen, Y.; Bai, Y.; Xie, X.; Ge, Q.-Y. MiR-486-5p Serves as a Good Biomarker in Nonsmall Cell Lung Cancer and Suppresses Cell Growth With the Involvement of a Target PIK3R1. *Front. Genet.* **2019**, *10*, 688. [[CrossRef](#)]
39. Pavel, A.B.; Campbell, J.D.; Liu, G.; Elashoff, D.; Dubinett, S.; Smith, K.; Whitney, D.; Lenburg, M.; Spira, A.; AEGIS Study Team. Alterations in Bronchial Airway miRNA Expression for Lung Cancer Detection. *Cancer Prev. Res. (Phila)* **2017**, *10*, 651–659. [[CrossRef](#)]
40. Xu, H.; Gao, L.; Che, T.; Du, W.; Li, Q.; Gao, Q.; Wang, Z. The effects of 12C6+ irradiation on cell cycle, apoptosis, and expression of caspase-3 in the human lung cancer cell line h1299. *Cancer Biother. Radiopharm.* **2012**, *27*, 113–118. [[CrossRef](#)]
41. Weinstein, L.B.; Steitz, J.A. Guided tours: From precursor snoRNA to functional snoRNP. *Curr. Opin. Cell Biol.* **1999**, *11*, 378–384. [[CrossRef](#)]
42. Li, D.; Luo, Y.; Gao, Y.; Yang, Y.; Wang, Y.; Xu, Y.; Tan, S.; Zhang, Y.; Duan, J.; Yang, Y. piR-651 promotes tumor formation in non-small cell lung carcinoma through the upregulation of cyclin D1 and CDK4. *Int. J. Mol. Med.* **2016**, *38*, 927–936. [[CrossRef](#)]
43. Yao, J.; Wang, Y.W.; Fang, B.B.; Zhang, S.J.; Cheng, B.L. piR-651 and its function in 95-D lung cancer cells. *Biomed. Rep.* **2016**, *4*, 546–550. [[CrossRef](#)] [[PubMed](#)]

44. Zhang, S.; Yao, J.; Shen, B.; Li, G.; Kong, S.; Bi, D.; Pan, S.; Cheng, B. Role of piwi-interacting RNA-651 in the carcinogenesis of non-small cell lung cancer. *Oncol. Lett.* **2018**, *15*, 940–946. [[CrossRef](#)] [[PubMed](#)]
45. Lafontaine, D.L.; Tollervey, D. The function and synthesis of ribosomes. *Nat. Rev. Mol. Cell Biol.* **2001**, *2*, 514–520. [[CrossRef](#)] [[PubMed](#)]
46. Zhou, H.; Wang, Y.; Lv, Q.; Zhang, J.; Wang, Q.; Gao, F.; Hou, H.; Zhang, H.; Zhang, W.; Li, L. Overexpression of Ribosomal RNA in the Development of Human Cervical Cancer Is Associated with rDNA Promoter Hypomethylation. *PLoS ONE* **2016**, *11*, e0163340. [[CrossRef](#)]
47. Tsoi, H.; Lam, K.C.; Dong, Y.; Zhang, X.; Lee, C.K.; Zhang, J.; Ng, S.C.; Ng, S.S.M.; Zheng, S.; Chen, Y.; et al. Pre-45s rRNA promotes colon cancer and is associated with poor survival of CRC patients. *Oncogene* **2017**, *36*, 6109–6118. [[CrossRef](#)]
48. Randerath, K.; Agrawal, H.P.; Randerath, E. tRNA alterations in cancer. *Recent Results Cancer Res.* **1983**, *84*, 103–120.
49. Shao, Y.; Sun, Q.; Liu, X.; Wang, P.; Wu, R.; Ma, Z. tRF-Leu-CAG promotes cell proliferation and cell cycle in non-small cell lung cancer. *Chem. Biol. Drug Des.* **2017**, *90*, 730–738. [[CrossRef](#)]
50. Kim, E.Y.; Jung, J.Y.; Kim, A.; Kim, K.; Chang, Y.S. Methionyl-tRNA synthetase overexpression is associated with poor clinical outcomes in non-small cell lung cancer. *BMC Cancer* **2017**, *17*, 467. [[CrossRef](#)]
51. Hoskins, A.A.; Moore, M.J. The spliceosome: A flexible, reversible macromolecular machine. *Trends Biochem. Sci.* **2012**, *37*, 179–188. [[CrossRef](#)]
52. Oh, J.M.; Venters, C.C.; Di, C.; Pinto, A.M.; Wan, L.; Younis, I.; Cai, Z.; Arai, C.; So, B.R.; Duan, J.; et al. U1 snRNP regulates cancer cell migration and invasion in vitro. *Nat. Commun.* **2020**, *11*, 1–8. [[CrossRef](#)]
53. Yang, H.; Jiang, Z.; Wang, S.; Zhao, Y.; Song, X.; Xiao, Y.; Yang, S. Long non-coding small nucleolar RNA host genes in digestive cancers. *Cancer Med.* **2019**, *8*, 7693–7704. [[CrossRef](#)] [[PubMed](#)]
54. Shen, Y.; Liu, S.; Fan, J.; Jin, Y.-H.; Tian, B.; Zheng, X.; Fu, H. Nuclear retention of the lncRNA SNHG1 by doxorubicin attenuates hnRNPC-p53 protein interactions. *EMBO Rep.* **2017**, *18*, 536–548. [[CrossRef](#)] [[PubMed](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

METHODOLOGY ARTICLE

Open Access



Cell-based reference samples designed with specific differences in microRNA biomarkers

P. Scott Pine^{1*}, Steven P. Lund^{2†}, Sanford A. Stass³, Debra Kukuruga³, Feng Jiang³, Lynn Sorbara⁴, Sudhir Srivastava⁴ and Marc Salit¹

Abstract

Background: We demonstrate the feasibility of creating a pair of reference samples to be used as surrogates for clinical samples measured in either a research or clinical laboratory setting. The reference sample paradigm presented and evaluated here is designed to assess the capability of a measurement process to detect true differences between two biological samples. Cell-based reference samples can be created with a biomarker signature pattern designed in silico. Clinical laboratories working in regulated applications are required to participate in proficiency testing programs; research laboratories doing discovery typically do not. These reference samples can be used in proficiency tests or as process controls that allow a laboratory to evaluate and optimize its measurement systems, monitor performance over time (process drift), assess changes in protocols, reagents, and/or personnel, maintain standard operating procedures, and most importantly, provide evidence for quality results.

Results: The biomarkers of interest in this study are microRNAs (miRNAs), small non-coding RNAs involved in the regulation of gene expression. Multiple lung cancer associated cell lines were determined by reverse transcription (RT)-PCR to have sufficiently different miRNA profiles to serve as components in mixture designs as reference samples. In silico models based on the component profiles were used to predict miRNA abundance ratios between two different cell line mixtures, providing target values for profiles obtained from in vitro mixtures. Two reference sample types were tested: total RNA mixed after extraction from cell lines, and intact cells mixed prior to RNA extraction. MicroRNA profiling of a pair of samples composed of extracted RNA derived from these cell lines successfully replicated the target values. Mixtures of intact cells from these lines also approximated the target values, demonstrating potential utility as mimics for clinical specimens. Both designs demonstrated their utility as reference samples for inter- or intra-laboratory testing.

Conclusions: Cell-based reference samples can be created for performance assessment of a measurement process from biomolecule extraction through quantitation. Although this study focused on miRNA profiling with RT-PCR using cell lines associated with lung cancer, the paradigm demonstrated here should be extendable to genome-scale platforms and other biomolecular endpoints.

Keywords: Reverse transcription PCR (RT-PCR), microRNA (miRNA), Reference samples, Biomarkers, Performance assessment, Measurement assurance

* Correspondence: p.scott.pine@nist.gov

†Equal contributors

¹Joint Initiative for Metrology in Biology, National Institute of Standards and Technology, 443 Via Ortega, Stanford, CA 94305, USA

Full list of author information is available at the end of the article



Background

The translation of a biomarker from discovery to clinical practice involves a transition from the research-oriented approach of the discovery laboratory to a procedure in an accredited clinical laboratory. However, most published cancer biomarkers never make it into clinical practice [1], with recent work highlighting the irreproducibility of results [2]. Demonstrating that methodology is accurately transferred is critical to the biomarker validation process. This process can be conceptualized by multiple phases of biomarker development [3]. Within the framework of the Early Detection Research Network (EDRN) of the National Cancer Institute [4], we are developing a measurement assurance paradigm for the first phase of this process, *preclinical exploratory studies*. Quality measurements during this phase are essential for the success of all subsequent phases of biomarker development.

Preclinical studies to develop cancer biomarkers typically involve comparing measurement results from tumor tissue to normal tissue and identifying features that distinguish the two classes. Reference samples, with known differences, enable laboratories to demonstrate their ability to detect potential biomarkers. Ideally, these biomarkers would be measured in one or more biofluids and provide a means for non-invasive screening at the earliest stages for the detection of cancer. Two EDRN laboratories are assessing microRNAs (miRNAs) found in sputum as potential biomarkers, which is the context for designing mixture reference samples described in the present paper.

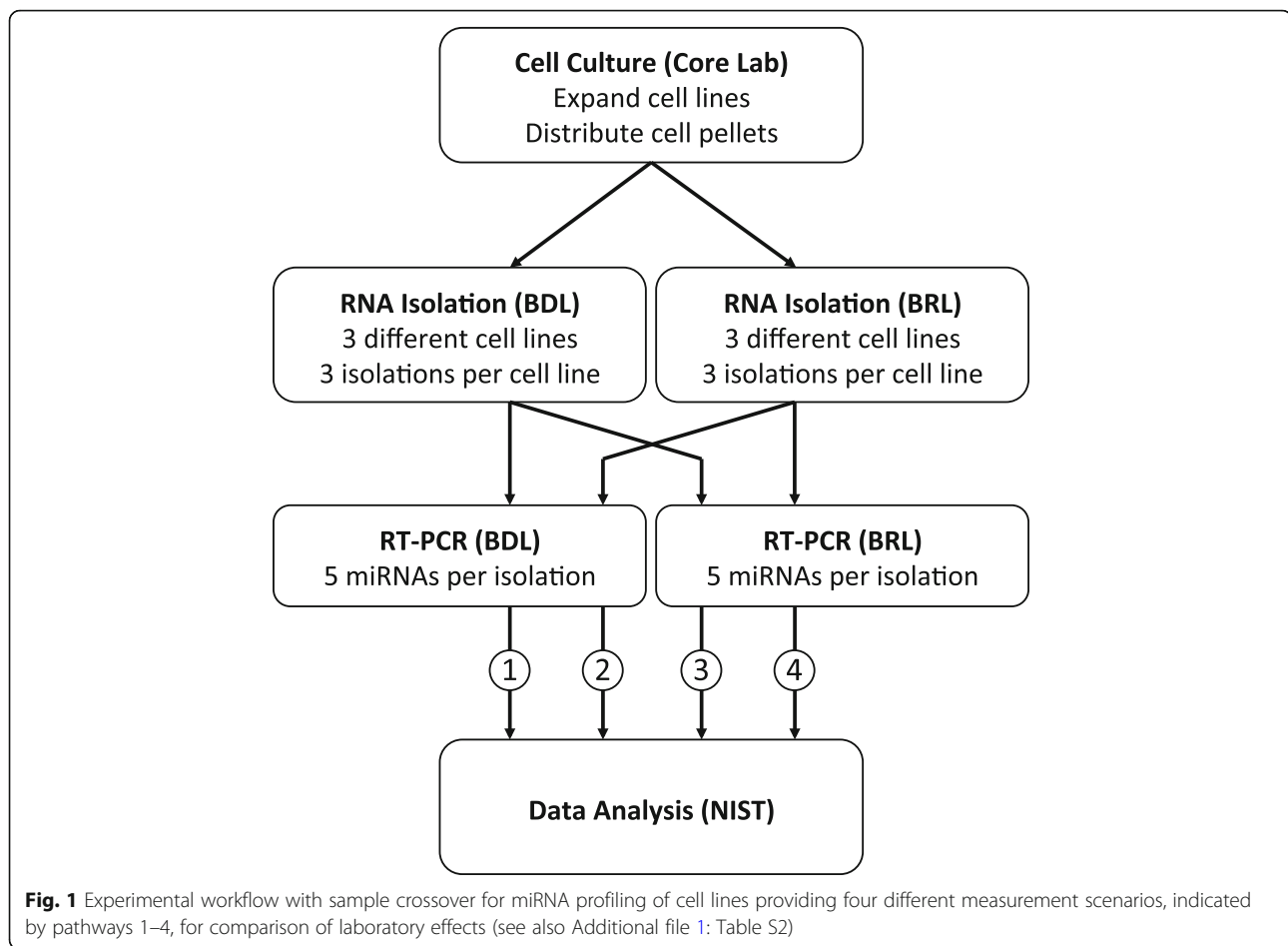
MicroRNAs are a class of small non-protein-coding RNAs that regulate the expression of hundreds of target genes, controlling biological functions involved in differentiation and development [5]. Dysregulation of miRNA expression may contribute to cancer development and progression [6]. Therefore, miRNAs are potential candidates as useful biomarkers.

Three cell lines with relatively different expression profiles for five miRNAs of interest were assessed as potential mixture components to create a pair of reference samples to act as surrogates for sputum samples measured in a non-invasive screening assay for lung cancer. Lung cancer cell lines from the NCI-H series (H226, H358, and H460) [7] were profiled for the expression of miR-21 [8], miR-126 [9], miR-210 [10], miR-375 [11], and miR-486 [12]. These miRNA were selected based upon their putative roles in cancer and previous screening results of the BDL that found that three of these miRNA (miR-21, miR-210, and miR-375) were *overexpressed* and two (miR-126 and miR-486) were *underexpressed* when comparing lung cancer and normal tissue, providing a potential panel of biomarkers for early detection [13, 14].

The miRNA profile of each reference sample need not mimic a particular biological state (i.e., normal or disease), but should provide the ability to assess technical performance of the measurement process used to discriminate based on the particular set of miRNAs of interest. Ideally, these reference samples should possess the following properties: 1) express all or most of the biomarkers of interest; 2) be regenerable for use over time as a process control; 3) be amenable to identical processing as clinical samples; 4) contain well-established relative differences in the biomarkers of interest; and 5) be commutable between measurement systems. Such samples could be used to assess repeatability within, and reproducibility between, laboratories for the entire measurement process.

In the first part of this study, a “crossover” design was used to assess reproducibility of miRNA profiles and parse out factors contributing to variability (see Fig. 1). Two different EDRN sites participated: one a biomarker development laboratory (BDL) engaged in clinical research; and the other a Clinical Laboratory Improvement Amendments (CLIA) compliant and College of American Pathology (CAP) accredited biomarker reference laboratory (BRL). Each laboratory used the same source of cells, which were grown in the core cell facility of the BRL. A standard operating procedure for RNA isolation was developed by the BRL for both laboratories, meeting the CLIA standards of the BRL. Each site measured the same set of miRNAs of interest using reverse transcription PCR (RT-PCR). Both sites measured these miRNAs in the total RNA samples they isolated, as well as those isolated at the other site, on three separate occasions. Datasets from each site were subsequently analyzed at the National Institute for Standards and Technology (NIST). This design permits evaluation of interlaboratory variability due to either RNA extraction or miRNA measurement.

To inform the design for the two reference samples, the miRNA measurements for each cell line were used to predict the ratio of miRNA abundance between samples formed by mixing the component cell lines in different proportions [15, 16]. To determine whether the mixture design was fit-for-purpose, the predicted ratios were compared with corresponding measurements from reference mixture samples prepared by two different methods: mixing total RNA extracted from the cell lines in the BRL, and mixing the cell lines before total RNA isolation by the BRL (see Fig. 2). The first approach provides RNA samples suitable for assessing each lab's ability to measure differential expression using their RT-PCR process. The second design provides reference samples suitable for assessing the entire measurement process of a lab.



Methods

Cell culture

Cell Lines were obtained from the American Type Culture Collection (ATCC®), Manassas, VA. H226 (ATCC# CRL-5826), H358 (ATCC# CRL-5807), and H460 (ATCC# HTB-177) were the cell lines selected by the biomarker discovery laboratory (BDL) [13, 14]. The core cell culture facility for the biomarker reference laboratory (BRL) expanded the cells and provided frozen cell pellets in RNAlater (Life Technologies, Grand Island, NY). Briefly, cells were thawed fresh at the lowest possible passage (< 25 passages) for each new batch, using the standard thawing method in 37 °C and dilution with culture medium into a new culture flask. The day after thawing, medium was replaced to completely remove residual DMSO. Cells did not exceed 80 to 85% confluence during cell expansion. T-225 flasks were seeded with 2.5 to 5.0 million cells, dependent on each cell line's growth characteristics. Cells were harvested using the standard PBS wash and trypsinization procedure. Recovered cells were counted 3X with an automated cell counter. For the initial profiling phase, a sufficient number of cells were expanded to provide for three separate isolations per lab. Separate expansions were

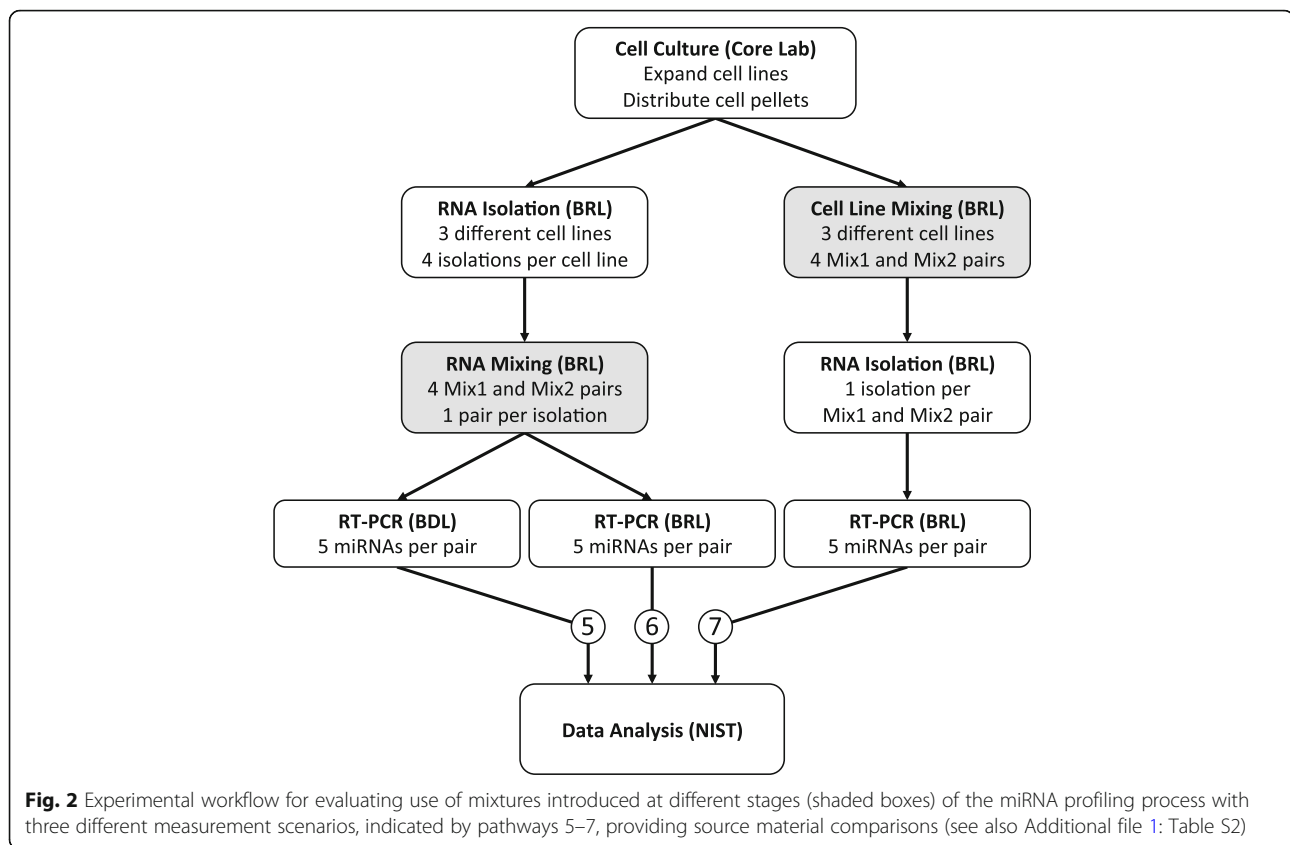
performed to provide cells for the RNA mixture experiments and the cell line mixture experiments.

RNA isolation

The BRL developed the following standard operating procedures for the isolation of total RNA enriched for miRNA from the cell pellets using the Qiagen miRNeasy Mini Kit. Briefly, cells are lysed in a denaturing buffer containing phenol / guanidine-thiocyanate designed to inhibit RNases to ensure the purification of intact RNA from cellular DNA and proteins. Samples are homogenized using the QIAshredder columns. Ethanol is then added to the sample to precipitate the RNA, and the sample is applied to an RNeasy mini spin column, where the RNA binds to the silica-based membrane. After washing contaminants away, RNA is eluted from the membrane in RNase-free water. Each lab performed three separate RNA isolations on different days on each component cell line.

RNA mixtures

For total RNA mixtures, a 3-component design with a 1:1 component and two reciprocal 3:1 components were



used. Total RNA isolated by the BRL was quantified using an Agilent BioAnalyzer and mixed by mass in the following proportions: Mix1 (1:3:1) and Mix2 (3:1:1), using H226, H358, and H460, respectively. In silico modeling was used to design the mixtures, selecting the cell line mixing proportion to use for each component that provided a useful distribution of ratios across the dynamic range (see [Results](#)).

Cell line mixtures

For cell line mixtures, intact cells were collected during log phase growth and counted in triplicate with an automated cell counter. Cells were then mixed (count:count:count) in the same proportions as the RNA mixtures: Mix1 (1:3:1) and Mix2 (3:1:1), using H226, H358, and H460, respectively.

RT-PCR measurement

For miRNA profiling of cell lines, each lab measured the RNA isolated within their lab as well as the RNA isolated by the other lab using RT-PCR. Human (Hsa-) specific kits (Applied Biosystems, Foster City, CA) for miR-16 (#000391), miR-21 (#000397), miR-126 (#002228), miR-210 (#000512), miR-486 (#001278), miR-375 (#000564) were used for reversed transcription to cDNA. The TaqMan miRNA assays were performed

using an ABI 7900 HT (Applied Biosystems, Foster City, CA) in the BRL and an IQ5 Multicolor Real-Time PCR Detection System (Biorad Laboratories, Hercules, CA) in the BDL. The same RT-PCR protocol was used for miRNA profiling of mixtures from RNA components or cell line components.

In silico modeling

Each analyte (among miR-21, miR-126, miR-210, miR-375, and miR-486) was modeled independently according to the following analysis. From the individual cell lines, let Y_{ceip} denote the average of three Cq measurements from within a single plate for RNA extracted from cell line c ($c = 1, 2, 3$; 1 = H226, 2 = H358, 3 = H460) during extraction replicate e ($e = 1, 2, 3$) at isolation lab i ($i = 1, 2$; 1 = BDL, 2 = BRL) and measured at PCR lab p ($p = 1, 2$; 1 = BDL, 2 = BRL). Under the assumption that miRNA abundance should be additive and linear, we can predict a Cq value for a hypothetical sample reflecting a designed mixture of the component cell line samples using the following equation:

$$\hat{M}_{meip} = -\log_2 \sum_{c=1}^3 2^{-Y_{ceip}} * \phi_{cm} \quad (1)$$

where $\hat{M}_{meip}$ is the predicted Cq measurement at PCR lab p of designed mixture m ($m = 1, 2$) hypothetically

created by combining the samples from isolation replicate e at isolation lab i at relative concentrations of ϕ_{1m} , ϕ_{2m} and ϕ_{3m} for component cell lines H226, H358 and H460, respectively. We refer to each of the four combinations of isolation lab and PCR lab as a different measurement process.

Statistical modeling

Data collected throughout the described experiments were analyzed in a Bayesian framework to evaluate the uncertainty related to average Cq values or their differences (e.g., average Δ Cq values). For each target miRNA in turn, the observed Cq values were analyzed using mixed effects models that include a fixed mean for each of 18 distinct combinations (see Additional file 1: Table S1) of source material, measurement method and target, as well as random effects for plate and isolation. The modeled average Cq values were used to evaluate and compare average Δ Cq values (Mix1–Mix2) for seven different measurement scenarios (see Additional file 1: Table S2).

Modeling results are presented in terms of marginal likelihoods, rather than posterior probabilities, to reduce the influence of subjective choices for prior distributions. A marginal likelihood reflects the chance of the observed data occurring assuming a particular value for a model parameter of interest (e.g., average Cq, average Δ Cq, or difference in average Δ Cq), after integrating across the modeled distribution of all other model parameters. We examine peak shapes formed by plotting marginal likelihood as a function of the value of the parameter of interest to assess whether or not particular factors appear to contribute to bias in these results. This approach is used to assess differences in Cq responses between samples (Mix1 vs. Mix2), laboratories (BDL vs. BRL), or source material (in silico vs. RNA vs. cells). In these plots, a value of 0 along the x-axis indicates “no difference” for the comparison while deviation from zero indicates a possible bias.

To examine the sensitivity of the reported results to modeling choices, the entire analysis was repeated using three model perturbations. All models were fit using Markov Chain Monte Carlo (MCMC) via the R package *rjags* [17, 18]. MCMC is a computational tool commonly used in Bayesian analysis that relies on simulation over many iterations to approximate closed-form computations that are difficult to evaluate because they involve complicated probability distributions. An introduction to Bayesian analysis in the context of the presented experiment and additional details of the three considered models are provided in the Additional file 1. The interested reader may also find a vast collection of more general references for Bayesian theory and practice, including the use of MCMC; see for instance [19].

Datasets and R code used to conduct the analysis are provided in Additional files 2, 3, 4 and 5.

Results

Relative miRNA expression levels in different cell lines

The expression profiles of five miRNAs of interest (miR-21, miR-126, miR-210, miR-375, and miR-486) [13, 14] and one control (miR-16) across three different cell lines were measured by both the BRL and BDL. Each laboratory performed three separate RNA isolations per cell line, split their samples and shared them with the other site. Each laboratory then used their own RT-PCR methods to measure the miRNA profiles of both sets of samples. This cross-over approach was used to determine if robust differences in miRNA abundance were consistently observed among the four possible measurement processes resulting from all combinations of isolation laboratory and PCR laboratory (see Fig. 1). Throughout this paper, the term Cq value refers to the average Cq (taken over 3 replicate wells) for a given miRNA in a given cell line, evaluated for each isolation and each process.

An analysis of variance (ANOVA) was performed on the Cq values in phase 1 (RNA extracted from pure cell lines). Fig. 3 shows the factors with the greatest mean squares (i.e., sum of squares divided by corresponding degrees of freedom). As expected, the biologically relevant factors of analyte, cell line and their interaction are the three most prominent components of variability. The confounding factor with the greatest attributed variability in Cq values was which lab conducted PCR analysis.

When evaluating log abundance ratios (i.e., differences in Cq values) across cell lines (or cell line mixtures), only the interactions involving cell line will matter. This is because effects that remain constant across cell lines will cancel when taking the difference (e.g., increasing the Cq value by 0.5 for all PCR measurements conducted at BRL will not impact the estimated log abundance ratio of a given miR between any two cell lines). From this perspective, the most substantial source of confounding variability is the factor associated with the interaction among analyte, cell line and PCR lab, which contributes, on average, about an order of magnitude less variability than do any of the biologically-relevant factors.

Figure 4a shows there was concordance among the four processes with respect to which of the miRNAs of interest were more abundant (lower Cq) in each cell line: H358 cells had the highest expression for miR-21 and miR-210; H226 cells had the highest expression of miR-126; and H460 cells had the highest expression of miR-375. Both H358 and H460 expressed similar levels of miR-486. MiR-21 also appeared more abundant in samples isolated by the BRL

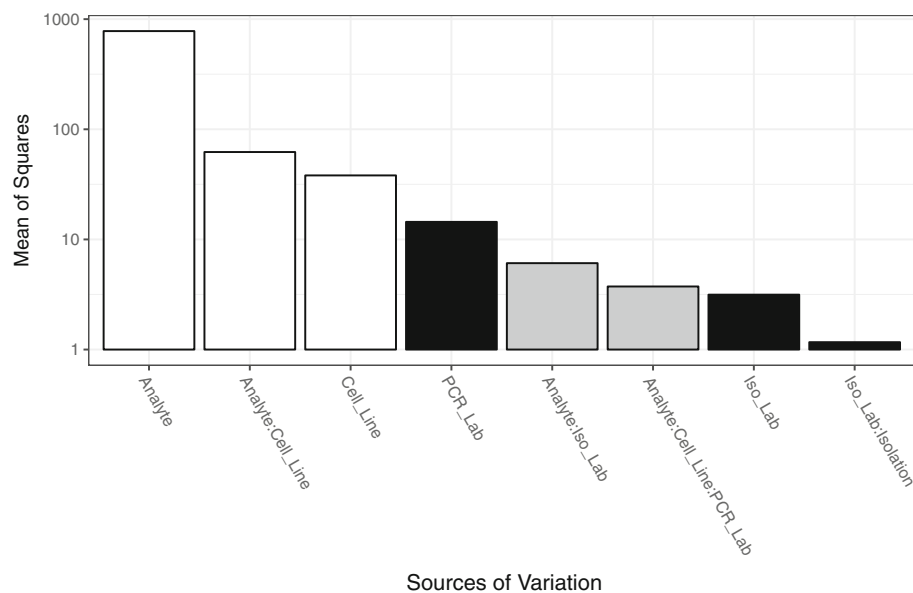


Fig. 3 Sources of variance in the miRNA profiling dataset. An analysis of variance in R [17] was used to identify the major sources of variability in the miRNA profiling of three cell lines by two laboratories using a cross-over design (see Fig. 1). A five-way ANOVA, testing all possible interactions, was applied to determine the contribution to variance by analyte, cell line, isolation laboratory, isolation process, and PCR laboratory. The input data for the model was the Cq for each of the four measurement processes for three isolations. The mean of squares is plotted on the y-axis on a log10 scale and is a measure of the average contribution of each factor or interaction to the variability in the experiment. Factors or interactions with a mean of squares less than one are not included. White bars correspond to biological factors or interactions, black bars correspond to laboratory factors or interactions, and grey bars correspond to interactions between biological and laboratory factors

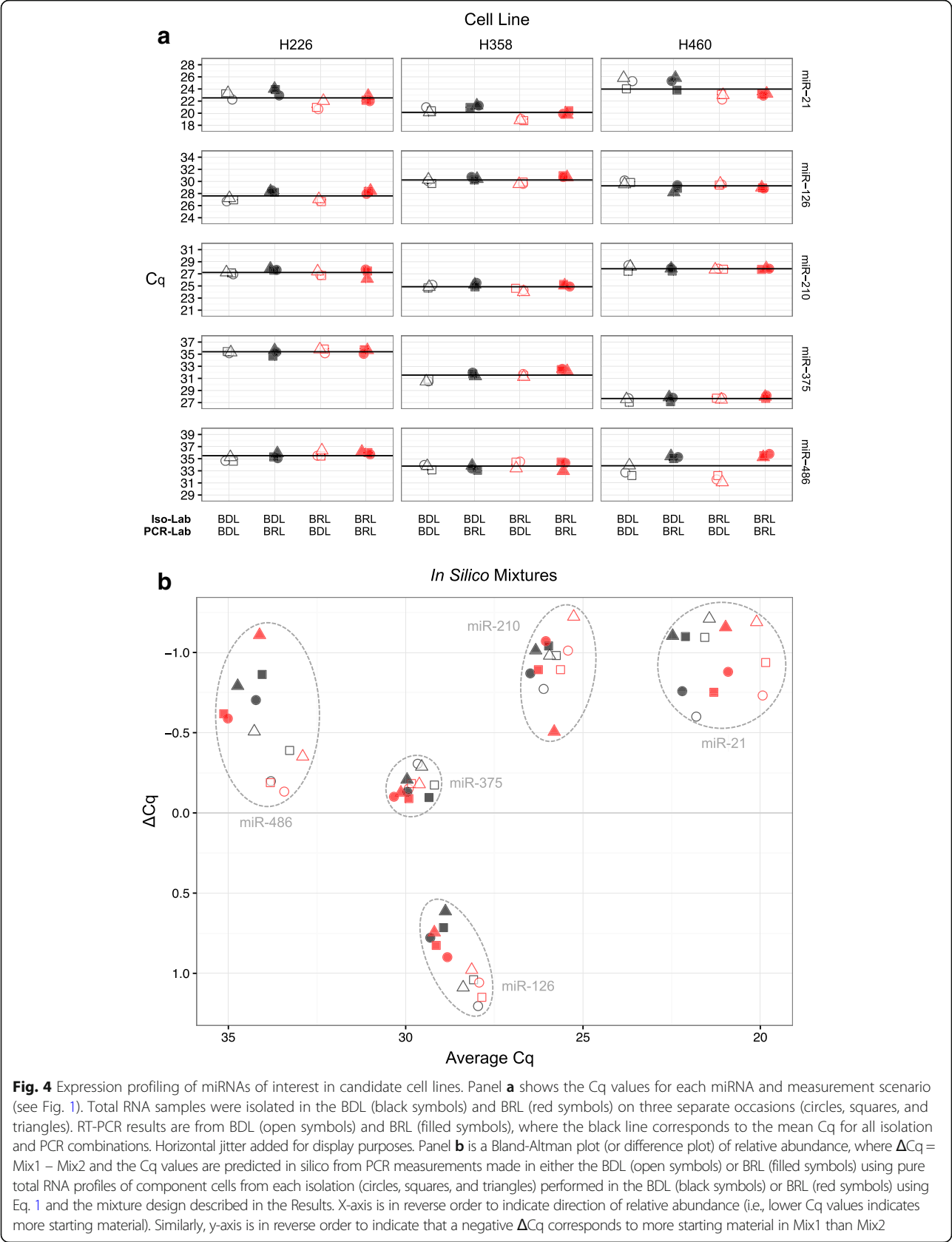
than from the BDL (red symbols and black symbols, respectively). Both miR-126 and miR-486 showed more difference with respect to which lab performed the PCR measurements, BDL (open symbols) or BRL (filled symbols). Both miR-375 and miR-486 had Cq values around 35 in H226 cells, a value indicating either single molecule detection or noise [20], indicating minimal, if any, signal contribution to the final mixtures. Overall, expression levels were sufficiently different between H226, H358, and H460 to proceed to in silico modeling for mixtures. ANOVA results of laboratory factors after the datasets are parsed by analyte and cell line are available in Additional file 6.

Two different mixture designs were selected on the basis of iterative in silico modeling via Eq. 1. The results from each iteration were viewed as a Bland-Altman plot or difference plot [21], where the predicted measurement difference between two mixtures is plotted as a function of the predicted average of measurements for both mixture. Mixture fractions of $\phi_{11} = 0.2$, $\phi_{21} = 0.6$, $\phi_{31} = 0.2$ were selected for mixture design 1 (Mix1); and $\phi_{12} = 0.6$, $\phi_{22} = 0.2$, $\phi_{32} = 0.2$ for mixture design 2 (Mix2). This design predicted a useful distribution of analyte abundances across the dynamic range as well as ΔCq (Mix1 – Mix2) between the samples ranging from -1 to 1 , i.e., two-fold up and down as shown in Fig. 4b.

Each measurement process demonstrated a consistent rank order of miRNA with respect to average Cq. Although the smallest difference in that dimension was observed between miR-126 and miR-375, the mixing design still predicted a useful separation between those two miRNAs with respect to ΔCq . The isolation lab differences observed with miR-21 in Fig. 4a produce a similar trend in Fig. 4b for the modeled mixtures, where the samples isolated in the BRL are predicted to be more abundant (lower average Cq) than those from the BDL. Because this trend was similar for each cell line, the modeled ΔCq are nearly the same (see also Fig. 9, In Silico). The trends for the PCR lab differences observed in Fig. 4a for miR-126 and miR-486 vary across combinations of cell line and analyte (see ANOVA results in Fig. 3 corresponding to interaction among analyte, cell line and PCR lab), so the modeling predicts the average of both Cq and ΔCq will be affected (see also Fig. 9, In Silico). The modeled mixtures, as shown in Fig. 4b, provided a sufficiently unique pattern of differences to proceed with preparation of actual mixtures of total RNA.

Reference samples of total RNA mixtures

Four separate total RNA isolations per cell line from separate expansions were performed in the BRL and



each isolation was used to prepare a single Mix1 and Mix2 pair (see Fig. 2) using the cell line mixture fractions selected from the in silico modeling. Mixture samples were each measured in triplicate at both PCR labs (see Fig. 5a). Both PCR laboratories detected a lower amount of miR-486 in Mix1 and lower amounts of miR-21 and miR-210 in both RNA mixtures derived from the fourth isolation (diamond symbols). This observed effect could be attributable to differences in growth conditions for the cell expansion used for the fourth isolation. However, the observed differences between Mix1 and Mix2 remained comparable with the sample design. For example, the average ($\pm$ SD) ΔCq for three replicate plates of miR-21 measured in the BDL was -1.03 ± 0.11 ,

-0.67 ± 0.16 , -0.85 ± 0.12 , and -0.69 ± 0.15 for isolations 1, 2, 3 and 4, respectively.

An ANOVA was performed on the Cq values in phase 2 (RNA mixtures). In this phase, all isolations occurred at the BRL, which precludes further assessments of variability due to isolation lab. Figure 6a shows the factors with the greatest mean of squares. Analyte was the most prominent component of variability. The confounding factors corresponding to which lab conducted PCR analysis and differences among the four replicate isolations were the next most prominent sources of variability. The most substantial sources of confounding variability with respect to evaluating log abundance ratios across mixtures are the interactions between mixture and

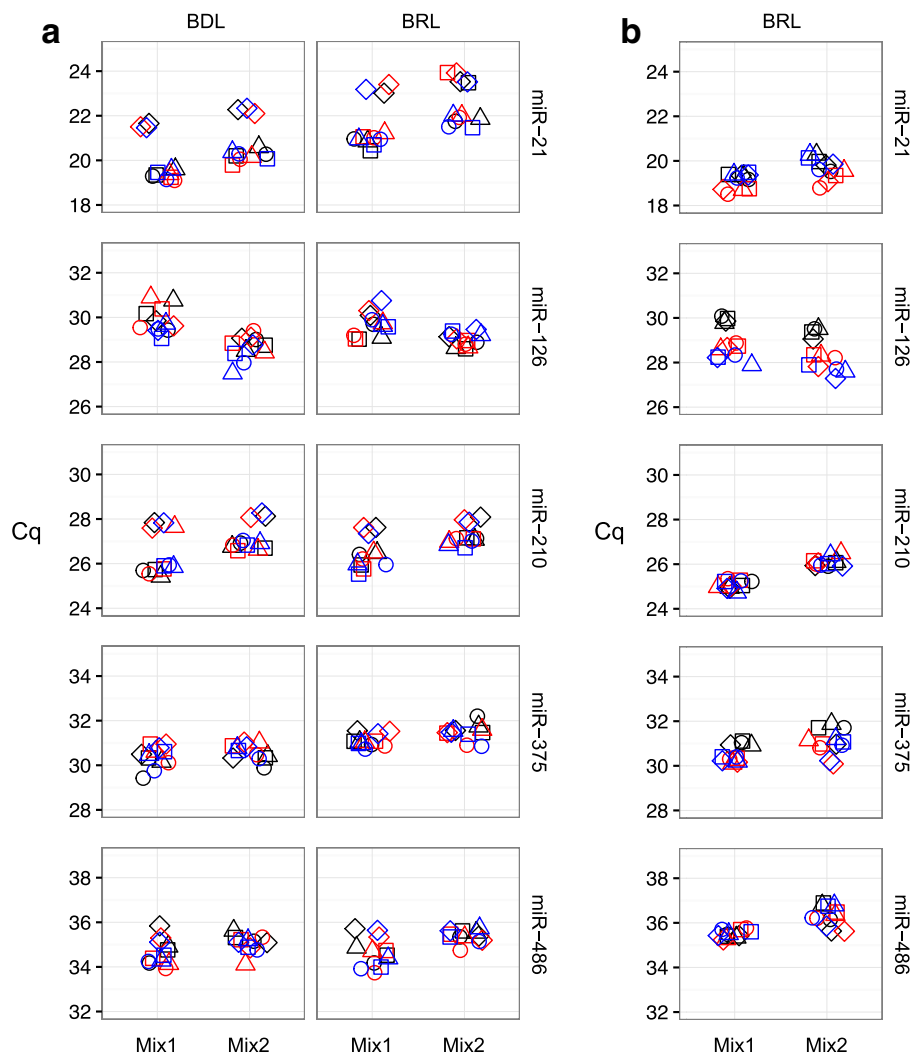


Fig. 5 Expression profiling of miRNAs of interest in RNA mixtures. Panels a and b

show the Cq values for total RNA mixtures and cell line mixtures, respectively. Mix1 and Mix2 were prepared volume:volume:volume, normalized for concentration, using pure total RNA isolated in the BRL (see Fig. 2, scenarios 5 and 6) or count:count:count from intact cells grown in the core facility (see Fig. 2, scenario 7) on four separate occasions (circles, squares, triangles, and diamonds). Each set of RNA mixture pairs (e.g. Mix1 circles and Mix2 circles) were split and measured in triplicate plates (black, red, and blue) by RT-PCR in both the BDL and BRL. For the cell line mixtures, only the BRL extracted RNA from each set of cell line mixture pairs and then measured by RT-PCR in triplicate plates. Horizontal jitter added for display purposes

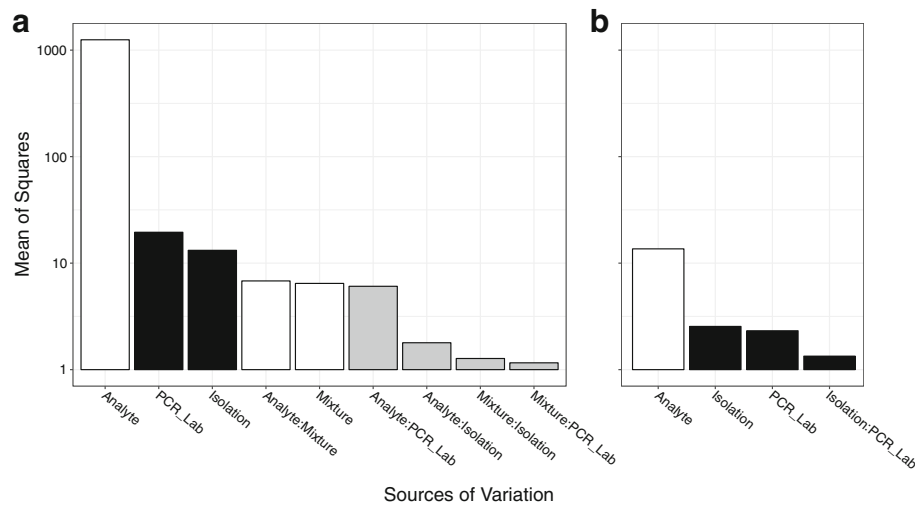


Fig. 6 Sources of variance in the total RNA mixture dataset. An analysis of variance in R [17] was used to identify the major sources of variability in measuring 4 RNA mixture pairs by two laboratories (see Fig. 2, pathways 5 and 6). A four-way ANOVA, testing all possible interactions, was applied to determine the contribution to variance by analyte, mixture, isolation process, and PCR laboratory. In panel a, the input data for the model was the Cq for each of the miRNA measured by each PCR laboratory for each mixture prepared from four isolations (see Fig. 5, panel a). In panel b, the input data was the Δ Cq between each mixture from the same dataset. The mean of squares is plotted on the y-axis on a log10 scale and is a measure of the average contribution of each factor or interaction to the variability in the experiment. Factors or interactions with a mean of squares less the one are not included. White bars correspond to biological factors or interactions, black bars correspond to laboratory factors or interactions, and grey bars correspond to interactions between biological and laboratory factors

isolation and between mixture and PCR lab. The degree of variability attributed to mixture and the interaction between mixture and analyte are largely controlled by the designed abundance ratios resulting from the chosen cell line proportions for each mixture. Figure 6b corresponds to an ANOVA conducted on Δ Cq (Mix1 – Mix2) values. The designed differences in Δ Cq values across the five analytes represent the most prominent component of variability.

Reference samples from cell line mixtures

To test the feasibility of using cell lines as surrogate samples, the three component cell lines were combined in the previously provided Mix1 and Mix2 proportions based upon cell counts to create four separate mixture pairs of cell suspensions. The mixture samples were then processed in triplicate by the BRL in the same manner as cells in a clinical specimen (see Fig. 5b). A PCR plate effect was observed for miR-21 and miR-126 when mixtures were considered separately. However, the observed differences between Mix1 and Mix2 remained comparable with the sample design. For example, the average ($\pm$ SD) Δ Cq for four isolations of miR-126 measured in the BRL was 0.56 ± 0.22 , 0.55 ± 0.26 , and 0.55 ± 0.30 , for plates 1, 2, and 3, respectively.

Detecting differences between reference sample pairs

Figure 7 shows the results for measuring the difference between samples applied to all data derived from either in

silico modeled mixes, RNA mixes, or cell mixes using two different methods. The delta Cq (Δ Cq) method was used to directly measure the relative difference between Mix1 and Mix2 using the mean Cq values from each miRNA target summarized by process (isolation lab, PCR lab, and sample type). In the delta-delta Cq ($\Delta\Delta$ Cq) method [22], each Cq value for a miRNA target was adjusted to produce a Δ Cq by subtracting the Cq value corresponding to miR-16 derived from the same sample (i.e., subject to the same experimental conditions of isolation, PCR lab) and measured on the same plate. The miR-16 adjusted values for each mixture were then compared ($\Delta\Delta$ Cq). Figure 7a shows the distribution of relative abundance and differences in a plot similar to Fig. 4b, with the x-axis normalized to the average miR-21 Cq value. Examining the residuals highlights there is more variability in the average relative Cq than in the Δ Cq (Fig. 7b). This is consistent with the ANOVA results shown in Figs. 3 and 6, which suggest the most prominent sources of confounding variation will cancel out when taking differences to compute Δ Cq values. Additionally, for these data, using miR-16 as a control ($\Delta\Delta$ Cq method) did not substantially reduce variability among multiple measurements of log abundance ratios compared to those derived directly from the Cq values of miRNA targets (Fig. 7, panels c and d compared to panels a and b, respectively).

Discussion

To characterize potential biases between labs, Fig. 8 portrays the marginal likelihood for the difference in

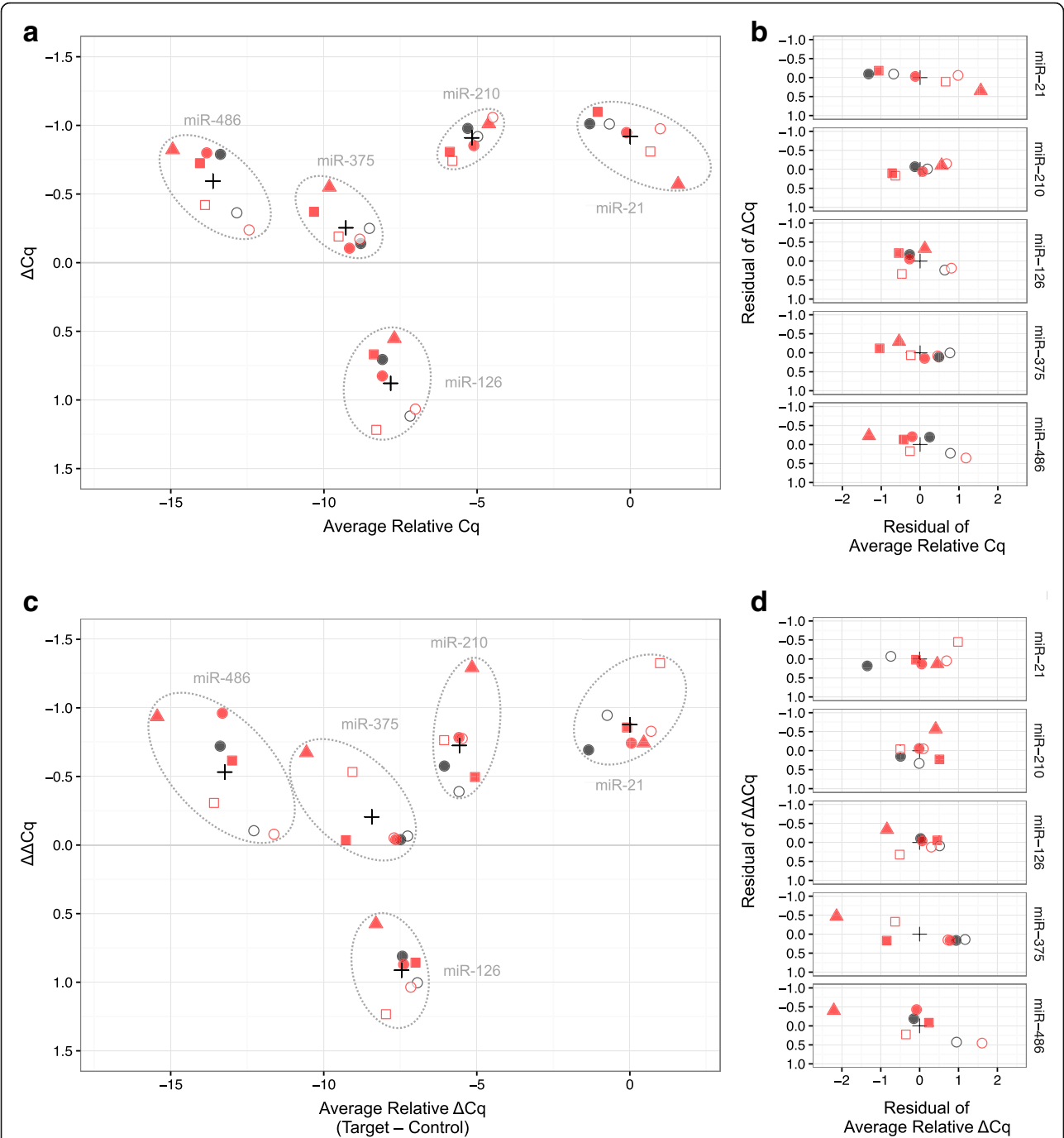


Fig. 7 Average difference in relative abundance between Mix1 and Mix2 for all measurement scenarios (see Figs. 1 and 2). Panel **a** is the Bland-Altman plot of relative abundance where $\Delta Cq = \text{Mix1} - \text{Mix2}$ and Cq values are rescaled relative to the average miR-21 value. Panel **c** is the Bland-Altman plot of relative abundance of Mix1 and Mix2 compared with $\Delta\Delta Cq$ result obtained when using miR-16 as the control [22]. Panels **b** and **d** display residuals from panels **a** and **c**, respectively (i.e., data for each miRNA are re-centered around corresponding black cross seen in panels **a** and **c**, respectively). The black cross denotes center of all values for a particular method of calculation (either ΔCq or $\Delta\Delta Cq$) for each miRNA. Summarized values for each miRNA were compared for each analysis method across the three samples: in silico mixes, RNA mixes, and cell mixes (circles, squares, and triangles, respectively). RNA isolations were performed in either the BDL (black symbols) or BRL (red symbols) and analyzed by RT-PCR in the BDL (open symbols) or BRL (filled symbols). Y-axis is in reverse order to indicate that a negative ΔCq corresponds to more starting material in Mix1 than Mix2

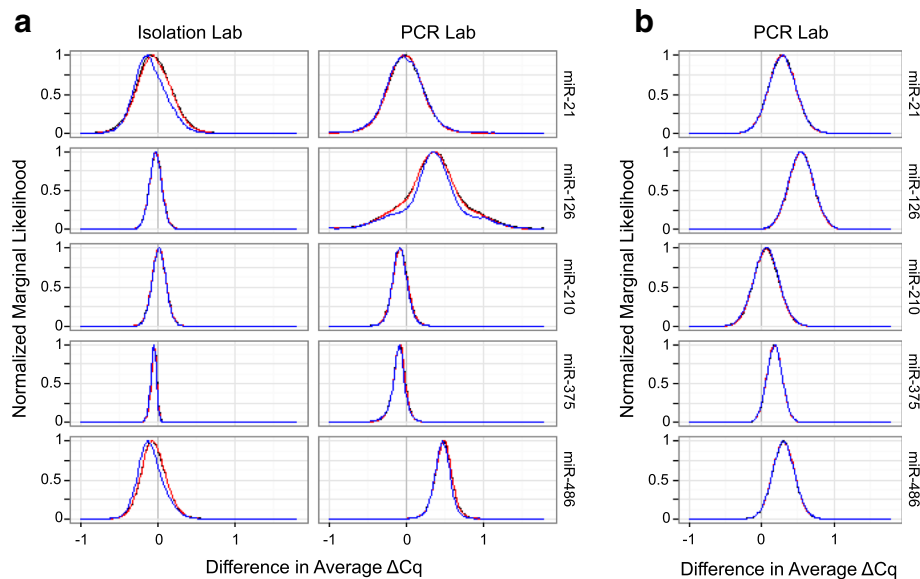


Fig. 8 Evaluation of differences in average ΔCq (Mix1–Mix2) between measurement scenarios differing only by measurement or isolation lab. The y-axis reflects the marginal likelihood (normalized so that the maximum is one). Panel **a** data based on average Cq values inferred for mixtures 1 and 2 derived using Eq. 1 and the average Cq of component cell lines. Isolation lab effect determined by the average difference in ΔCq corresponding to samples isolated at BDL versus those isolated at BRL, for the in silico mixtures. PCR lab effect determined by the average difference in ΔCq corresponding to samples measured at BDL versus those measured at BRL, for the in silico mixtures. Panel **b** data based on mixtures of RNA isolated and prepared in the BRL. PCR lab effect determined by the average difference in ΔCq corresponding to samples measured at BDL versus those measured at BRL. Black, red, and blue lines correspond to model variations 1, 2, and 3, respectively, as described in Additional file 1

average ΔCq values (Mix1 – Mix2) among six measurement scenarios that share source material but differ in either PCR lab or isolation lab, the seventh scenario using cell line mixtures was only measured by one lab (see Figs. 1 and 2 and Additional file 1:Table S2). In panel a, in silico modeling of Mix1 and Mix2 (see Methods) was used to predict ΔCq values, and the differences in the predicted ΔCq values between the BDL and BRL, either as the isolation lab or as the PCR lab, were used to assess laboratory bias for those two steps of the measurement process. These modeling efforts suggest that the likelihood of the observed data is largest under models for which biases between isolation labs are near zero, supporting a hypothesis that different labs isolating RNA from the same source can provide material with similar miRNA ratio profiles.

The likelihoods of the observed data for miR-21, miR-210, and miR-375, respectively, are largest under models for which biases due to PCR lab are near zero. However, for both miR-126 and miR-486 and for both the in silico mixtures (Fig. 8a) and the RNA mixtures (Fig. 8b), the likelihoods are largest under models for which the average ΔCq value corresponding to PCR measurements conducted at the BDL are roughly 0.5 greater than the average ΔCq value corresponding to PCR measurements conducted at the BRL. These interpretations are consistent with the spatial separation observed in Fig. 4b when

comparing open symbols (BDL) and filled symbols (BRL) for miR-126 and miR-486.

If the measurements in phases 1 and 2 had been deemed insufficiently reproducible, additional measurements could be obtained to confirm the apparent biases along with efforts to identify differences in how PCR measurements were conducted at each lab for each combination of cell line and either miR-126 or miR-486. For the proof-of-concept purposes of the present study, the exhibited reproducibility was deemed adequate and no further attempts were made to investigate or control the observed variability.

Although the source material for each of these measurement scenarios was obtained from different expansions of the cell lines, the direction of fold-change between samples for each target miRNA is consistent and the order of relative abundance is the same among the miRNA targets: miRNA-486 < miRNA-375 < miRNA-210 < miR-21 (see Fig. 7).

To characterize potential biases in ΔCq among source materials, Fig. 9 portrays the marginal likelihood for the average ΔCq values (Mix1 – Mix2) among the seven measurement scenarios. The top row of graphs is based on the in silico modeled ΔCq values from the cell line profiles, and the bottom row includes both the RNA mixtures and cell line mixtures. In each plot, the height of the likelihood curve reflects the chance of the

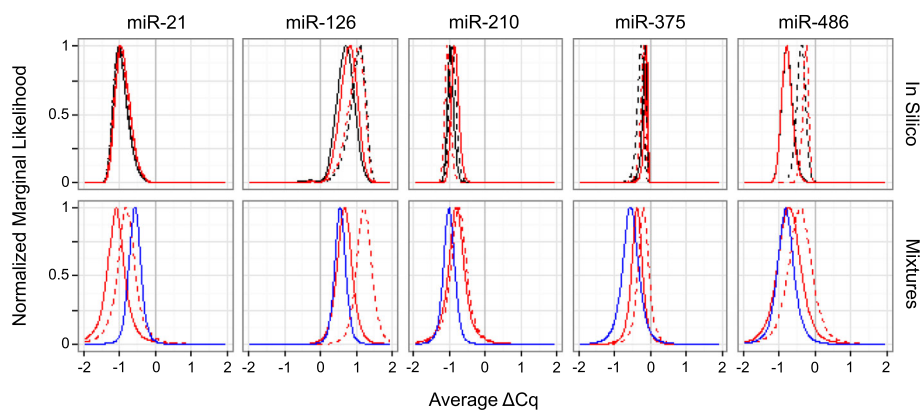


Fig. 9 Comparison of average ΔCq (Mix1 – Mix2) across experimental workflows. The y-axis reflects the marginal likelihood (normalized so that the maximum is one). The top row of graphs compares ΔCq results from model 1 across measurement processes (see Fig. 1) using in silico modeled Cq values (see Eq. 1) for each mixture. Black and red lines correspond to samples isolated at BDL and BRL, respectively. Dashed and solid lines correspond to RT-PCR measurements from BDL and BRL, respectively. The bottom row of graphs compares ΔCq results from model 1 between mixture paradigms (see Fig. 2). Red lines correspond to RNA isolated in the BRL then mixed, blue lines correspond to RNA isolated in the BRL after mixing cells. Dashed and solid lines correspond to RT-PCR measurements from BDL and BRL, respectively

observed data occurring given a model for which the true average ΔCq value is fixed at the corresponding value along the x-axis, after integrating across the modeled distribution of all other model parameters. For each miR, the peaks are either all above or all below zero, corroborating that the direction of change (negative or positive ΔCq) is consistent for each miRNA across all source material and measurement combinations. For miR-210 and miR-375, all seven peaks appear to substantially overlap meaning there is little evidence of bias among any of the measurement scenarios. For miR-126 and miR-486, there appears to be some peak separation between measurement scenarios that differ with respect to which lab conducted PCR but strong overlap among peaks corresponding to measurement scenarios that differ only by source material. The greatest variability in average ΔCq associated with source material appears to occur within miR-21, and, even then, the peaks still overlap. This statistical analysis does not conclusively indicate that all three source materials produce equivalent ΔCq values on average, nor does it conclusively refute that claim. That is, using cell lines as a regenerable source for designing cell and RNA mixture reference materials exhibits potential and warrants further study.

The reported results based on marginal likelihoods appear consistent across the three considered variations of the statistical model, indicating they are not unique to one specific subjective modeling choice. Further consideration of additional models not evaluated in this analysis may produce results that differ substantially. Analyzing additional samples would further reduce the uncertainty underlying these assessments.

To the extent that variability in average ΔCq among the source materials may be present, some of this difference

might be due to drift in cell line profiles, since cells used for in silico modeling were cultured several months before those used to produce the RNA and cell mixtures and the individual cell lines were not separately profiled again at the time the mixtures were prepared. Preparing a sufficiently large “batch” of each cell line from one expansion to partition into source material for individual cell line profiles, total RNA mixtures, and cell line mixtures might reduce variability among observed ΔCq values for each analyte. In addition, determining the miRNA fraction per cell line may also allow for further refinement of the target values derived from cell line profiles and provide a more quantitative assessment of measurement performance. A difference in mRNA fraction has been shown to affect the expected log ratios in other gene expression experiments [23]. A similar use of synthetic miRNA spike-ins introduced into these RNA mixtures could help determine if there are differences in the miRNA fraction in total RNA. Further stabilizing ΔCq between replicate sample pairs may require additional optimization of cell growth conditions or identification of cell lines with greater differences among the miRNA profiles.

Conclusions

Cell line mixtures can be prepared with designed-in differences between samples in a reproducible process and may provide suitable reference samples to assess the entire measurement process from RNA extraction through miRNA measurement. The selection of potential cell lines may be based on previous laboratory experience with material or derived from the literature (including cell repository descriptions). The screening of candidate cell lines by multiple laboratories was useful in establishing robust differences in expression for a selected set of

miRNAs. In silico modeling of these levels informed the final mixture design for both RNA mixtures and mixtures of cells. Identifying cell lines with even greater differences between miRNA profiles might allow for mixtures that provide a broader range of designed-in ratios. RNA mixtures derived from these cell lines can also serve as well-matched reference samples for measurement assurance of the post-RNA isolation stages of gene expression measurement systems.

This testing paradigm should be generalizable to other biomarkers and other laboratories and provide a model for establishing both intra- and inter-laboratory measurement assurance and demonstrate translatability of biomarker assays between research development laboratories and clinical reference laboratories.

Additional files

Additional file 1: Intro to Bayesian analysis. (PDF 1844 kb)

Additional file 2: Cell line profile Cq values. (CSV 13 kb)

Additional file 3: RNA mixture Cq values. (CSV 18 kb)

Additional file 4: Cell line mixture Cq values. (CSV 8 kb)

Additional file 5: Code for running MCMC. (PDF 41 kb)

Additional file 6: Phase1 ANOVA parsed by analyte and cell line. (PDF 78 kb)

Abbreviations

BDL: biomarker discovery laboratory; BRL: biomarker reference laboratory; CAP: College of American Pathology; CLIA: Clinical Laboratory Improvement Amendments; EDRN: Early Detection Research Network; MCMC: Markov Chain Monte Carlo; miRNA: microRNA; NIST: National Institute of Standards and Technology; RT-PCR: reverse transcription polymerase chain reaction

Acknowledgements

Authors would like to thank LaToya Lewis and Atieno Akumu for performing RT-PCR assays in the BRL and BDL, respectively; and Rena Lapidus and Brandon Cooper of the cell culture core lab for maintaining and distributing cell lines.

Funding

The work presented in this manuscript is jointly designed, executed and written under the auspices of an Interagency Agency Agreement between the National Cancer Institute and the National Institute of Standards and Technology, both of which are Federal Agencies supported by the funds from US Government. SAS and FJ were supported by an NCI-EDRN Cooperative Agreement Grant: U24CA115091.

Availability of data and materials

The datasets supporting the conclusions of this article are included within the article and its additional files.

Author's contributions

PSP, MS, and LS designed the study. SAS, DK, and FJ developed standard operating procedures for sample handling, and acquired and processed the data. PSP developed metrics and visualizations. SPL provided statistical analysis. PSP and SPL drafted the manuscript. All authors participated in the revision process and provided final approval.

Disclaimer

Certain commercial entities, equipment or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it

intended to imply that the entities, materials or equipment are necessarily the best available for the purpose.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Joint Initiative for Metrology in Biology, National Institute of Standards and Technology, 443 Via Ortega, Stanford, CA 94305, USA. ²Statistical Engineering Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA. ³Department of Pathology, University of Maryland School of Medicine, Baltimore, MD 21201, USA. ⁴Division of Cancer Prevention, National Cancer Institute, Rockville, MD 20850, USA.

Received: 10 July 2017 Accepted: 7 March 2018

Published online: 20 March 2018

References

- Kern SE. Why your new cancer biomarker may never work: recurrent patterns and remarkable diversity in biomarker failures. *Cancer Res.* 2012; 72(23):6097–101.
- Plant AL, Locascio LE, May WE, Gallagher PD. Improved reproducibility by assuring confidence in measurements in biomedical research. *Nat Methods.* 2014;11(9):895–8.
- Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, et al. Phases of biomarker development for early detection of cancer. *J Natl Cancer Inst.* 2001;93(14):1054–61.
- Srivastava S, Kramer BS. Early detection cancer research network. *Lab Invest.* 2000;80(8):1147–8.
- Hammond SM. An overview of microRNAs. *Adv Drug Deliv Rev.* 2015;87:3–14.
- Reddy KB. MicroRNA (miRNA) in cancer. *Cancer Cell Int.* 2015;15:38.
- Gazdar AF, Girard L, Lockwood WW, Lam WL, Minna JD. Lung cancer cell lines as tools for biomedical discovery and research. *J Natl Cancer Inst.* 2010;102(17):1310–21.
- Pfeffer SR, Yang CH, Pfeffer LM. The role of miR-21 in Cancer. *Drug Dev Res.* 2015;76(6):270–7.
- Ebrahimi F, Gopalan V, Smith RA, Lam AK. miR-126 in human cancers: clinical roles and current perspectives. *Exp Mol Pathol.* 2014;96(1):98–107.
- Lu J, Xie F, Geng L, Shen W, Sui C, Yang J. Potential role of MicroRNA-210 as biomarker in human cancers detection: a meta-analysis. *Biomed Res Int.* 2015;2015:303987.
- Yan JW, Lin JS, He XX. The emerging role of miR-375 in cancer. *Int J Cancer.* 2014;135(5):1011–8.
- Wang J, Tian X, Han R, Zhang X, Wang X, Shen H, Xue L, Liu Y, Yan X, Shen J, Mannoor K, Deepak J, Donahue JM, Stass SA, Xing L, Jiang F. Downregulation of miR-486-5p contributes to tumor progression and metastasis by targeting protumorigenic ARHGAP5 in lung cancer. *Oncogene.* 2014;33(9):1181–9.
- Yu L, Todd NW, Xing L, Xie Y, Zhang H, Liu Z, et al. Early detection of lung adenocarcinoma in sputum by a panel of microRNA markers. *Int J Cancer.* 2010;127(12):2870–8.
- Xing L, Todd NW, Yu L, Fang H, Jiang F. Early detection of squamous cell lung cancer in sputum by a panel of microRNA markers. *Mod Pathol.* 2010; 23(8):1157–64.
- Pine PS, Boedigheimer M, Rosenzweig BA, Turpaz Y, He YD, Delenstarr G, et al. Use of diagnostic accuracy as a metric for evaluating laboratory proficiency with microarray assays using mixed-tissue RNA reference samples. *Pharmacogenomics.* 2008;9(11):1753–63.
- Pine PS, Rosenzweig BA, Thompson KL. An adaptable method using human mixed tissue ratiometric controls for benchmarking performance on gene expression microarrays in clinical laboratories. *BMC Biotechnol.* 2011;11:38.
- R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2016. URL <http://www.R-project.org/>.
- Plummer M. Rjags: Bayesian graphical models using MCMC. R package version. 2016:4–6. <https://CRAN.R-project.org/package=rjags>.

19. Brooks S, Gelman A, Jones G, Meng XL, editors. Handbook of Markov chain Monte Carlo: Chapman and Hall/CRC; 2011.
20. Mestdagh P, Feys T, Bernard N, Guenther S, Chen C, Speleman F, Vandesompele J. High-throughput stem-loop RT-qPCR miRNA expression profiling using minute amounts of input RNA. *Nucleic Acids Res.* 2008; 36(21):e143.
21. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res.* 1999;8(2):135–60.
22. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2[−]Delta Delta C(T) method. *Methods.* 2001;25(4):402–8.
23. Munro SA, Lund SP, Pine PS, Binder H, Clevert DA, Conesa A, et al. Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. *Nat Commun.* 2014;5:5125.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit



ORIGINAL ARTICLE

Integrated analysis of miRNAs and DNA methylation identifies miR-132-3p as a tumor suppressor in lung adenocarcinoma

Yun Su¹, Amol Shetty² & Feng Jiang³ ¹ Department of Surgery, Nanjing University of Chinese Medicine, Nanjing, China² Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland, USA³ Department of Pathology, University of Maryland School of Medicine, Baltimore, Maryland, USA**Keywords**

DNA methylation; epigenetics; lung cancer; microRNA.

Correspondence

Feng Jiang, Department of Pathology,
University of Maryland School of Medicine,
10 S. Pine St., Baltimore, MD 21201, USA.
Tel: +01 410 706 4854
Fax: +01 410 706 7717
E-mail: fjiang@som.umaryland.edu

Received: 25 March 2020;

Accepted: 6 May 2020.

doi: 10.1111/1759-7714.13497

Thoracic Cancer **11** (2020) 2112–2124**Abstract**

Background: Aberrant miRNA expression and DNA methylation are two major epigenetic events in lung adenocarcinoma (LUAD). We conducted a combined analysis of the molecular changes in LUAD.

Methods: We analyzed differentially expressed miRNAs and methylated CpG loci in 489 LUAD tissues versus 49 normal lung tissues of the Cancer Genome Atlas (TCGA). The results were validated in cell lines and xenograft mouse models and additional pairs of 36 LUAD and 36 normal lung tissues.

Results: A total of 125 differentially expressed miRNAs and 145 differentially methylated CpG loci were identified in the LUAD versus normal lung tissues of TCGA data. Expression of the 22 miRNAs was inversely correlated with the 47 differentially methylated sites located in the miRNAs. Molecular and cellular function analysis showed that the abnormally methylated miRNAs were mainly involved in cell-to-cell signaling and interaction in airway cells. The DNA methylation status and altered expressions of miRNAs and their target genes were confirmed in 36 pairs of lung tumor and noncancerous lung tissues. Furthermore, aberrant miRNA expressions or DNA methylations alone could be involved in tumorigenesis of LUAD via different pathways. In addition, elevated miR-132-3p expression, reduced expression of its targeted gene (ZEB2), and decreased cell proliferation was observed in lung cancer cells treated with DNA methyltransferase inhibitor. Moreover, in vitro and in vivo analyses showed that miR-132-3p downregulation via DNA methylation promoted tumorigenicity of lung cancer by directly regulating ZEB2.

Conclusions: The interaction between two epigenetic aberrations could have important functions in LUAD. miR-132-3p might act as a tumor suppressor in the tumorigenicity of LUAD.

Key points**Significant findings of the study:**

- Systemically investigating relationship between aberrant miRNA expression and DNA methylation in lung cancer could improve understanding of lung tumorigenesis and develop diagnostic and therapeutic targets.

What this study adds:

- Three forms of relationships between the two epigenetic changes are defined.
- miR-132-3p is further identified as a tumor suppressor in lung cancer.

Introduction

Non-small-cell lung cancer (NSCLC) is the number one cancer killer in men and women. NSCLC mainly consists of three histological subtypes: lung adenocarcinoma (LUAD), squamous cell carcinoma (SCC), and large cell carcinoma (LCC). LUAD is the most common type of lung cancer, accounting for 40% of all NSCLCs. Molecular changes, particularly altered epigenetic processes, can lead to malignant cellular transformation.¹ Through transcriptional silencing of critical cancer-associated genes, DNA methylation plays a vital role in cancer pathogenesis.¹ As another major component of epigenome, microRNA (miRNA) is a key regulator at the post-transcriptional levels that modulate transcriptome changes.² Furthermore, miRNAs regulate many cancer-related genes associated with different biological processes, including proliferation, apoptosis, development, and tumorigenesis.² Previous studies³⁴ have shown that altered methylation of miRNAs are associated with dysregulation of the genes and contribute to the development and progression of lung cancer.

The Cancer Genome Atlas (TCGA) contains multiomics data at different molecular levels and provides more than 10 000 patients with nearly 33 types of cancer.⁵ Integrated analysis of the multiomics data would deepen our understanding of the complex mechanisms underlying tumorigenesis.⁵ In this study, we investigated the epigenetic relationship between miRNA and DNA methylation in the pathogenesis of LUAD by analyzing TCGA multigenomic data. The results were further validated in cell lines and xenograft mouse models and additional LUAD tumor and the matched noncancerous lung tissues.

Methods

Data collection and data preprocessing

Genome-wide DNA methylation, miRNA expression, and clinical information of the patients were retrieved from TCGA data portal (<https://tcga-data.nci.nih.gov/tcga/>, accessed March 2016). The inclusion criterion was the specimens with DNA methylation, miRNA expression, and clinical information. The platforms for DNA methylation were Human methylation 27 with 27 578 probes and Human Methylation 450 with 485 577 probes. The platforms for miRNA were Agilent two-channel array with 722 human mature miRNA expressions and Illumina HiSeq_miRNASeq with 1036 human mature miRNA expressions. Finally, 489 LUAD tissues and 49 matched normal tissues were selected from 489 lung cancer patients. The detailed clinical characteristics of the patients are listed in Table 1. To analyze miRNA and DNA methylation data, we removed outlier samples with TCGAanalyze_Preprocessing function of TCGAAbiolinks using a Spearman correlation cutoff of 0.6. The datasets for GC-content [47] and library size were

Table 1 Clinical characteristics of 489 patients diagnosed with LUAD in TCGA

	Number of cases (%)
Age at diagnosis	
<60	133 (27.2%)
>60	356 (72.8%)
Sex	
Female	196 (40.1%)
Male	293 (59.9%)
Smoking status	
10 < Pack-years	222 (45.4%)
10 ≥ Pack-years	221 (45.2%)
NA	46 (9.4%)
Stage	
Stage I	340 (73.0%)
Stage II	73 (14.9%)
Stage III–IV	58 (12.1%)

LUAD, lung adenocarcinoma; TCGA, the Cancer Genome Atlas.

normalized by using the TCGAanalyze_Normalization. The miRNA data between LUAD and normal tissues were analyzed using limma package in R. The fold changes (FCs) in the expression of individual miRNAs were calculated and miRNAs with $\log_2$ (FC) >1.0 and $P < 0.05$ were significant. We used the MEDIPS package (version 1.24.0) for the analysis and comparison of DNA methylation datasets of LUAD and normal lung tissues.⁶ P -values <0.05 and $\log_2$ (FC) ≥1 were considered to show differentially methylated CpG island loci.

Lung tissue specimens

With a study protocol (Protocol number: HP00040666) approved by the Institutional Review Board of the University of Maryland Baltimore, we obtained 36 frozen LUAD tumor tissues and the matched noncancerous lung tissues from a tissue bank of the University Medical Center. We reviewed complete medical records of the patients. Clinical diagnosis of lung cancer was made using histopathologic examinations of specimens obtained by CT-guided transthoracic needle biopsy, transbronchial biopsy, video-assisted thoracoscopic surgery (VATS), or surgical resection. The surgical pathologic staging was determined according to the TNM classification of the International Union Against Cancer with the eighth American Joint Committee on Cancer and the International Staging System for Lung Cancer. Histopathological classification was determined according to the World Health Organization classification. Clinical characteristics of the surgically-resected tissues are shown in Table 2.

Analysis of DNA methylation in miRNAs

We extracted the transcription start site (TSS) data of miRNAs from the Functional Annotation of the

Table 2 Clinical characteristics of surgically-resected lung tumor diagnosed with LUAD and matched noncancerous normal lung tissues of 36 patients

Age	66 (SD 9.4)
Sex	
Female	10
Male	26
Smoking status	
Pack-years	30.3 (SD 9.2)
Stage	
Stage I	24
Stage II	6
Stage III-IV	6

LUAD, lung adenocarcinoma; SD, standard deviation.

Mammalian Genome5 database and tMethyl-CpG binding domain-sequencing data. We considered 1000 bp upstream and downstream as coding sequence regions of miRNAs.⁷

Analysis of DNA methylation by quantitative methylation specific PCR (qMSP)

We performed DNA isolation and qMSP as described in our previous publications.^{8, 9} Briefly, we extracted DNA from the specimens using DNeasy kit (Qiagen, Valencia, CA). We eluted DNA with 50 µL of elution buffer (10 mmol/L Tris-Cl, pH 8.5) (Sigma-Aldrich Corporation). DNA was quantified using the Quantifiler Human DNA Quantification kit (Applied Biosystems, Foster City, CA). Bisulfite conversion was carried out on DNA using the Zymo EZ DNA Methylation Kit (Zymo Research, Irvine, CA) according to the manufacturer's protocol. qMSP was done by using a Lightcycler system (Roche Applied Science, Mannheim, Germany)¹⁰. Hypermethylation and hypomethylation markers for the genes were designed. Cycle threshold (Ct) values for each gene were determined. We normalized Ct values of the target genes in relation to that of myogenic differentiation antigen 1 (MYOD1), which was used as control for the qMSP analysis.^{8, 9} By using the formula: $2^{-(Ct(MYOD1) - Ct(target\ gene))} \times 100$, we computed ratio value to decide the relative level of methylation of the genes in each sample as previously described.¹⁰

Analysis of expressions of the miRNAs and the targeted genes by using quantitative reverse transcription PCR (RT-qPCR)

We extracted RNA by using a protocol established in our previous reports.² The purity and concentration of RNA were determined from OD260/280 readings by using a dual-beam Ultraviolet (UV) spectrophotometer (Eppendorf

AG, Hamburg, Germany). RNA integrity was determined by capillary electrophoresis by using the RNA 6000 Nano Lab-on-a-Chip kit and the Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA). We evaluated the expressions of miRNAs and the targeted genes by qRT-PCR with Taqman miRNA assays (Applied Biosystems, Foster City, CA).² Briefly, RNA was applied for reverse transcription (RT) by using the Applied Biosystems 9700 Thermocycler (Applied Biosystems) with miRNA-specific looped primer and TaqMan MicroRNA RT Kit (Applied Biosystems). The reaction included 50 nM stem-loop RT primer, $\times 1$ RT buffer, 0.25 mM each of deoxyrinonucleotides, and 3.33 U/µL MultiScribe reverse transcriptase in a total volume of 15 µL. The 20 µL PCR reaction included RT product, $\times 1$ TaqMan Universal PCR Master Mix (Applied Biosystems), and the corresponding primers and Taqman probe for the target genes. The reactions were incubated in a 94 well plate at 95°C for 15 minutes, followed by 45 cycles of 95°C for 15 seconds and 60°C for one minute. We normalized Ct values of the target miRNAs or the targeted genes in relation to that of U6 and determined relative expression of a gene in a given sample using the eq. $2^{-\Delta Ct}$, where $\Delta Ct = Ct(targeted\ genes) - Ct(U6)$. U6 has been considered one of the suitable endogenous controls for the quantification of miRNA and protein-coding genes in tissue specimens.¹¹ Two interplate controls and one no-template control were carried along in each experiment. All experiments were performed at least three times.

Lentivirus infection

Recombined LV-GV369-miR-132-3p-EGFP and the respective negative control vector were commercially manufactured by GeneChem Co, Ltd. (Shanghai, China). We transduced the lentiviral vectors into A549 cells (ATCC, Manassas, VA) as described in our previous studies.^{12–18} Briefly, we seeded (2×10^5 cells/mL) in 6 well plates and incubated the cells for 24 hours to reach 50% confluence, which were replaced with infection medium containing lentiviral vectors at a multiplicity of infection (MOI) of 20 plaque-forming units/cell. Successfully infected cells were GFP positive under a fluorescent microscope after 48 hours. We evaluated efficiency of miR-132-3p overexpression by PCR.

Western blot analysis

We performed western blotting as described in our previous studies^{12, 14, 17} with antibodies against ZEB2 or NOVA1 (Santa Cruz Biotechnology, Inc., Santa Cruz, CA) or β -actin (Sigma-Aldrich Co., St Louis, MO).

Treatment of cancer cells with 5-aza-2'-deoxycytidine and miRNA inhibitor

Two human LUAD cell lines (A549 and H1975) were obtained from the ATCC. We seeded the cell lines at a density of 5×10^6 /well in six-well plates. The cells were cultured for 24 hours and treated with 10 μ M of the DNA demethylation agent 5-aza-2'-deoxycytidine (5-Aza-CdR; Sigma-Aldrich Co) and phosphate buffered saline (PBS) (Sigma-Aldrich Co., St Louis, MO), respectively. The cells were also treated with either 200 nmol/L of miR-132 inhibitor or inhibitor negative control (Thermo Fisher Scientific Inc., Halethorpe, MD), with Lipofectamine 2000 reagent, respectively (Thermo Fisher Scientific Inc.), according to the manufacturer's instructions. Transfection efficiencies were determined by qRT-PCR.

Cell proliferation assay

Cell proliferation was performed using the methylthiazol tetrazolium (MTT) assay as described in our previous studies.^{14, 16, 18} Briefly, the cells were counted and seeded in six-well plates at a density of 1000 cells per well after transfection. Cells were allowed to grow for 24, 36, and 72 hours. The cells were then stained using crystal violet and each colony with a minimum of 50 cells and counted at each time point. All experiments were repeated three times, each sample in triplicate. Data represented the mean ($\pm$ SD) of three independent experiments.

Cell cycle analysis by flow cytometry

DNA content was analyzed by using flow cytometry as previously described.^{14, 18} Briefly, harvested cells were washed twice with 5mM EDTA/PBS and fixed with 95–100% cold ethanol and kept at 4°C overnight. Cells were incubated with RNase A (50 μ g/mL), and then stained with propidium iodide (PI, 10 mg/mL) (Sigma) for DNA content analysis on a FACScan flow cytometer (Becton Dickinson). Statistical analysis of variance (ANOVA) and least significant difference *t*-tests were used to compare cell cycle data of cells with and without specific siRNA treatments. All experiments were repeated three times, each sample in triplicate. Data represented the mean ($\pm$ SD) of three independent experiments.

Transmigration and wound healing assays

Cells were plated in medium without serum in the top chamber of a transwell (Corning, NY) in order to determine migration and invasion. The bottom chamber contained standard medium with 10% FBS. After incubation

for 24 hours, the cells that had migrated to the lower surface of the membrane were fixed with formalin and stained with crystal violet. The migrating cells were examined microscopically and determined by counting the migrating/invasive cells in five randomly selected fields using an Olympus BX41 microscope. Photomicrographs were taken using a Qcolor5 digital camera system fitted to this microscope. To determine whether the inhibition of cell migration by forced expression of miR-132-3p was due to the inhibition of cell proliferation, the cells with forced expression of miR-132-3p were treated with aphidicolin (1 mg/mL) (Sigma-Aldrich), a proliferation inhibitor. A wound healing assay was performed in 12 well plates (1×10^5 per well). When the cells were grown to 90%–95% confluences, transfection of cancer cells were performed. Wound lines were created manually by scratching the monolayer with a sterile 200 μ l pipette tip and migration of the cells was assessed after 24, 48, and 72 hours. Pictures were taken using a Nikon inverted phase-contrast microscope (Nikon, Melville, NY). The distance between the parallel lines was measured using ImageJ software. All experiments were carried out at least three times. All experiments were repeated three times, each sample in triplicate. Data represented the mean ($\pm$ SD) of three independent experiments.

Tumorigenicity assays in nude mice

The animal study was performed with the approval of the University of Maryland Baltimore under code IACUC# 0516007. Briefly, five athymic Balb/c, Nu/Nu mice per group were inoculated with 1×10^6 A549 cells infected with LV-GV369-miR-132-3p-vector and 1×10^6 A549 cells with control vector, respectively. All the animals were monitored regularly, and tumor growth was measured at regular intervals. The mice were observed for four weeks and then euthanized under deep anesthesia with pentobarbital (Sigma). We calculated volume of the tumors by using formula (length [mm]) $\times$ (width [mm]) 2×0.52 . The tumor size was represented by mean $\pm$ SD mm³.

Pathway analyses

We use Ingenuity Pathway Analysis (IPA) (Ingenuity Systems, Mountain view, CA) to identify the most significant pathways of the genes, as previously described.¹⁹ The gene list files containing gene symbols, FC, *P*-values were uploaded to IPA for the analysis.

Statistical analysis

All data were expressed as mean $\pm$ standard deviation (SD). We used unpaired *t*-test to analyze expression levels

of miRNAs and methylation status in LUAD and corresponding normal tissues. We used chi-square analysis and Pearson's correlation test to evaluate correlation between methylation status and expression levels of miRNAs. We performed the chi-square and *t*-tests to evaluate the relationship between the molecular changes and clinical features. A *P*-value of less than 0.05 was considered statistically significant.

Results

Differentially expressed miRNAs in LUAD tissues versus normal lung tissues

Multomics data of 489 LUAD and 49 matched non-cancerous lung tissues of lung cancer patients in the TCGA were analyzed. The detailed clinical characteristics of the patients are listed in Table 1. From the cohort of the TCGA data, we identified a total of 125 miRNAs that were differentially expressed between the LUAD tissues and matched normal tissues ($P < 0.05$ and $FC > 1.0$) (Table S1). A total of 89 miRNAs were significantly upregulated and 36 were downregulated in the LUAD tissues versus normal lung tissues. Furthermore, we obtained 36 frozen LUAD tumor tissues and the matched non-cancerous lung tissues from a tissue bank of the University Medical Center (Table 2). We selected eight miRNAs (five upregulated and three downregulated miRNAs) and used qRT-PCR to validate the expression levels in the 36 LUAD tissues and the 36 matched normal lung tissues. The eight miRNAs exhibited a significantly different level in the LUAD tissues compared with normal lung tissues (all $P < 0.05$) (Table 3). Moreover, the eight miRNAs had changes in the tissue specimens by qRT-PCR in the same direction as by deep sequencing analysis in TCGA. Therefore, the results generated from the independent set of lung tumor and noncancerous lung tissues confirmed the findings from TCGA data.

Differentially methylated loci in LUAD tissues versus normal lung tissues

A total of 145 CpG-islands were differentially methylated in the 489 LUAD tissues versus 49 noncancerous lung tissues (Table S2). Of the methylation loci, 110 and 35 were hypermethylated and hypomethylated in tumor samples, respectively, compared with normal lung tissues (all $P < 0.05$). We tested four loci (cg07533148, cg04317399, cg07307078, and cg02919422) for the methylation status in additional 36 LUAD tissues and the matched normal lung tissues. The four methylated loci had the same changes in the tissue specimens as in TCGA data (Fig 1). Therefore, the findings produced from the different tissue specimens validated the discoveries from TCGA data.

Relationship between miRNA and DNA methylation in LUAD

To identify miRNA-DNA methylation site pairs where DNA methylation sites were located within 1000 bps from the miRNA upstream and downstream regions, we extracted the TSS data of miRNAs from the FANTOM5 database. We then used Spearman correlation and MA to analyze the correlation between DNA methylation and miRNA expression. Of the 125 differentially expressed miRNAs, 22 (17.6%) were inversely correlated with differentially methylated loci within 1000 bps from the miRNA upstream and downstream regions (Table 4) (all $P < 0.05$). Of the 145 differentially methylated loci, 47 (32.4%) had opposite association with the 22 miRNAs (Table 4) (all $P < 0.05$).

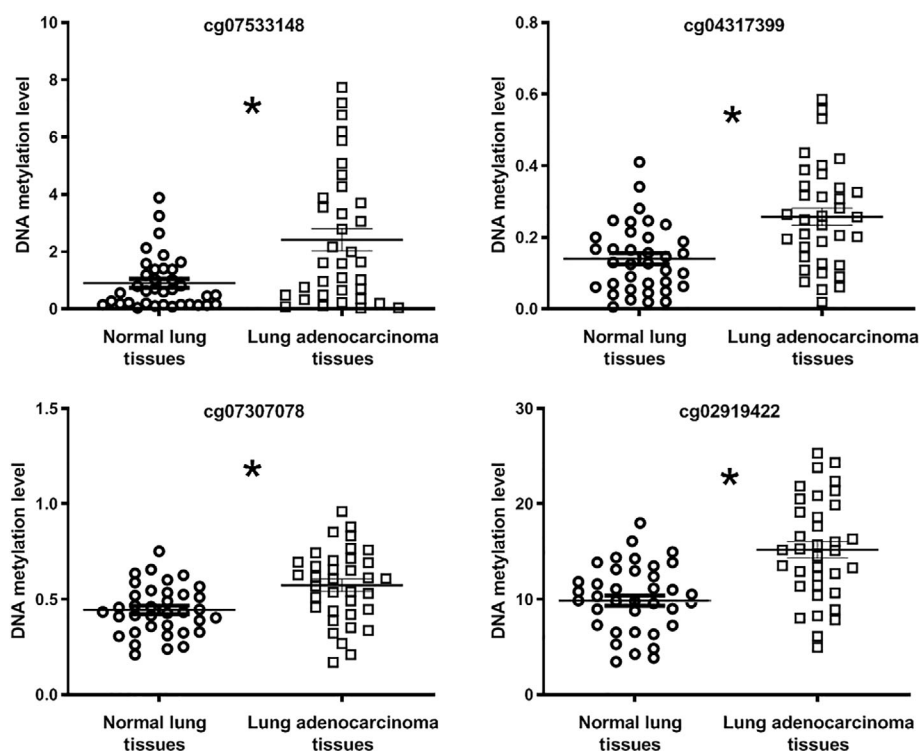
There were more than one differentially methylated sites located in either upstream or downstream regions of miRNAs (Table 4). The 22 miRNAs might be epigenetically regulated by the DNA methylation in the genes. Of the 22 miRNAs, nine miRNAs were downregulated by hypermethylation, and 13 miRNAs were upregulated by the hypomethylation. However, of the 125 differentially expressed miRNAs, 103 (82.4%) were not associated with altered DNA methylation (Table S3). Furthermore, of

Table 3 Expression of eight miRNAs in stage I adenocarcinoma and normal lung tissues measured by RT-PCR miRNAs

	Mean $\pm$ SD in normal tissues	Mean $\pm$ SD in adenocarcinoma tissues	<i>P</i> -value
miR-577	0.234 51 $\pm$ 0.0242345	0.785 15 $\pm$ 0.021453	0.006564
miR-196a-5p	2.435 54 $\pm$ 0.746 75	11.5645 $\pm$ 2.574	0.0006546
miR-182-3p	0.04546 $\pm$ 0.01356	0.123 454 $\pm$ 0.02446	0.0067434
miR-147b	0.223 576 $\pm$ 0.1232	0.542 343 $\pm$ 0.03454	0.00054534
miR-183-3p	3.476 $\pm$ 0.2574	6.2554 $\pm$ 1.4545	0.00956545
miR-184	4.3451 $\pm$ 0.564 34	0.234 354 3 $\pm$ 0.042352	0.00012436
miR-486-5p	212.29 $\pm$ 32.46	17.23 $\pm$ 8.2874	2.69E-03
miR-30a-3p	327.48 $\pm$ 36.2538	22.16 $\pm$ 5.028	4.57E-07

RT-PCR, reverse transcription-polymerase chain reaction; SD, standard deviation.

Figure 1 Quantitative methylation specific PCR analysis of four differentially methylated loci in 36 lung adenocarcinoma tissues and 36 normal lung tissues confirmed data from the TCGA database. All four methylation loci (cg07533148, cg04317399, cg07307078, and cg02919422) were hypermethylated in lung adenocarcinoma tissues compared with the matched normal lung tissues (*all $P < 0.05$).



187 differentially methylated loci in LUAD tissues, 98 (67.6%) were not associated with expression of the miRNAs (Table S4).

To investigate if the direct interaction between miRNA and DNA methylation could affect expression of target genes of the miRNAs, we used miRNA enrichment to predict potential target genes of miR-132-3p and -339. ZEB2 and NOVA1 were the top predicted genes of the two miRNAs, respectively.

We then validated the bioinformatic findings in the 36 LUAD tissues and 36 normal lung tissues. As shown in Fig 2, hypermethylation and reduced expression of miR-132 and -339 and elevated expression of the target genes were validated in the LUAD tissues compared with normal lung tissues. Therefore, the interaction between the two epigenetic events and their impact on their target genes were experimentally validated.

Inhibiting DNA methylation in LUAD cells could elevate miR-132-3p expression and reduce expression of the target gene

To confirm whether expression of the miR-132-3p was affected by the methylation status, we treated two LUAD cell lines (A549 and H1975) with 5-Aza-dC. In silico prediction of CpG islands of miR-132-3p and the DNA methylation levels in LUAD cells are shown in Fig. The A549 cancer cells treated with 5-Aza-dC had a significantly lower methylation

level compared with the cells without the treatment ($P = 0.00016$). Furthermore, 5-Aza-dC treatment could elevate the expression of miR-132-3p in the two cell lines (Fig 4a). In addition, the 5-Aza-dC treatment suppressed the expression level of ZEB2 (Fig 4b), a direct target gene of miR-132-3p. Moreover, cell proliferation was reduced in the cancer cells treated with 5-Aza-dC compared to cancer cells treated with PBS (Fig 4c). However, miR-132-3p inhibitor increased cell proliferation of the cancer cells (Fig 4c). Furthermore, the effect of 5-Aza-dC on reduction of cell proliferation was significantly diminished by miR-132-3p inhibitor. Therefore, the hypermethylation of the miRNAs could regulate expression of the miRNAs, which in turn, upregulated the corresponding targeted genes, and cell proliferation.

miR-132-3p could reduce in vitro tumorigenicity of LUAD cells

A549 cells with ectopically expressed miR-132-3p had more than 85% transfection efficiency at 48 hours. Expression level of miR-132-3p was significantly higher in cells with ectopically expressed miR-132-3p compared with cells with negative controls at 48 hours (Fig 5a). Furthermore, the expression level of ZEB2, a direct target gene of miR-132-3p, was considerably lower in the cells with ectopically expressed miR-132-3p compared with cells with control vector (Fig 5b). In addition, forced expression of miR-132-3p could reduce cell viability compared to cells with

Table 4 Epigenetic interactions between miRNAs and in lung adenocarcinoma

miRNAs	Expression level in tumors	Site ID	Methylation	P-value	Distance	Location in miRNA
mir-1236	Down	cg21754201	Hyper	0.0108	-222	Downstream
		cg25355006	Hyper	0.0040	-94	Downstream
mir-1276	Down	cg04091325	Hyper	0.0060	-702	Downstream
mir-132	Down	cg08457620	Hyper	<0.0001	-179	Downstream
mir-143	Down	cg17736336	Hyper	0.0120	273	Upstream
mir-147b	Over	cg23803468	Hypo	0.0010	-964	Downstream
mir-200a	Over	cg07213830	Hypo	<0.0001	-834	Downstream
		cg14288281	Hypo	0.0010	-351	Downstream
mir-200b	Over	cg20256117	Hypo	0.0120	-988	Downstream
mir-219-1	Down	cg02538046	Hyper	0.0190	-757	Downstream
		cg03000593	Hyper	0.0010	428	Upstream
		cg04678230	Hyper	<0.0001	401	Upstream
		cg07245868	Hyper	0.0010	535	Upstream
		cg08771019	Hyper	0.0030	445	Upstream
		cg10134527	Hyper	<0.0001	575	Upstream
		cg13027595	Hyper	<0.0001	406	Upstream
		cg14096569	Hyper	<0.0001	526	Upstream
		cg14258935	Hyper	0.0080	304	Upstream
		cg14288848	Hyper	0.0230	717	Upstream
		cg18144560	Hyper	<0.0001	520	Upstream
		cg20185718	Hyper	0.0220	273	Upstream
		cg21330831	Hyper	<0.0001	498	Upstream
		cg25954512	Hyper	<0.0001	569	Upstream
		cg26646118	Hyper	<0.0001	590	Upstream
		cg27168291	Hyper	<0.0001	593	Upstream
		cg27368379	Hyper	0.0020	262	Upstream
		cg12195211	Hyper	0.005	23	Upstream
mir-339	Down	cg01717376	Hypo	<0.0001	-234	Downstream
mir-375	Over	cg07718444	Hypo	<0.0001	168	Upstream
mir-429	Over	cg11618529	Hyper	0.0300	-301	Downstream
mir-516a-1	Over	cg12646585	Hypo	0.0080	38	Upstream
mir-548d-1	Over	cg14018100	Hypo	0.0030	545	Upstream
		cg22123387	Hypo	0.0190	-272	Downstream
mir-590	Over	cg13210403	Hypo	0.0080	885	Upstream
mir-625	Over	cg19619721	Hypo	0.0170	512	Upstream
		cg02928928	Hypo	<0.0001	-739	Downstream
mir-675	Over	cg07435282	Hypo	0.0030	-671	Downstream
		cg14836313	Hypo	0.0250	-995	Downstream
		cg19273253	Hypo	0.0070	433	Upstream
		cg19837124	Hypo	<0.0001	398	Upstream
mir-744	Down	cg25237720	Hyper	0.0010	900	Upstream
mir-877	Down	cg16979445	Hyper	0.0040	445	Upstream
mir-9-2	Over	cg17224769	Hypo	0.0020	526	Upstream
mir-92b	Over	cg02892624	Hypo	0.0030	-536	Downstream
		cg17421062	Hypo	0.0040	-462	Downstream
mir-937	Over	cg10817223	Hypo	0.0020	169	Upstream
mir-940	Down	cg07032866	Hyper	0.0050	979	Upstream

control vector (Fig 5c). The cells with ectopic expression of miR-132-3p had a lower cell proliferation comprised with cells with negative control vector (Fig 5d). Moreover, forced expression of miR-132-3p in lung cancer decreased their colony formation, migration, and invasion (Fig 5e–g). Finally, forced expression of miR-132-3p induced the percentage of cells in the G1/G0 phase by more than 10%, whereas it decreased the percentage of S phase at least 10%

compared with cancer cells with a lower level of the miRNA (all $P \leq 0.05$) (Fig 5h).

miR-132-3p could reduce in vivo tumorigenicity of LUAD cells

Tumors were visible as early as six days in the mice that received negative control vector cancer cells. On 15 days

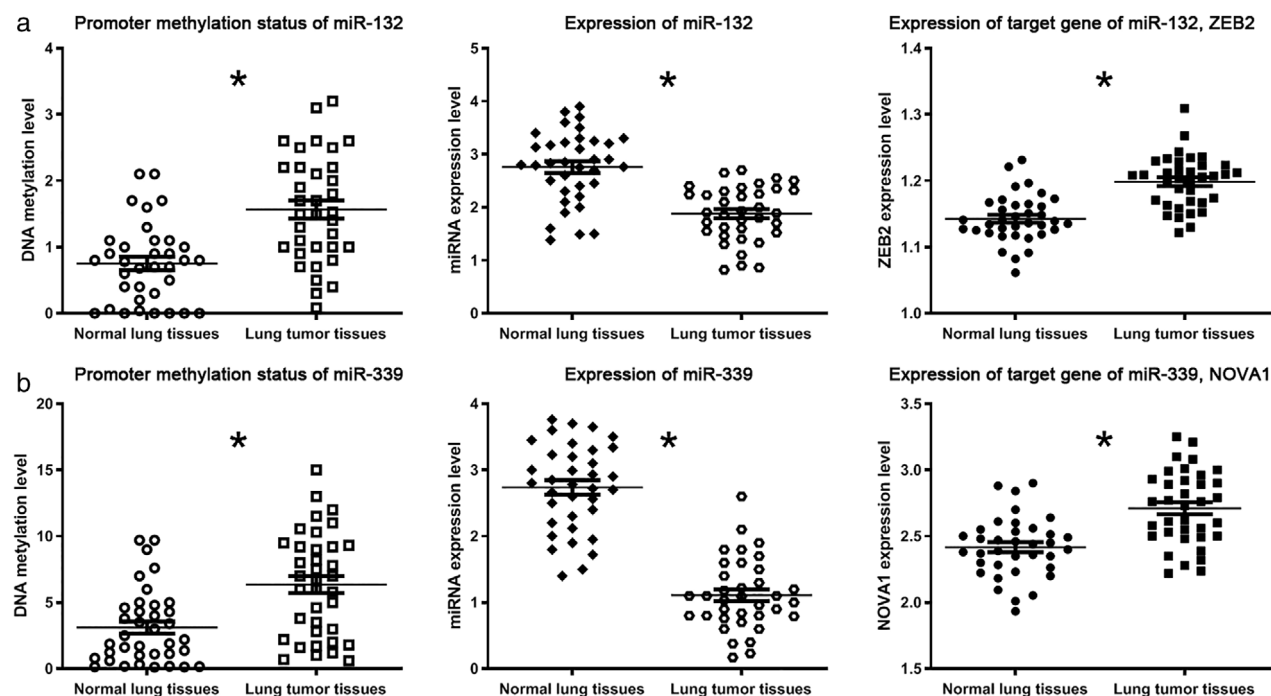
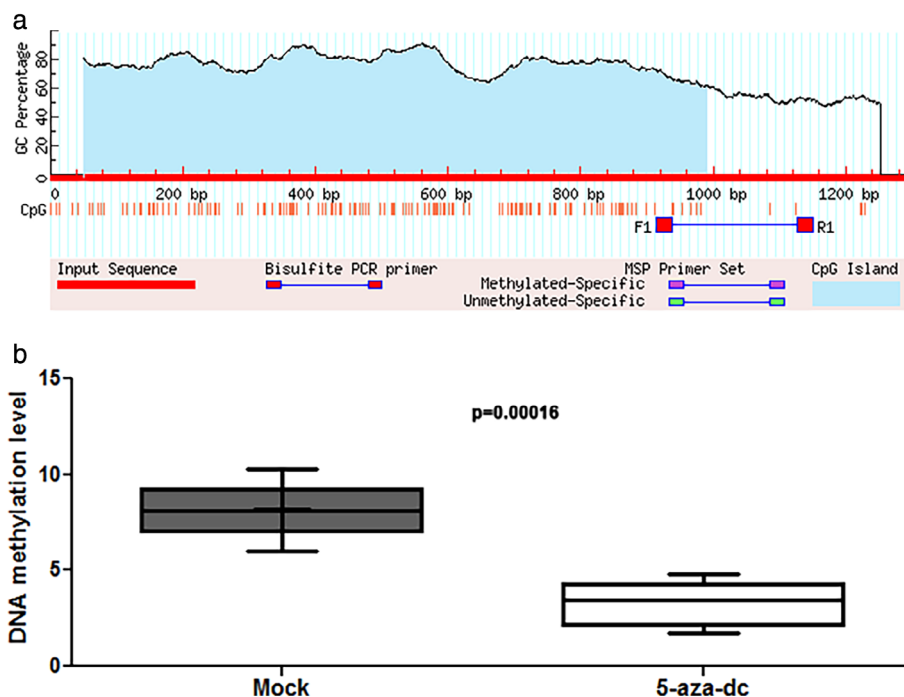


Figure 2 Quantitative methylation specific PCR analysis of methylated loci of miRs-132 and 339 and RT-qPCR analysis of the miRNAs in the 36 lung adenocarcinoma tissues and 36 normal lung tissues. (a) miR-132-3p was significantly hypermethylated and downregulated in lung adenocarcinoma tissues compared with the matched normal lung tissues. However, the target gene of miR-132-3p, ZEB2, was overexpressed in the paired lung adenocarcinoma tissues versus normal lung tissues. (b) miR-339 was considerably hypermethylated and downregulated in lung adenocarcinoma tissues versus the matched normal lung tissues. However, the target gene of miR-339, NOVA1, was overexpressed in the paired lung adenocarcinoma tissues versus normal lung tissues (*all $P < 0.05$).

Figure 3 In silico prediction of CpG islands of miR-132-3p and the DNA methylation levels in LUAD cells treated with or without 5-Aza-dC. (a) CpG islands in miR-132-3p and location of the primers for qMSP. (b) the A549 cancer cells treated with 5-Aza-dC had a low methylation level compared with the cells without the treatment (* $P = 0.00016$).



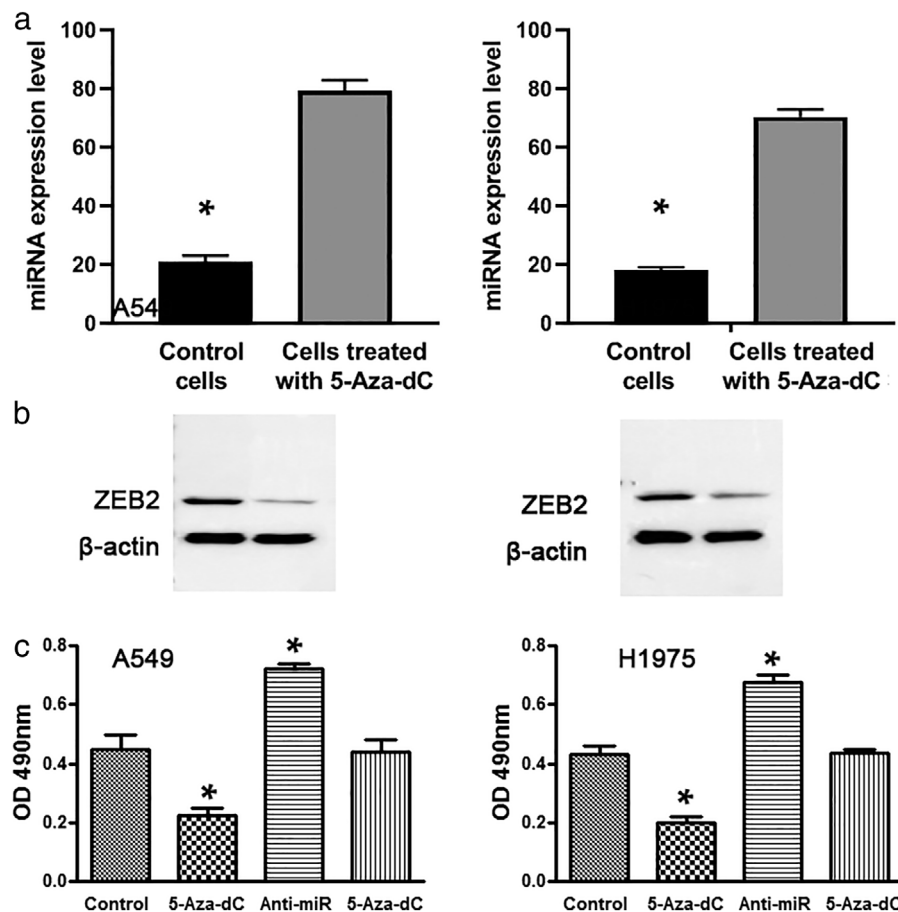


Figure 4 The treatment with 5-Aza-dC increased the expression level of miR-132-3p and decreased the level of ZEB2 protein and cell proliferation in two lung cancer cell lines, A549 and H1975. The cells were treated with 5 μ M 5-Aza-dC. Each sample was subjected to (a) qRT-PCR for detection of miR-132-3p. (b) Western blot for protein expression of ZEB2. (c) Cell proliferation was performed on the cells treated with 5-Aza-dC or miR-132-3p inhibitor or their combination for 48 hours, respectively. Cell proliferation was reduced or increased in the cancer cells treated with 5-Aza-dC or miR-132-3p, respectively, compared with cancer cells treated with PBS (*, all $P \leq 0.05$). However, the effect of 5-Aza-dC on cell proliferation was counteracted by miR-132-3p inhibitor.

post-injection, tumors appeared in all the five mice injected with the A549 cells with negative control vector, whereas they were observed in only two of the five mice injected with ectopically expressed miR-132-3p. Furthermore, the tumors generated from the A549 cells with negative control vector were significantly larger compared to those produced from the cancer cells with ectopically expressed miR-132-3p at the end of observation (28 days) ($180.33 \pm 21.36 \text{ mm}^3$ vs. $68.67 \pm 13.01 \text{ mm}^3$, $P = 0.002$) (Fig 5i). These in vivo findings in ectopic xenograft mouse models are consistent with the in vitro observations, and hence support that miR-132-3p could function as a tumor suppressor tumorigenicity of lung cancer.

Bioinformatic molecular and cellular functional and pathway analyses

To better understand the biological function of the relationships between the two epigenetic events in lung carcinogenesis, we used Ingenuity Pathway Analysis (IPA) to identify possible regulatory networks, by which the

miRNAs with altered DNA methylation might be involved in LUAD. The 22 miRNAs with abnormal hypermethylation were mostly associated with molecular and cellular functions, including cell-to-cell signaling and interaction, cellular compromise, cellular development, cellular growth and proliferation, and cell cycle (Table 5). Furthermore, the highest enriched pathways of the miRNAs were IL-17A signaling in airway cells, PEDF signaling, Wnt/ β -catenin signaling, and JAK family kinases in IL-6-type cytokine signaling, and role of macrophages, fibroblasts and endothelial cells in rheumatoid arthritis (Table 5).

The 98 differentially methylated loci, which were independent of the DNA methylation/miRNA/mRNA (protein) axis, were assigned to molecular and cellular functions, cellular development, cellular growth and proliferation, cellular function and maintenance, and cell morphology (Table 5). Furthermore, the differentially methylated loci were highly associated with canonical pathways of glutamate receptor signaling, CREB signaling in neurons, amyotrophic lateral sclerosis signaling, gap junction signaling, and G α i signaling (Table 5).

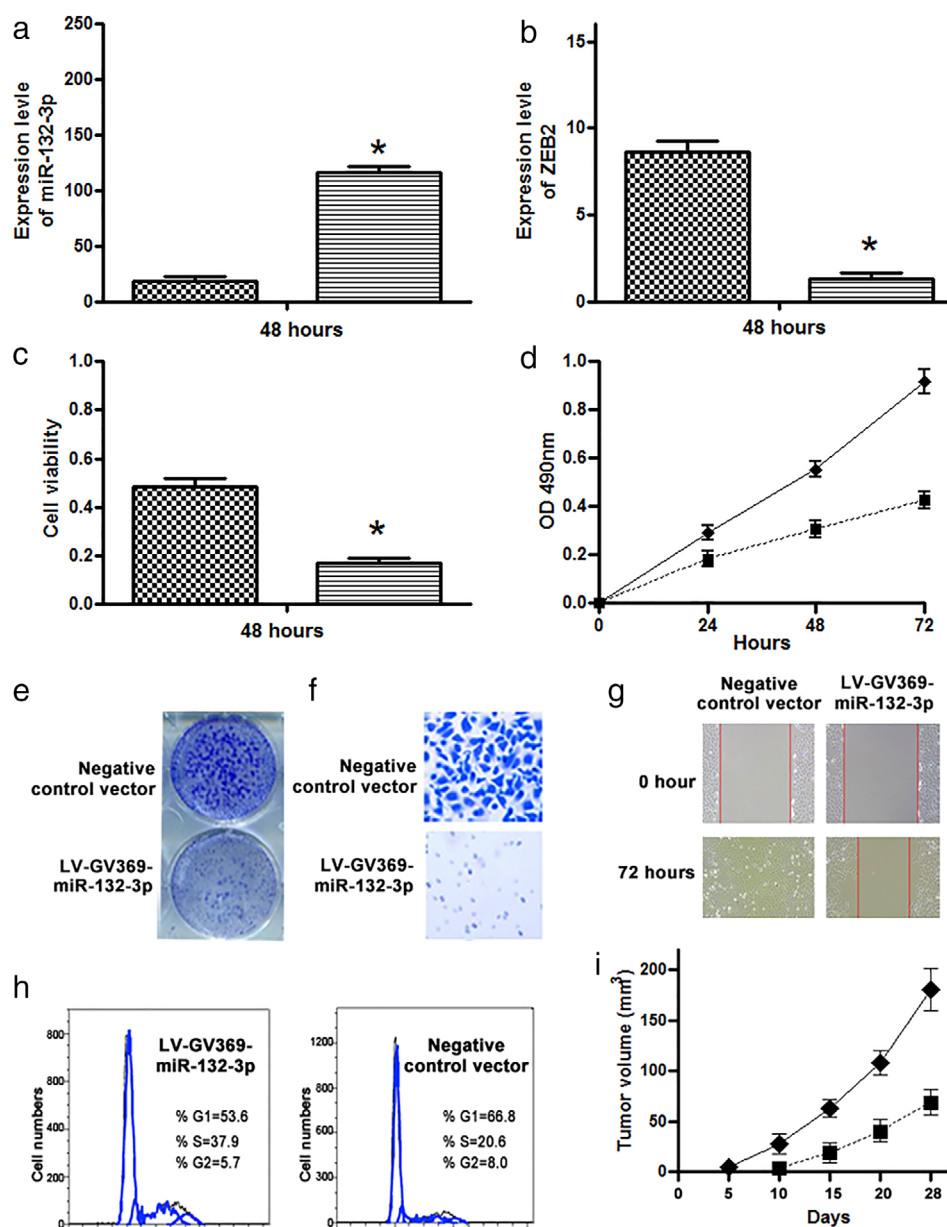


Figure 5 Forced expression of miR-132-3p could inhibit the tumorigenicity of A549 lung cancer cells. (a) Expression level of miR-132-3p was significantly higher in cells with ectopically expressed miR-132-3p versus negative controls at 48 hours. (b) Ectopic expression of miR-132-3p reduced the expression level of ZEB2, a direct target gene of miR-132-3p. (c) Forced expression of miR-132 can significantly reduce cell viability. (d) Cells with ectopic expression of miR-132-3p had a lower cell proliferation comprised with cells with negative control vector. (e) Forced expression of miR-132-3p decreased cell colony formation. (f) Transwell migration assay shows that forced expression of miR-132-3p could constrain migration and invasion of A549 cancer cells. (g) In the wound-healing assays, cancer cells with forced expression of miR-132-3p showed a slower gap closure compared with cells transfected with control vector. The figure only shows the results of A549 cells from the time points 0 and 72 hours, respectively. (h) Forced expression of miR-132-3p in lung cancer cells elevated the percentage of the cancer cells in the G1/G0 phase and reduced the percentage of S phase (*all $P < 0.05$). (i) Measurement of the tumor size over the time in vivo experiment shows that tumor growth in the mice injected with cancer cells with forced expression of miR-132-3p was significantly lower than in the mice injected with the cells transfected with control vector. Tumor sizes (mean $\pm$ SD³) were measured at the indicated intervals and plotted. (a) (▨) Negative controls vector, (▤) LV-GV369-miR-132-3p; (b) (▨) Negative controls vector, (▤) LV-GV369-miR-132-3p; (c) (▨) Negative controls vector, (▤) LV-GV369-miR-132-3p; (d) (—◆—) Negative controls vector, (—■—) LV-GV369-miR-132-3p; (i) (—◆—) Negative controls vector, (—■—) LV-GV369-miR-132-3p.

Table 5 Molecular and cellular functional and pathway analyses

	Molecular and cellular functions	P-value	Canonical pathways	P-value
The 22 miRNAs with abnormal hypermethylation	1) Cell-to-cell signaling and interaction	0.00927	1) IL-17A signaling in airway cells	0.00155
	2) Cellular compromise	0.0135	2) PEDF signaling	0.00315
	3) Cellular development	0.0018	3) Wnt/ β -catenin signaling	0.00342
	4) Cellular growth and proliferation	0.0036	4) Role of JAK family kinases in IL-6-type cytokine signaling	0.00357
	5) Cell cycle	0.0238	5) Role of macrophages, fibroblasts and endothelial cells in rheumatoid arthritis	5.86E-05
The 103 differentially expressed miRNAs, which are not associated with the differentially methylated loci	1) Cellular movement	2.33E-06	1) Glucocorticoid receptor signaling	8.59E-06
	2) Cellular assembly and organization	6.56E-06	2) Glioblastoma multiforme signaling	1.38E-05
	3) Cellular function and maintenance	3.23E-05	3) Hepatic fibrosis/hepatic stellate cell activation	2.53E-05
	4) Cellular development	7.66E-05	4) Myc mediated apoptosis signaling	5.86E-05
	5) Cellular growth and proliferation	0.000119	5) Estrogen-mediated S-phase entry	6.14E-05
The 98 differentially methylated loci, differentially expressed miRNAs	1) Cellular development	0.000402	1) Glutamate receptor signaling	0.000227
	2) Cellular growth and proliferation	0.00228	2) CREB signaling in neurons	0.000783
	3) Cellular function and maintenance	0.00271	3) Amyotrophic lateral sclerosis signaling	0.0017
	4) Cell morphology	0.00269	4) Gap junction signaling	0.0039
			5) Gai signaling	0.00425

The 103 differentially expressed miRNAs not epigenetically regulated by DNA methylation, mainly have molecular and cellular functions of cellular movement, cellular assembly and organization, cellular function and maintenance, cellular development, and cellular growth and proliferation (Table 5). The differentially expressed miRNAs might play important roles in LUAD through pathways of glucocorticoid receptor signaling, glioblastoma multiforme signaling, hepatic fibrosis/hepatic stellate cell activation, myc mediated apoptosis signaling, and estrogen-mediated s-phase entry (Table 5).

Therefore, there might be three forms of relationships between the two epigenetic events in tumorigenesis of LUAD: (i) synergetic interaction of microRNAs and DNA methylation contribute to the tumorigenesis by the DNA methylation/miRNA/mRNA (protein) axis. (ii) dysregulation of miRNAs; or (iii) abnormal DNA methylation alone might be involved in LUAD development and progression via different pathways, respectively.

Discussion

Investigating the relationship of the DNA methylation and altered expressions of miRNAs could improve our understanding of lung tumorigenesis, and hence provide potential diagnostic biomarkers for the malignancy.²⁰ For instance, determination of miRNA gene methylation status

could be used as a molecular biomarker to detect cancer at an early stage and predict cancer prognosis.²¹ Furthermore, epigenetic alterations could be reversible, making epigenetic regulation attractive in terms of developing new therapeutic approaches.²² Since miRNAs can be inactivated by epigenetic mechanisms, drug-induced reduction of DNA methylation might be a promising therapeutic strategy.²³ In this study, by analyzing TCGA-LUAD data, we identified 125 differentially expressed miRNAs and 145 differentially methylated loci in lung tumor tissues versus normal lung tissues. Interestingly, direct interaction between 22 miRNAs and 53 methylation loci were found, suggesting that the miRNAs might be epigenetically regulated by DNA methylation in tumorigenesis of LUAD. Furthermore, hypermethylation and reduced expression of certain miRNAs and increased expression of the target genes were validated in additional surgically-resected LUAD tissues.

Downregulation of miR-132-3p has been observed in several human malignancies, including lung cancer.^{24–27} For instance, miR-132-3p has been suggested to suppress migration and invasion of lung cancer cells by targeting ZEB2.²⁶ However, although the studies^{24–26} suggested that dysregulation of miR-132-3p might have important functions in tumorigenesis, how it contributes to lung cancer remains unknown. Here, we confirmed that miR-132-3p was considerably downregulated in lung tumor tissues and

ZEB2 could be a direct and functional target of miR-132-3p in LUAD. More importantly, our data suggested that elevated expression of miR-132-3p, reduced expression of ZEB2, and increased cell proliferation were observed in lung cancer cells treated with DNA methyltransferase inhibitor. Furthermore, forced expressed miR-132-3p in LUAD cells reduced their tumorigenicity in vitro and in vivo. Therefore, altered methylation status of miRNAs could regulate expression of the miRNAs, which in turn upregulates the corresponding targeted genes and cell proliferation. miR-132-3p could act as a tumor suppressor gene whose downregulation contributes to the progression of lung cancer by directly targeting ZEB2. ZEB2 is a transcription factor and plays vital roles in the TGF- β signaling cascade and is involved in multiple cellular functions.²⁸ ZEB2 could act as a transcriptional repressor of E-cadherin and plays a role in epithelial-mesenchymal transition, which leads to specific morphological and phenotypic alterations in tumor cells during cancer metastasis.²⁹ In addition, our molecular and cellular function analysis showed that the abnormally methylated miRNAs were mainly involved in cell-to-cell signaling and interaction in airway cells, and further support that the direct interaction between two epigenetic aberrations could have important functions in LUAD.

We also identified another two relationships between the two molecular events. A total of 98 methylation loci were not associated with altered miRNA expressions, whereas 103 miRNAs had no relation with the differentially methylated loci. Interestingly, molecular and cellular functional and pathway analyses suggested the two additional relationships between the molecular changes might contribute to tumorigenesis of LUAD via different pathways. Therefore, our results might gain a deeper understanding of the role of epigenetics in lung cancer.

Some limitations exist in this study. (i) We only investigated the effect of CpG island methylation on miRNA expression. However, there is a mutual regulation loop between miRNAs and DNA methylation in human malignancies.²⁰ Since a single miRNA can regulate multiple target genes, miRNAs could also regulate DNA methylation by modulating methylation-related critical proteins.²⁰ In a different study, we are investigating the mechanism, by which aberrant DNA methylation could be regulated by dysregulation of miRNAs in lung tumorigenesis. (ii) The focus of this study is LUAD. However, 85% of lung cancers are NSCLCs, which consist of three main subtypes: LUAD, SCC, and LCC. Our ongoing study is to investigate the relationship between the two epigenetic changes in tumorigenesis of SCC and LCC.

In conclusion, the interaction between miRNA and DNA methylation could have important functions in lung

tumorigenesis. miR-132-3p downregulation via DNA methylation could promote tumorigenicity of lung cancer by directly regulating ZEB2. Nevertheless, further investigation of the interaction of the two important epigenetic events is needed to improve our understanding of lung tumorigenesis, and hence provide potential diagnostic and therapeutic targets of lung cancer.

Acknowledgments

We thank the University of Maryland Marlene and Stewart Greenebaum Comprehensive Cancer Center Biostatistics Shared Service to provide statistical services.

Disclosure

The authors declare no conflict of interest.

References

- Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. *Carcinogenesis* 2010; **31**: 27–36.
- Liao J, Shen J, Leng Q, Qin M, Zhan M, Jiang F. MicroRNA-based biomarkers for diagnosis of non-small cell lung cancer (NSCLC). *Thorac Cancer* 2020; **11** (3): 762–8.
- Wang J, Yu XF, OUYang N *et al.* Role of DNA methylation regulation of miR-130b expression in human lung cancer using bioinformatics analysis. *J Toxicol Environ Health A* 2019; **82**: 935–43.
- Liu H, Zhao H. Prognosis related miRNAs, DNA methylation, and epigenetic interactions in lung adenocarcinoma. *Neoplasia* 2019; **66**: 487–93.
- Network TC. Corrigendum: Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 2013; **494**: 506.
- Shivakumar M, Lee Y, Bang L, Garg T, Sohn KA, Kim D. Identification of epigenetic interactions between miRNA and DNA methylation associated with gene expression as potential prognostic markers in bladder cancer. *BMC Med Genomics* 2017; **10**: 30.
- Lizio M, Abugessaisa I, Noguchi S *et al.* Update of the FANTOM web resource: Expansion to provide additional transcriptome atlases. *Nucleic Acids Res* 2019; **47**: D752–8.
- Su Y, Fang H, Jiang F. Integrating DNA methylation and microRNA biomarkers in sputum for lung cancer detection. *Clin Epigenetics* 2016; **8**: 109.
- Su Y, Fang HB, Jiang F. An epigenetic classifier for early stage lung cancer. *Clin Epigenetics* 2018; **10**: 68.
- Gaykalova DA, Vatapalli R, Wei Y *et al.* Outlier analysis defines zinc finger gene family DNA methylation in tumors and saliva of head and neck cancer patients. *PLOS One* 2015; **10**: e0142148.

- 11 Li H, Jiang Z, Leng Q *et al.* A prediction model for distinguishing lung squamous cell carcinoma from adenocarcinoma. *Oncotarget* 2017; **8**: 50704–14.
- 12 Fan T, Li R, Todd NW *et al.* Up-regulation of 14-3-3zeta in lung cancer and its implication as prognostic and therapeutic target. *Cancer Res* 2007; **67**: 7901–6.
- 13 Lin Y, Holden V, Dhilipkannah P, Deepak J, Todd NW, Jiang F. A non-coding RNA landscape of bronchial epitheliums of lung cancer patients. *Biomedicine* 2020; **8** (4): 88.
- 14 Mei YP, Liao JP, Shen J *et al.* Small nucleolar RNA 42 acts as an oncogene in lung tumorigenesis. *Oncogene* 2012; **31**: 2794–804.
- 15 Pang W, Tian X, Bai F *et al.* Pim-1 kinase is a target of miR-486-5p and eukaryotic translation initiation factor 4E, and plays a critical role in lung cancer. *Mol Cancer* 2014; **13**: 240.
- 16 Wang J, Tian X, Han R *et al.* Downregulation of miR-486-5p contributes to tumor progression and metastasis by targeting protumorigenic ARHGAP5 in lung cancer. *Oncogene* 2014; **33**: 1181–9.
- 17 Jiang F, Qiu Q, Khanna A *et al.* Aldehyde dehydrogenase 1 is a tumor stem cell-associated marker in lung cancer. *Mol Cancer Res* 2009; **7**: 330–8.
- 18 Mannoor K, Shen J, Liao J, Liu Z, Jiang F. Small nucleolar RNA signatures of lung tumor-initiating cells. *Mol Cancer* 2014; **13**: 104.
- 19 Yin L, Chang C, Xu C. G2/M checkpoint plays a vital role at the early stage of HCC by analysis of key pathways and genes. *Oncotarget* 2017; **8**: 76305–17.
- 20 Wang S, Wu W, Claret FX. Mutual regulation of microRNAs and DNA methylation in human cancers. *Epigenetics* 2017; **12**: 187–97.
- 21 Konno M, Koseki J, Asai A *et al.* Distinct methylation levels of mature microRNAs in gastrointestinal cancers. *Nat Commun* 2019; **10**: 3888.
- 22 Ellis L, Atadja PW, Johnstone RW. Epigenetics in cancer: Targeting chromatin modifications. *Mol Cancer Ther* 2009; **8**: 1409–20.
- 23 Nasr MA, Salah RA, Abd Elkodous M, Elshenawy SE, El-Badri N. Dysregulated microRNA fingerprints and methylation patterns in hepatocellular carcinoma, cancer stem cells, and mesenchymal stem cells. *Front Cell Dev Biol* 2019; **7**: 229.
- 24 Guo H, Zhang X, Chen Q, Bao Y, Dong C, Wang X. miR-132 suppresses the migration and invasion of lung cancer cells by blocking USP9X-induced epithelial-mesenchymal transition. *Am J Transl Res* 2018; **10**: 224–34.
- 25 Li Y, Zu L, Wang Y, Wang M, Chen P, Zhou Q. miR-132 inhibits lung cancer cell migration and invasion by targeting SOX4. *J Thorac Dis* 2015; **7**: 1563–9.
- 26 You J, Li Y, Fang N *et al.* MiR-132 suppresses the migration and invasion of lung cancer cells via targeting the EMT regulator ZEB2. *PLOS One* 2014; **9**: e91827.
- 27 Zhang JX, Zhai JF, Yang XT, Wang J. MicroRNA-132 inhibits migration, invasion and epithelial-mesenchymal transition by regulating TGFbeta1/Smad2 in human non-small cell lung cancer. *Eur Rev Med Pharmacol Sci* 2016; **20**: 3793–801.
- 28 Dai YH, Tang YP, Zhu HY *et al.* ZEB2 promotes the metastasis of gastric cancer and modulates epithelial mesenchymal transition of gastric cancer cells. *Dig Dis Sci* 2012; **57**: 1253–60.
- 29 Li MZ, Wang JJ, Yang SB *et al.* ZEB2 promotes tumor metastasis and correlates with poor prognosis of human colorectal cancer. *Am J Transl Res* 2017; **9**: 2838–51.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Table S1. Differentially expressed miRNAs between lung adenocarcinoma tissues versus normal tissues.

Table S2. The differentially methylated loci in lung adenocarcinoma tissues versus normal lung tissues.

Table S3. Differentially expressed miRNAs between lung adenocarcinoma tissues and normal tissues.

Table S4 The differentially methylated loci were not associated with expression of the miRNAs.

ORIGINAL ARTICLE

MicroRNA-based biomarkers for diagnosis of non-small cell lung cancer (NSCLC)

Jipei Liao¹, Jun Shen¹, Qixin Leng¹, Meng Qin¹, Min Zhan² & Feng Jiang¹ ¹ Department of Pathology, University of Maryland School of Medicine, Baltimore, Maryland, USA² Departments of Epidemiology & Public Health, University of Maryland School of Medicine, Baltimore, Maryland, USA**Keywords**

Biomarkers; diagnosis; lung cancer; microRNA; plasma; sputum.

Correspondence

Feng Jiang, Department of Pathology, The University of Maryland School of Medicine, 10 South Pine Street, MSTF 7th floor, Baltimore, MD 21201-1192, USA.
Tel: +01 410 706 4854
Fax: +01 410 706 7717
Email: fjiang@som.umaryland.edu

Received: 13 December 2019;

Accepted: 12 January 2020.

doi: 10.1111/1759-7714.13337

Thoracic Cancer **11** (2020) 762–768**Abstract**

Background: The development of biomarkers for the early detection of non-small cell lung cancer (NSCLC) is clinically important. We have developed miRNA biomarkers in sputum and plasma, respectively, for NSCLC. Herein, we evaluate whether integrated analysis of the miRNAs across the different types of specimens could improve the early detection of NSCLC.

Methods: Using reverse transcription PCR, we determined expressions of two miRNAs (miRs-31-5p and 210-3p) in sputum and three miRNAs (miRs-21-5p, 210-3p, and 486-5p) in plasma of a training cohort of 76 NSCLC patients and 72 cancer-free smokers. The results were validated in a testing cohort of 56 NSCLC patients and 55 cancer-free smokers.

Results: The panels of two sputum miRNAs and three plasma miRNAs had 65.8–75.0% sensitivities and 83.3–87.5% specificities for diagnosis of NSCLC in the training cohort. The individual sputum or plasma miRNA panel had a higher sensitivity for squamous cell carcinoma or adenocarcinoma of the lung, respectively. From the miRNAs, we optimized an integrated panel of biomarkers consisting of two sputum miRNAs (miRs-31-5p and 210-3p) and one plasma miRNA (miR-21-5p) that had higher sensitivity (85.5%) and specificity (91.7%) for diagnosis of NSCLC compared with the individual panels alone. Furthermore, the performance of the integrated panel of biomarkers was independent of histology and stage of NSCLC, and patients' age, sex, and ethnicity. The performance of the integrated panel of biomarkers was confirmed in the testing cohort.

Conclusions: Integrating biomarkers across different body fluids would synergistically improve the early detection of NSCLC.

Key points

- Lung cancer is a heterogeneous disease and develops from complex aberrations.
- Integrating sputum and plasma miRNAs has higher accuracy than when they are used alone

Introduction

A total of 85% of lung cancers are non-small cell lung cancer (NSCLC), which mainly consists of two histological types: adenocarcinoma (AC) and squamous cell carcinoma (SCC).¹ The National Lung Screening Trial (NLST)

shows that low-dose CT (LDCT) could detect NSCLC at the early stage and significantly reduce mortality.² However, LDCT has a high false-positive rate (or a low specificity) in the early detection of lung cancer. Therefore, the development of body fluid-based biomarkers for precisely

diagnosing lung cancer at an early stage is clinically important.

It is well accepted that molecular changes in epithelial cells collected from the normal-appearing mainstem bronchus of smokers could be developed as biomarkers for NSCLC.³ Sputum contains bronchial epithelial cells from the lungs and lower respiratory tract. Therefore, examination of the exfoliated bronchial airway epithelial cells in sputum can identify the lung tumor-related molecular alterations, and thus provide a noninvasive and specific tool for diagnosis of NSCLC.³ MicroRNAs (miRNAs) are a class of small noncoding RNAs of ~22nt in length.⁴ Dysregulations of miRNAs can drive multiple processes of tumorigenesis by regulating cell cycle, apoptosis, and migration.⁴ For example, we have shown that down-regulations of miR-486-5p contribute to the development of NSCLC by regulating Pim-1 and ARHGAP5.^{5–8} We and others have also demonstrated that aberrant miRNA expressions detected in sputum biologically reflect those in primary lung cancer.^{9–15} We have recently developed a panel of two sputum miRNA (miR-31-5p and miR-210-3p) biomarkers for NSCLC with 65% sensitivity and 89% specificity.¹⁶ Furthermore, efforts have been made to develop cell-free (eg, plasma) biomarkers by detecting the circulating molecules directly released from lung tumors. Plasma miRNAs released from primary lung tumors or the circulating cancer cells might also provide potential biomarkers for NSCLC. We have developed a panel of three plasma miRNA (miRs-21-5p, 210-3p, and 486-5p) biomarkers that have 75% sensitivity and 85% specificity for diagnosis of NSCLC.¹⁷

Lung cancer is a heterogeneous disease and develops from complex molecular aberrations. The objective of this study was to evaluate whether integrated analysis of the miRNAs in sputum and plasma could improve the early detection of NSCLC.

Methods

Specimen collection and preparation

The Institutional Review Boards (IRB) at University of Maryland Baltimore approved this study. Written informed consent was obtained from all participants. Sputum and blood samples were collected from the participants before they had any treatment as previously described.^{7,9,16,18–26} The preparation of airway bronchial epitheliums from sputum was performed using a protocol developed in previous studies.^{7,9,16,18–23,25,26} Plasma was prepared from blood by using the standard operating protocols developed by The NCI-Early Detection Research Network.²⁷

Quantitative reverse transcription PCR (qRT-PCR) for determination of miRNA expression levels

RNA was extracted from the clinical specimens by using our established protocols.^{9,16} The expressions of two miRNAs (miRs-31-5p and 210-3p) in sputum, and three miRNAs (miRs-21-5p, 210-3p, and 486-5p) in plasma were determined by using Taqman miRNA assays (Applied Biosystems, Foster City, CA).^{7,9,16,19–21,28} The levels of miRNAs were measured by using a threshold cycle method with U6 as an internal control to normalize the expression of each gene.^{7,9,16,21,29–31} We included controls in each experiment: RNA extracted from a H358 NSCLC cell line (a positive control for RT, preamplification and ddPCR), genomic DNA (a positive control for genomic DNA detection), and nuclease-free water (control for contamination). All experiments were repeated three times.

Statistical analysis

We expected that the study had $\geq 90\%$ power at ≤ 1 false positives to detect miRNAs with fold change ≥ 2 between cancer patients and control subjects. To achieve the statistical criteria, ≥ 55 specimens from each group were required. Statistical analysis of RT-qPCR data was done using Statistical Analysis System software version 6.12 (SAS Institute, Cary, NC). All *P*-values were two-sided, and a *P*-value of < 0.05 was considered statistically significant. Furthermore, contingency table and logistic regression analysis were applied to determine the associations between the expression levels of the miRNAs and both clinicopathologic and demographic characteristics of the cases and controls. In addition, we analyzed the significantly associated miRNAs by using logistic regression models with constrained parameters as least absolute shrinkage and selection operator (LASSO) to eliminate the less relevant variables.³² We estimated functions of the combined miRNA biomarkers by logistic regression with or without adjustment for known risk factors for NSCLC. Receiver-operator characteristic (ROC) curve analysis was undertaken using expression level for each miRNA from cancer patients and cancer-free controls by Analyze-it software (Analyze-it Software, Leeds, UK). We also generated a 95% confidence interval for the difference in the area under the ROCs (AUCs) by using the bootstrap. We established the optimal cutoff value by using the Youden index. The results of the training cohort were blindly validated in the testing cohort by using leave-one-out cross validation. Moreover, to compare different miRNA biomarker panels, we computed their AUCs to determine the sensitivity and specificity as previously described.¹⁷

Results

Patients

We enrolled patients between the ages of 55–80 who had $\geq$ a 30 pack-year smoking history and were former smokers (quit within 15 years). Exclusion criteria included pregnancy, current pulmonary infection, surgery within six months, radiotherapy within one year, and life expectancy of $<$ one year. A total of 132 NSCLC patients and 127 cancer-free smokers were recruited. No difference of age, race, gender and smoking status was observed in the lung cancer cases versus cancer-free smokers. The cases and controls were randomly split into two cohorts: a training cohort and a testing cohort. The training cohort comprised 76 lung cancer patients and 72 cancer-free smokers. The training cohort was used to simultaneously analyze sputum and plasma miRNAs, and hence develop an integrated panel of miRNA biomarkers for NSCLC. The testing cohort included 56 lung cancer patients and 55 cancer-free smokers and was used for blindly confirming the results generated from the training cohort. Detailed demographic and clinical characteristics of the two cohorts are shown in Tables 1 and 2.

Individual sputum miRNAs and plasma miRNAs could distinguish lung cancer patients from cancer-free controls

In the training cohort, the two sputum miRNAs (miRs-31-5p and 210-3p) and three miRNAs (miRs-21-5p, 210-3p, and 486-5p) exhibited a higher expression level in

Table 1 Characteristics of NSCLC patients and cancer-free smokers in a training cohort

	NSCLC cases (<i>n</i> = 76)	Controls (<i>n</i> = 72)	<i>P</i> -value
Age	67.45 (SD 12.35)	66.63 (SD 10.37)	0.33
Sex			0.45
Female	27	25	
Male	49	37	
Race			0.43
White	55	52	
African American	21	20	
Pack-years (median)	35.46	35.22	0.38
Stage			
Stage I	25		
Stage II	23		
Stage III–IV	28		
Histological type			
Adenocarcinoma	43		
Squamous cell carcinoma	33		

NSCLC, non-small cell lung cancer.

Table 2 Characteristics of NSCLC patients and cancer-free smokers in a testing cohort

	NSCLC cases (<i>n</i> = 56)	Controls (<i>n</i> = 55)	<i>P</i> -value
Age	66.87 (SD 11.36)	66.06 (SD 10.78)	0.39
Sex			0.40
Female	20	19	
Male	36	36	
Race			0.45
White	41	39	
African American	15	16	
Pack-years (median)	35.68	35.13	0.38
Stage			
Stage I	18		
Stage II	17		
Stage III–IV	21		
Histological type			
Adenocarcinoma	31		
Squamous cell carcinoma	25		

NSCLC, non-small cell lung cancer.

NSCLC patients versus cancer-free smokers (all $P < 0.001$) (Table S1). Combined analysis of the two sputum miRNAs as a panel produced 0.82 AUC with 65.8% sensitivity and 87.5% specificity for diagnosis of NSCLC (Table 3). The expression level of both miR-21-5p and miR-210-3p in sputum was correlated to SCC ($P < 0.05$). Therefore, the sputum miRNAs showed a higher sensitivity for diagnosis of SCC compared with AC (78.8% vs. 55.8%, $P = 0.02$), while having the same specificity (87.5%). However, the sputum miRNAs did not have a specific relationship with stage of lung cancer, and patients' age, race, and gender (All $P > 0.05$), but smoking history ($P = 0.04$).

The analysis of the three plasma miRNAs had an AUC of 0.85, producing 75.0% sensitivity and 83.3% specificity for diagnosis of NSCLC (Table 3). The expression levels of the plasma miRNAs were related with AC. As a result, combined use of the three plasma miRNAs created a higher sensitivity (81.4.8% vs. 66.7%, $P = 0.03$) and the same specificity (83.3%) for diagnosis of AC vs. SCC. The plasma biomarkers did not have relationship with stage of NSCLC, and patients' age, sex, ethnicity, and smoking history of the participants (All > 0.05).

Integrated analysis of two miRNA biomarkers and one plasma miRNA biomarker has a synergistic effect on the early detection of NSCLC

From the multiple sputum and plasma miRNA biomarkers, we used logistic regression models with constrained

Table 3 Comparison of the integrated panel of two sputum and one plasma miRNAs with individual panels of sputum and plasma miRNAs in a training cohort

	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Two sputum miRNAs	0.816 (0.756–0.882)	65.79% (54.01%–76.29%)	87.50% (77.59%–94.12%)
The three plasma miRNAs	0.852 (0.779–0.902)	75.00% (63.74%–84.23%)	83.33% (72.70%–91.08%)
Integrated 2 sputum and 1 plasma miRNAs	0.913 (0.864–0.956)	85.53% (75.58%–92.55%)	91.67% (82.74%–96.88%)

AUC, the area under receiver operating characteristic curve; CI, confidence interval.

parameters as in LASSO to identify and optimize biomarkers. Two sputum miRNAs (miRs-31-5p and 210-3p) and one plasma miRNA (miRNA-21-5p) were finally selected as the best and incorporated into an algorithm, producing 0.91 AUC for distinguishing lung cancer from cancer free subjects (Figure S1). $U_1 = 18.847 + 2.812 \times \log(\text{miR} - 205 - 5p/u6) - 5.608 \times \log(\text{miR145}/U6) + 5.386 \times \log(\text{miR} - 210/U6) - 7.734 \times \log(\text{miR126}/U6)$

Integrated analysis of the three biomarkers by using this algorithm could produce a greater AUC than the individual sputum biomarker panel or plasma biomarker panel (0.91 vs 0.82 and 0.85, all $P < 0.05$) (Table 3). Using Youden's index, we set the optimal cutoff at 3.52. Integrated use of the two sputum miRNAs and one plasma miRNA produced higher sensitivity (85.5%) and specificity (91.7%) than did the sputum and plasma biomarkers (all $P < 0.05$) (Table 3). Furthermore, combined use of all the five biomarkers (two miRNA and three plasma biomarkers) did not produce higher sensitivity and specificity compared with the integrated panel of the three biomarkers ($P > 0.05$). In addition, the estimated Pearson correlation among levels of the three miRNAs was very low (all $P > 0.05$), suggesting that the integration had

complementary classification for diagnosis of lung cancer. Moreover, in contrast to either the sputum or plasma miRNA biomarker panel used alone, integrated analysis of the three biomarkers across the different body fluids did not show a special association with histology of lung cancer. In addition, the integrated panel of biomarkers had no association with stage of lung cancer, and patients' age, race, and gender (all $P > 0.05$).

Validating the synergistic effect of the integrated panel of three miRNAs across the body fluids for diagnosis of NSCLC in the testing cohort

The two sputum miRNAs (miRs-31-5p and 210-3p) and one plasma miRNA (miR-21-5p) were analyzed in the specimens of 56 lung cancer patients and 55 cancer-free smokers. The integration of the three biomarkers across sputum and plasma showed higher sensitivity (83.9% vs. 64.3% and 75.0%) and specificity (90.9% vs. 87.3% and 83.6%) than did the individual sputum miRNA panel and plasma miRNA panel (all $P < 0.05$) (Fig 1). Furthermore, integrated analysis of the three biomarkers did not display a particular relationship with histological types and stage of NSCLC, and patients' age, race, and gender (all $P > 0.05$), supporting the synergistic value of this integrated panel of biomarkers for the early detection of lung cancer.

Discussion

Most lung SCCs are located centrally, usually in the larger bronchi that join the trachea to the lung, whereas ACs often arise in peripheral lung tissue and originate from the smaller airways of the lungs. Sputum and plasma have different characteristics as surrogate material for molecular diagnosis of NSCLC. In particular, sputum miRNAs provide cell-based biomarkers that are developed from molecular changes of the exfoliated bronchial epitheliums of large airways or main bronchial where SCC more commonly exists. In contrast, cell-free plasma miRNA biomarkers are based on detecting the circulating molecules directly released from lung tumors and the floating cancer cells. Therefore, sputum biomarkers could be more

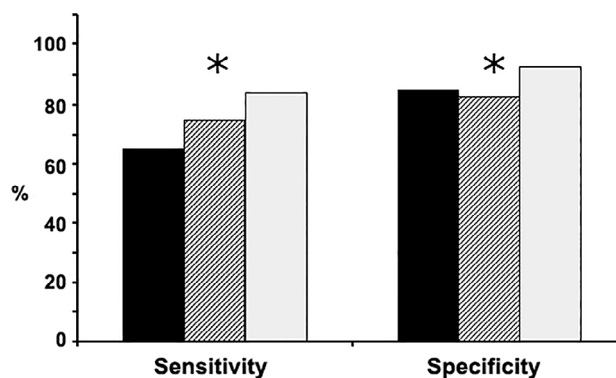


Figure 1 Compares the integrated panel of the three biomarkers with the individual sputum miRNA panel and plasma miRNA panel for lung cancer diagnosis in the testing cohort. The integration of the three biomarkers across sputum and plasma had higher sensitivity and specificity compared with the individual sputum miRNA panel and plasma miRNA panel. * $P < 0.05$. ■ Sputum biomarkers, ▨ plasma biomarkers, □ integrated sputum and plasma biomarkers.

sensitive for SCCs, whereas plasma biomarkers might be more sensitive for ACs. Given that altered miRNAs in sputum and plasmas contribute to lung tumorigenesis via the different mechanisms, integrating the miRNAs across the different body fluids may have a synergetic effect. Indeed, here we demonstrate that the combined analysis of the two sputum (cell-based) miRNA biomarkers and one plasma (cell-free circulating) miRNA biomarkers generates a better performance than did a single category of the miRNA biomarkers. Furthermore, no statistical association exists among the two sputum miRNAs and one plasma miRNA, further supporting that the integration of the two types of miRNA biomarkers might have a synergistic value. Through a complementary manner, integrating sputum miRNAs with plasma miRNAs could diagnose lung cancer, irrespective of the histology and stage. This finding is consistent with our previous discovery,³³ in which combined analysis of miRNA expressions in circulating peripheral blood mononucleated cells and sputum might have a synergistic effect for lung cancer detection. Moreover, integrated analysis of the two sputum and one plasma biomarkers does not show a statistical relationship with patients' age, race, and gender. The findings would be clinically important if the integrated biomarker panel is used for identifying early stage lung cancer.

Dysregulations of miRs-21-5p, 31-5p, and 210-3p have been reported to be associated with a variety of malignancies, including NSCLC.^{34–36} For example, the miR-21 family has an important oncogenic role, whose overexpression contributes to tumor growth and metastasis, and is associated with poor prognosis and chemosensitive tumors.³⁴ Furthermore, anti-miR-21 has been demonstrated to suppress cell growth of breast cancer through downregulation of the antiapoptotic factor, B-cell lymphoma 2 (Bcl-2). miR-210-3p has been reported to stimulate a hypoxic phenotype and upsurge radioresistance in NSCLCs.³⁵ Hypoxia-induced miR-210-3p can regulate tumor cell susceptibility to cytolytic T-lymphocyte-mediated lysis by a mechanism involving its downstream targets PTPN1, HOXA1, and TP53I11. Aberrant miR-210-3p expression in body fluids has been shown to assist in the diagnosis of several types of malignancies, including lung cancer.³⁷ miR-31-5p downregulation in breast cancer is related to tumor metastases, whereas elevated miR-31-5p expression in colorectal cancer may be associated with late stage tumors.³⁶ miR-31-5p dysregulation endorses KRAS mutation-driven NSCLC, and hence plays a key role in lung tumorigenesis.³⁸ Our current observations that the elevated expression levels of the three miRNAs in sputum and plasma of NSCLC patients further support their important roles in lung carcinogenesis.

The study may have some limitations. First, clinically useful biomarkers should have very high sensitivity and specificity for diagnosis of lung cancer. Integrated use of

three miRNAs has a sensitivity of 85% and a specificity of 91%, which are still not sufficient to be used in the clinics for the early detection of lung cancer. We are aiming to identify new lung tumor-associated miRNAs of the exfoliated bronchial epitheliums of airway in sputum by using whole genome next-generation sequencing, which may provide additional biomarkers to be added in the integrated approach to more precisely diagnose lung cancer. Second, using LDCT for the early detection of lung cancer can reduce the mortality.³⁹ However, LDCT always produces over-diagnosis or a false positive rate.³⁹ It would be interesting to evaluate whether the integrated panel of miRNA biomarkers could complement LDCT for the early detection of lung cancer by specifically reducing its over-diagnosis. Third, small cell lung cancer (SCLC) accounts for approximately 15% of all lung cancers.¹ However, specimens from SCLC patients were not available in this current study. Diagnostic potential of the biomarkers for the early detection of SCLC remains unknown. In our ongoing study, we are collecting sputum and plasma samples from patients diagnosed with SCLC to evaluate whether these biomarkers could be useful for the diagnosis of SCLC.

In conclusion, integrating sputum and plasma miRNA biomarkers has a synergistic value, and hence presents a potential approach for the early detection of NSCLC. Nevertheless, a large multicenter trial is required to confirm the integrated method before it can be translated in laboratory settings.

Acknowledgments

This work was supported in part by the Geaton and JoAnn DeCesaris Family Foundation and University of Maryland Marlene & Stewart Greenebaum Comprehensive Cancer Center Pilot Grant Program (F.J.).

Disclosure

The authors declare no conflict of interest.

References

- 1 Torre LA, Bray F, Siegel RL *et al.* Global cancer statistics, 2012. *CA Cancer J Clin* 2015; **65**: 87–108.
- 2 Aberle DR, Adams AM, Berg CD *et al.* Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 2011; **365**: 395–409.
- 3 Belinsky SA, Liechty KC, Gentry FD *et al.* Promoter hypermethylation of multiple genes in sputum precedes lung cancer incidence in a high-risk cohort. *Cancer Res* 2006; **66**: 3338–44.
- 4 Croce CM, Calin GA. miRNAs, cancer, and stem cell division. *Cell* 2005; **122**: 6–7.

- 5 Wang J, Tian X, Han R *et al*. Downregulation of miR-486-5p contributes to tumor progression and metastasis by targeting protumorigenic ARHGAP5 in lung cancer. *Oncogene* 2014; **33**: 1181–9.
- 6 Pang W, Tian X, Bai F *et al*. Pim-1 kinase is a target of miR-486-5p and eukaryotic translation initiation factor 4E, and plays a critical role in lung cancer. *Mol Cancer* 2014; **13**: 240.
- 7 Xing L, Todd NW, Yu L, Fang H, Jiang F. Early detection of squamous cell lung cancer in sputum by a panel of microRNA markers. *Mod Pathol* 2010; **23**: 1157–64.
- 8 Belinsky SA, Leng S, Wu G *et al*. Gene methylation biomarkers in sputum and plasma as predictors for lung cancer recurrence. *Cancer Prev Res (Phila)* 2017; **10**: 635–40.
- 9 Xie Y, Todd NW, Liu Z *et al*. Altered miRNA expression in sputum for diagnosis of non-small cell lung cancer. *Lung Cancer* 2010; **67**: 170–6.
- 10 Fusco N, Fumagalli C, Guerini-Rocco E. Looking for sputum biomarkers in lung cancer secondary prevention: Where are we now? *J Thorac Dis* 2017; **9**: 4277–9.
- 11 Bagheri A, Khorshid HRK, Tavallaie M *et al*. A panel of noncoding RNAs in non-small-cell lung cancer. *J Cell Biochem* 2018; **23**: 35–8.
- 12 He RQ, Cen WL, Cen JM *et al*. Clinical significance of miR-210 and its prospective Signaling pathways in non-small cell lung cancer: Evidence from gene expression omnibus and the cancer genome atlas data mining with 2763 samples and validation via real-time quantitative PCR. *Cell Physiol Biochem* 2018; **46**: 925–52.
- 13 Bagheri A, Khorram Khorshid HR, Mowla SJ *et al*. Altered miR-223 expression in sputum for diagnosis of non-small cell lung cancer. *Avicenna J Med Biotechnol* 2017; **9**: 189–95.
- 14 Razzak R, Bedard EL, Kim JO *et al*. MicroRNA expression profiling of sputum for the detection of early and locally advanced non-small-cell lung cancer: A prospective case-control study. *Curr Oncol* 2016; **23**: e86–94.
- 15 Kim JO, Gazala S, Razzak R *et al*. Non-small cell lung cancer detection using microRNA expression profiling of bronchoalveolar lavage fluid and sputum. *Anticancer Res* 2015; **35**: 1873–80.
- 16 Shen J, Liao J, Guarnera MA *et al*. Analysis of MicroRNAs in sputum to improve computed tomography for lung cancer diagnosis. *J Thorac Oncol* 2014; **9**: 33–40.
- 17 Shen J, Liu Z, Todd NW *et al*. Diagnosis of lung cancer in individuals with solitary pulmonary nodules by plasma microRNA biomarkers. *BMC Cancer* 2011; **11**: 374.
- 18 Cohen JD, Li L, Wang Y *et al*. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* 2018; **359**: 926–30.
- 19 Yu L, Shen J, Mannoor K, Guarnera M, Jiang F. Identification of ENO1 as a potential sputum biomarker for early-stage lung cancer by shotgun proteomics. *Clin Lung Cancer* 2014; **15**: 372–378.e1.
- 20 Li N, Ma J, Guarnera MA, Fang H, Cai L, Jiang F. Digital PCR quantification of miRNAs in sputum for diagnosis of lung cancer. *J Cancer Res Clin Oncol* 2014; **140**: 145–50.
- 21 Anjuman N, Li N, Guarnera M, Stass SA, Jiang F. Evaluation of lung flut in sputum samples for molecular analysis of lung cancer. *Clin Transl Med* 2013; **2**: 15.
- 22 Yu L, Todd NW, Xing L *et al*. Early detection of lung adenocarcinoma in sputum by a panel of microRNA markers. *Int J Cancer* 2010; **127**: 2870–8.
- 23 Jiang F, Todd NW, Li R, Zhang H, Fang H, Stass SA. A panel of sputum-based genomic marker for early detection of lung cancer. *Cancer Prev Res (Phila)* 2010; **3**: 1571–8.
- 24 Jiang F, Todd NW, Qiu Q, Liu Z, Katz RL, Stass SA. Combined genetic analysis of sputum and computed tomography for noninvasive diagnosis of non-small-cell lung cancer. *Lung Cancer* 2009; **66**: 58–63.
- 25 Qiu Q, Todd NW, Li R *et al*. Magnetic enrichment of bronchial epithelial cells from sputum for lung cancer diagnosis. *Cancer* 2008; **114**: 275–83.
- 26 Li R, Todd NW, Qiu Q *et al*. Genetic deletions in sputum as diagnostic markers for early detection of stage I non-small cell lung cancer. *Clin Cancer Res* 2007; **13**: 482–7.
- 27 Tuck MK, Chan DW, Chia D *et al*. Standard operating procedures for serum and plasma collection: Early detection research network consensus statement standard operating procedure integration working group. *J Proteome Res* 2009; **8**: 113–7.
- 28 Ma J, Lin Y, Zhan M, Mann DL, Stass SA, Jiang F. Differential miRNA expressions in peripheral blood mononuclear cells for diagnosis of lung cancer. *Lab Invest* 2015; **95**: 1197–206.
- 29 Xing L, Su J, Guarnera MA *et al*. Sputum microRNA biomarkers for identifying lung cancer in indeterminate solitary pulmonary nodules. *Clin Cancer Res* 2015; **21**: 484–9.
- 30 Su J, Liao J, Gao L *et al*. Analysis of small nucleolar RNAs in sputum for lung cancer diagnosis. *Oncotarget* 2015; **7**: 5131–42.
- 31 Shen J, Todd NW, Zhang H *et al*. Plasma microRNAs as potential biomarkers for non-small-cell lung cancer. *Lab Invest* 2011; **91**: 579–87.
- 32 Su Y, Guarnera MA, Fang H, Jiang F. Small non-coding RNA biomarkers in sputum for lung cancer diagnosis. *Mol Cancer* 2016; **36**: 15.
- 33 Su J, Leng Q, Lin Y *et al*. Integrating circulating immunological and sputum biomarkers for the early detection of lung cancer. *Biomark Cancer* 2018; **10**: 1179.
- 34 Zhu S, Wu H, Wu F, Nie D, Sheng S, Mo YY. MicroRNA-21 targets tumor suppressor genes in invasion and metastasis. *Cell Res* 2008; **18**: 350–9.
- 35 Grosso S, Doyen J, Parks SK *et al*. MiR-210 promotes a hypoxic phenotype and increases radioresistance in human lung cancer cell lines. *Cell Death Dis* 2013; **4**: e544.
- 36 Wang S, Hu J, Zhang D, Li J, Fei Q, Sun Y. Prognostic role of microRNA-31 in various cancers: A meta-analysis. *Tumour Biol* 2014; **35**: 11639–45.
- 37 Chang W, Lee CY, Park JH *et al*. Survival of hypoxic human mesenchymal stem cells is enhanced by a positive feedback loop involving miR-210 and hypoxia-inducible factor 1. *J Vet Sci* 2013; **14**: 69–76.

- 38 Edmonds MD, Boyd KL, Moyo T *et al.* MicroRNA-31 initiates lung tumorigenesis and promotes mutant KRAS-driven lung cancer. *J Clin Invest* 2016; **126**: 349–64.
- 39 Patz EF Jr, Pinsky P, Gatsonis C *et al.* Overdiagnosis in low-dose computed tomography screening for lung cancer. *JAMA Intern Med* 2014; **174**: 269–74.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Table S1 Expression levels of the sputum miRNAs and plasma miRNAs in NSCLC patients versus cancer-free controls of the training cohort. **Figure S1** The receiver-operator characteristic (ROC) curve of sputum and plasma miRNA biomarkers for lung cancer diagnosis. Two sputum miRNAs (miRs-21-5p and 210-3p) and one plasma miRNA (miRNA-31-5p) were selected as the best ones and incorporated into an algorithm: $P = \frac{e^x}{1 + e^x}$, where $x = 6.62 + 1.84 \times \log(\text{miR-21-5p}) - 2.6 \times \log(\text{miR-21-3p}) - 2.36 \times \log(\text{miR-31-5p})$. The algorithm produced the area under the ROC (AUC) of 0.91 for distinguishing lung cancer cases from controls.

Sputum long non-coding RNA biomarkers for diagnosis of lung cancer

Chhavi Gupta^{a,1}, Jian Su^{a,1}, Min Zhan^b, Sanford A. Stass^a and Feng Jiang^{a,*}

^aDepartment of Pathology, University of Maryland School of Medicine, Baltimore, MD, USA

^bDepartments of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, MD, USA

Abstract.

BACKGROUND: Analysis of molecular changes in sputum may help diagnose lung cancer. Long non-coding RNAs (lncRNAs) play vital roles in various biological processes, and their dysregulations contribute to the development and progression of lung tumorigenesis. Herein, we determine whether aberrant lncRNAs could be used as potential sputum biomarkers for lung cancer.

METHODS: Using reverse transcription PCR, we measure expressions of lung cancer-associated lncRNAs in sputum of a discovery cohort of 67 lung cancer patients and 65 *cancer-free smokers with benign diseases* and a validation cohort of 59 lung cancer patients and 60 *cancer-free smokers with benign diseases*.

RESULTS: In the discovery cohort, four of the lncRNAs displayed a significantly different level in sputum of lung cancer patients *vs. cancer-free smokers with benign diseases* (all $P < 0.001$). From the four lncRNAs, three lncRNAs (SNHG1, H19, and HOTAIR) are identified as a biomarker panel, producing 82.09% sensitivity and 89.23% specificity for diagnosis of lung cancer. Furthermore, the biomarker panel has a higher sensitivity (82.09% *vs.* 52.24%, $P = 0.02$) and a similar specificity compared with sputum cytology (89.23% *vs.* 90.77%, $P = 0.45$). In addition, the lncRNA biomarker panel had a higher sensitivity (87.50% *vs.* 70.07%, $p = 0.03$) for diagnosis of squamous cell carcinoma compared with adenocarcinoma of the lung, while maintaining the same specificity (89.23%). The potential of the sputum lncRNA biomarkers for lung cancer detection is confirmed in the validation cohort.

CONCLUSION: We have for the first time shown that the analysis of lncRNAs in sputum might be a noninvasive approach for diagnosis of lung cancer.

Keywords: Diagnosis, lung cancer, sputum, lncRNA, biomarkers

1. Introduction

Approximately 155,870 Americans will die from lung cancer each year, more than the other three leading cancers combined (breast, prostate, and colorectal cancers). Over 85% lung cancers are non-small cell lung cancers (NSCLC). NSCLC mainly consists of adenocarcinoma (AC) and squamous cell carcinoma (SCC). Tobacco smoking is the major cause of NSCLC. The early detection of lung cancer in a large

randomized trial using low-dose CT (LDCT) has revealed a 20% reduction in mortality as compared to chest X-rays [1]. Therefore, lung cancer early detection can increase curability and save lives. LDCT is recommended to be used for lung cancer early detection among smokers in the USA [2,3]. However, LDCT is associated with over-diagnosis, excessive cost, and radiation exposure [2,3]. The development of non-invasive biomarkers that can accurately and cost-effectively diagnose early stage lung cancer remains to be clinically imperative [2,3].

It has been well accepted that lung tumors develop from a field defect characterized by molecular abnormalities resulted from repeated exposure of the entire airway to the tobacco carcinogens [4,5]. Kadara et al. [6,7] found that the molecular alterations in the

¹These authors contributed equally to this work.

*Corresponding author: Feng Jiang, Department of Pathology, University of Maryland School of Medicine, 10 South Pine Street, MSTF 7th floor, Baltimore, MD 21201, USA. E-mail: fjiang@som.umaryland.edu.

large bronchial airways reflected those in primary lung tumors in the distal lung, regardless of the anatomic location relative to the tumors. Spira et al. [8,9] demonstrated that molecular changes in epithelial cells collected from the normal-appearing mainstem bronchus of smokers could be developed as biomarkers for NSCLC. Sputum contains bronchial epithelial cells from the lungs and lower respiratory tract. Based on the field cancerizations, examination of the exfoliated bronchial epitheliums of airway in sputum might detect the lung tumor-related molecular alterations, and hence provide a non-invasive and specific means for diagnosis of NSCLC [6–8,10–12].

Long non-coding RNAs (lncRNAs) are transcripts of longer than 200 nucleotides in length, and have important functions in various biological processes [13, 14]. Dysfunctions of lncRNAs play vital roles in the development of lung cancer [15]. Specifically, lncRNAs can regulate different molecular signaling pathways via changing gene expression, and therefore, are implicated in numerous mechanisms of lung carcinogenesis [16–20]. Since aberrant lncRNAs detected in clinical specimens are associated with lung tumorigenesis, exploration of the molecules for clinical diagnosis is a growing area of research [15,21,22]. For examples, Liu et al. found that an lncRNA, HOTAIR, was significantly up-regulated in NSCLC tissues and might be a prognostic biomarker for the disease [23]. Furthermore, we have found that expression levels of SNHG1 and RMRP are reliably measured in plasma, and the lncRNAs may provide cell-free circulating biomarkers for lung cancer [24]. In addition, we recently showed that integrated analysis of lncRNAs, miRNAs, and mRNAs in plasma could have a synergistic impact on lung cancer early detection [25]. However, examination of aberrant lncRNA levels in sputum, which is surrogate material for noninvasive and specific diagnosis of lung cancer has not been investigated.

We recently identified six lncRNAs whose aberrant expression level in surgically resected tumor tissues was associated with lung cancer [26,27]. The objective of this study was to investigate whether the lung cancer-associated lncRNAs could be used as sputum biomarkers for distinguishing lung cancer patients from smokers with benign diseases.

2. Materials and methods

2.1. Study population

The study protocol was approved by the Institutional Review Board of the University of Maryland Medi-

cal Center. Written informed consent forms were obtained from all participants. The participants were recruited at the point of their referral for suspected lung cancer between the ages of 55–80. Exclusion criteria included pregnancy, current pulmonary infection, surgery within 6 months, radiotherapy within 1 year, and life expectancy of < 1 year. Clinical diagnosis of lung cancer was made using histopathologic examinations of specimens obtained by CT-guided transthoracic needle biopsy, transbronchial biopsy, videotape-assisted thoracoscopic surgery, or surgical resection. The surgical pathologic staging was determined according to the TNM classification of the International Union Against Cancer with the 8th American Joint Committee on Cancer and the International Staging System for Lung Cancer. Histopathological classification was determined according to the World Health Organization classification. A total of 251 subjects including 126 lung cancer patients and 125 *cancer-free smokers with benign diseases* were recruited. The 126 lung cancer patients were diagnosed with NSCLC consisting of 71 stage I cases, 33 stage II cases, and 22 stage III–IV cases. Seventy-six cases were AC and 50 were SCC of the lungs. *All the 125 cancer-free individuals were smokers who served as control group participants. These control participants had* granulomatous inflammation ($n = 89$), nonspecific inflammatory changes ($n = 23$) or lung infections ($n = 13$). The cancer-free individuals had been followed for at least 2 years, and none had any evidence of cancer. No difference of age, gender, and smoking status was observed in the lung cancer cases vs. *cancer-free smokers* (All $p > 0.05$). The lung cases and *cancer-free smokers* were randomly split into a training cohort and a testing cohort by using a random number generator. The training cohort consisted of 67 lung cancer patients and 65 *cancer-free smokers*. The testing cohort comprised 59 lung cancer patients and 60 *cancer-free smokers*. The demographic and clinical characteristics of the two cohorts are presented in Tables 1 and 2.

2.2. Sputum collection, preparation, and sputum cytological study

Sputum was collected and processed from the subjects as previously described [10,28–43], before they received any treatment (e.g., surgery, preoperative adjuvant chemotherapy, and radiotherapy). The sputum samples were centrifuged at 800 g for 10 min. The cell pellets were mixed with phosphate-buffered saline solution (PBS) (Sigma-Aldrich, St. Louis, MO. Cytospin

Table 1
Characteristics of NSCLC patients and cancer-free smokers in a training cohort

	NSCLC cases (n = 67)	Cancer-free individuals (n = 65)	P-value
Age	68.39 (SD 11.36)	66.75 (SD 11.32)	0.38
Sex			0.46
Female	26	23	
Male	41	42	
Race			0.45
White	44	43	
African American	23	22	
Pack-years (median)	35.89	35.47	0.39
Stage			
Stage I	40		
Stage II	17		
Stage III–IV	10		
Histological type			
Adenocarcinoma	43		
Squamous cell carcinoma	24		

Abbreviations: NSCLC, non-small cell lung cancer.

Table 2
Characteristics of NSCLC patients and cancer-free smokers in a training cohort

	NSCLC cases (n = 59)	Cancer-free individuals (n = 60)	P-value
Age	68.57 (SD 12.03)	66.28 (SD 11.60)	0.39
Sex			0.43
Female	23	22	
Male	36	38	
Race			0.47
White	39	40	
African American	20	20	
Pack-years (median)	36.02	35.59	0.40
Stage			
Stage I	31		
Stage II	16		
Stage III–IV	12		
Histological type			
Adenocarcinoma	33		
Squamous cell carcinoma	26		

Abbreviations: NSCLC, non-small cell lung cancer.

slides were prepared from the cell pellets and underwent Papanicolaou staining for evaluating whether the specimens were representative of deep bronchial cells. All sputum samples were of lower respiratory origin as indicated by the presence of macrophages and bronchial epithelial cells. Cytologic diagnosis was made using the classification of Saccomanno [44]. Positive cytology included both carcinoma *in situ* and invasive carcinoma [28,29]. The remaining sputum cells were stored at -80°C until being tested.

2.3. The analysis of lncRNAs in sputum by reverse transcription-PCR (RT-PCR)

RNA was extracted from cell pellets of sputum as previously described [31–34,36,37]. The purity and concentration of RNA were determined by OD260/280 readings using a dual beam UV spectrophotometer (Eppendorf AG, Hamburg, Germany). RNA integrity was determined by capillary electrophoresis using the RNA 6000 Nano Lab-on-a-Chip kit and the Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA, USA). The expression levels of the six lncRNAs (SNHG1, MALAT1, HOTAIR, H19, MEG3, and RMRP) were determined in sputum by using RT-PCR with Taqman assays (Applied Biosystems, Foster City, CA, USA) as previously described [24,25]. The cycle threshold (Ct) was defined as the number of cycles required for the fluorescent signal to cross the threshold. An internal control gene, U6, was analyzed in parallel in the specimens. Relative expression of a targeted lncRNA in a given sample was computed using the equation $2^{-\Delta\text{Ct}}$, where $\Delta\text{Ct} = \text{Ct (targeted lncRNA)} - \text{Ct (U6)}$. All assays were performed in triplicates. Furthermore, two interplate controls and one no-template control were carried along in each experiment. The no template control for RT was RNease free water instead of RNA sample input, and no template control for PCR was RNease free water instead of RT products input.

2.4. Statistical analysis

We used univariate analysis to identify the lncRNAs whose expression levels were associated with NSCLC. We analyzed the significantly associated lncRNAs by using logistic regression models with constrained parameters as least absolute shrinkage and selection operator (LASSO) to eliminate the less relevant variables [45–47,47]. We estimated functions of the combined lncRNA biomarkers by logistic regression with or without adjustment for known risk factors for NSCLC. The performances of the lncRNAs were subjected to receiver-operator characteristic (ROC) curve. We also generated a 95% confidence interval for the difference in the area under the ROCs (AUCs) by using the bootstrap [48]. We established the optimal cut-off value by using the Youden index [49,50]. The results of the training cohort were blindly validated in the testing cohort by using leave-one-out cross validation [51]. To compare different approaches for lung cancer diagnosis, we computed their AUCs to determine the sensitivity and specificity as previously described [52].

Table 3
Expression levels of sputum lncRNAs in the training cohort of lung cancer patients and cancer-free individuals

lncRNAs	Mean (SEM) of level in NSCLC patients	Mean (SEM) of level in cancer-free individuals	P-value	AUC (95% CI)
SNHG1	6.65 (1.08)	5.29 (1.13)	< 0.001	0.80 (0.73 to 0.87)
H19	3.39 (2.63)	2.25 (0.65)	< 0.001	0.77 (0.68 to 0.83)
HOTAIR	1.08 (0.56)	0.29 (0.09)	< 0.001	0.73 (0.66 to 0.83)
MEG3	1.53 (0.92)	3.02 (1.56)	< 0.001	0.72 (0.63 to 0.78)

Abbreviations: SEM, the standard error of the mean; AUC, the area under receiver operating characteristic curve; CI, confidence interval.

Table 4
Diagnostic values of the sputum lncRNA biomarkers in training cohort and testing cohort

	The training cohort		The testing cohort	
	Sensitivity (CI)	Specificity (CI)	Sensitivity (CI)	Specificity (CI)
NSCLC	82.09% (70.80% to 90.39%)	89.23% (79.06% to 95.56%)	81.36% (69.09% to 90.31%)	88.33% (77.43% to 95.18%)
AC	79.07% (63.96% to 89.96%)	89.23% (79.06% to 95.56%)	75.76% (57.74% to 88.91%)	88.33% (77.43% to 95.18%)
SCC	87.50% (67.64% to 97.34%)	89.23% (79.06% to 95.56%)	88.46% (69.85% to 97.55%)	88.33% (77.43% to 95.18%)

Abbreviations: NSCLC, non-small cell lung cancers; AC, adenocarcinoma; SCC, squamous cell carcinoma; CI, confidence interval.

3. Results

3.1. Developing a panel of sputum lncRNA biomarkers for the early detection of lung cancer

All targeted six lncRNAs had ≤ 32 Ct values in each sputum sample of the training cohort, and therefore were reliably detectable in the specimens. No product was synthesized in the negative control samples. As shown in Table 3, four (SNHG1, H19, and HOTAIR, and MEG3) of the six lncRNAs displayed a significantly different level in sputum of patients with lung cancer vs. *cancer-free individuals*, producing 0.72–0.80 AUCs (all $P < 0.01$). Of the four lncRNAs, three (SNHG1, H19, and HOTAIR) had a higher, whereas one (MEG3) had a lower expression level in sputum of lung cancer patients vs. *cancer-free individuals* (All $P < 0.01$) (Table 3). We used logistic regression models with constrained parameters as in LASSO to develop a panel of lncRNA biomarkers for lung cancer. SNHG1, H19, and HOTAIR were selected as the best biomarkers and incorporated into an algorithm, which produced 0.90 AUC for distinguishing lung cancer cases from cancer-free subjects (Supplementary Fig. 1). Furthermore, the estimated correlations among expression levels of the three lncRNAs were low (All $P > 0.05$), implying that the diagnostic values of the lncRNAs were complementary to each other. Using Youden's index, we set up optimal cut-off at 0.36. Based on the cut-off, combined analysis of the three lncRNAs generated 82.03% sensitivity and 89.23% specificity for diagnosis of lung cancer (Table 4). Sputum cytology has 52.24% sensitivity and 90.77% specificity. Therefore, the sputum biomarkers had a higher sensitivity (82.03%) compared with spu-

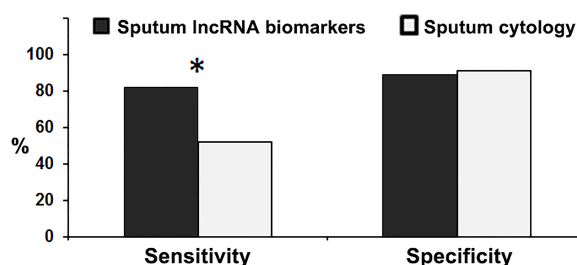


Fig. 1. Sensitivity and specificity of sputum lncRNA biomarkers and cytology for diagnosis of lung cancer. The sputum lncRNA biomarker panel has a higher sensitivity and a similar specificity compared with sputum cytology ($P = 0.02$). *, $P < 0.05$.

tum cytology (52.24%) ($P = 0.02$), while maintaining a similar specificity (89.23% vs. 90.77%, $p = 0.45$) (Fig. 1). Interestingly, the panel of the lncRNA biomarkers had a higher sensitivity (87.50%) for SCC compared with AC of the lungs (79.07%), while maintaining the same specificity (89.23%) (Table 4). Therefore, the sputum biomarker panel could better differentiate SCC cases from *cancer-free individuals* compared with AC cases. In addition, the three lncRNAs did not display statistical differences of sensitivity and specificity between early and advanced stages (stage I–II vs. stage III–IV) ($P > 0.05$). Moreover, the sputum biomarkers had no special association with the age, gender and ethnicity of the participants (All $p > 0.05$). The expression levels of the sputum lncRNAs were associated with smoking history of lung cancer patients at the edge of significance ($P = 0.05$).

3.2. Validating the sputum lncRNA markers in an independent cohort of lung cancer patients and cancer-free individuals

To evaluate the diagnostic performance of the spu-

tum biomarkers, the three lncRNAs (SNHG1, H19, and HOTAIR) were assessed in sputum of additional 59 NSCLC patients and 60 *cancer-free individuals*. The three genes used in combination could differentiate the NSCLC patients from *cancer-free individuals* with 81.36% sensitivity and 88.33% specificity (Table 4). Furthermore, the panel of the lncRNA biomarkers had a higher sensitivity for SCC (88.46%) compared with AC (75.76%) of the lungs ($P = 0.03$) with the same specificity (88.33%). In addition, the three sputum lncRNA biomarkers didn't show association with the age, gender and ethnicity of the participants (All $p > 0.05$). Taken together, the results generated from the validation cohort confirmed the potential of combined use of the three lncRNAs as a sputum biomarker panel for the early detection of lung cancer.

4. Discussion

As a mirror to lung diseases, sputum is noninvasively obtained and contains bronchial epithelial cells from the lungs and lower respiratory tract, and thus has the advantages as surrogate material for specifically diagnosing lung cancer. Herein, we for the first time report the feasibility of detection of lncRNA in sputum and develop a panel of sputum lncRNA biomarkers for lung cancer. Furthermore, the biomarker panel has a higher sensitivity and a similar specificity compared with sputum cytology, a standard clinical model. Like other sputum molecular biomarkers, the sputum lncRNA biomarkers display a higher diagnostic performance for SCC than AC of the lungs. Interestingly, the sensitivity and specificity of the biomarkers are independent of stage of lung cancer. Therefore, the discovery might be an important characteristic if they are employed for the diagnosis of NSCLC, particularly SCC, at the early stage. *Nevertheless, the potential of the sputum lncRNA biomarkers for the early detection of lung cancer need to be prospectively validated in a large cohort study.*

The three lncRNAs (SNHG1, H19, and HOTAIR) have diverse and important functions in lung tumorigenesis through regulating different molecular pathways. For instance, elevated expression of SNHG1 was frequently observed in lung cancer tissues [53]. Furthermore, SNHG1 could promote NSCLC progression of lung cancer via miR-101-3p/SOX9/Wnt/ β -catenin regulatory network and miR-145-5p/MTDH axis [53,54]. In addition, SNHG1 plays an oncogenic role in lung cancer through ZEB1 signaling pathway by inhibiting

TA63 [55]. H19 is an imprinted oncofetal lncRNA, which is expressed in the embryo, down-regulated at birth and then reappears in tumors [56]. H19 plays important function in cell proliferation and differentiation, epithelial to mesenchymal transition, and mesenchymal to epithelial transition [57]. H19 has also the versatile and functions depending on its context within the process of tumor progression [58]. Interestingly, lung tumor-released H19 could promote gefitinib resistance via packaging into exosomes in lung cancer cells [59]. HOTAIR represses gene expression through recruitment of chromatin modifiers. The increased HOTAIR expression has been founded in lung cancer, and the expression level correlates with metastasis and poor prognosis. Moreover, HOTAIR promotes proliferation, survival, invasion, metastasis, and drug resistance in lung cancer cells through miR-217/DACH1 signaling pathway [60,61]. Our current study extends the previous findings by developing the lung cancer-associated lncRNA as sputum biomarkers that might be useful in the early detection of lung cancer.

There are some limitations in this present study. 1), the size of the cohorts is small. We will perform a large and prospective study to validate the diagnostic value of the sputum lncRNA biomarkers for the early detection of lung cancer. 2), the panel of the three biomarkers is developed from six lung cancer-associated lncRNAs. Although showing promise, the three biomarkers' diagnostic performance is not high enough to be used in clinical settings. By searching published data, we recently found 21 lncRNAs whose malfunction was characterized in lung tumorigenesis [19,60,62–86]. Our ongoing study is evaluating if the additional lncRNAs might be added into the biomarker panel to improve the diagnosis of lung cancer. 3), the objective of this present study is to develop sputum biomarkers for differentiating lung cancer from cancer-free smokers who have benign diseases. Toward that end, we collect sputum from lung cancer patients and cancer-free smokers with benign diseases who serve as control individuals. Sputum of healthy controls is not available in this study. Therefore, we are not able to detect differences of lncRNAs in sputum between smokers with benign diseases and healthy controls. In future, it might be important to collect sputum from nonsmokers who don't have any benign disease to determine if there is different lncRNA changes in sputum between the individuals with the smokes with benign diseases and the healthy controls.

In sum, sputum lncRNA biomarkers are developed that could potentially be used as a noninvasive ap-

proach for differentiating early stage NSCLC patients from cancer-free smokers. Nevertheless, undertaking a prospective study to further validate the sputum biomarkers in a large cohort is required.

Acknowledgments

This work was supported in part by VA Merit Award I01 CX000512, the Geaton and JoAnn DeCesaris Family Foundation, NIH/NCI-1R21CA240556, University of Maryland Marlene & Stewart Greenebaum Comprehensive Cancer Center Pilot Grant Program, and DoD-Idea Development Award (F.J.), and NIH/NCI-Early Detection Research Network-5U24CA11509 (S.S.).

Conflict of interest

The authors declare no conflict of interest.

References

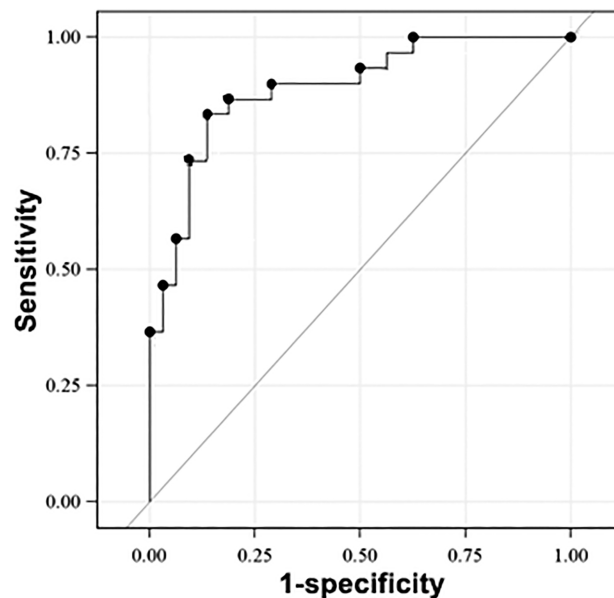
- [1] D.R. Aberle, A.M. Adams, C.D. Berg, W.C. Black, J.D. Clapp, R.M. Fagerstrom, I.F. Gareen, C. Gatsonis, P.M. Marcus and J.D. Sicks, Reduced lung-cancer mortality with low-dose computed tomographic screening, *N Engl J Med* **365** (2011), 395–409.
- [2] E.F. Patz, Jr., P. Pinsky, C. Gatsonis, J.D. Sicks, B.S. Kramer, M.C. Tammemagi, C. Chiles, W.C. Black and D.R. Aberle, Overdiagnosis in low-dose computed tomography screening for lung cancer, *JAMA Intern Med* **174** (2014), 269–274.
- [3] D.R. Aberle, C.D. Berg, W.C. Black, T.R. Church, R.M. Fagerstrom, B. Galen, I.F. Gareen, C. Gatsonis, J. Goldin, J.K. Gohagan, B. Hillman, C. Jaffe, B.S. Kramer, D. Lynch, P.M. Marcus, M. Schnall, D.C. Sullivan, D. Sullivan and C.J. Zylak, The national lung screening trial: overview and study design, *Radiology* **258** (2011), 243–253.
- [4] X. Tang, H. Shigematsu, B.N. Bekele, J.A. Roth, J.D. Minna, W.K. Hong, A.F. Gazdar and Wistuba, II, EGFR tyrosine kinase domain mutations are detected in histologically normal respiratory epithelium in lung cancer patients, *Cancer Res* **65** (2005), 7568–7572.
- [5] L.M. Solis, C. Behrens, M.G. Raso, H.Y. Lin, H. Kadara, P. Yuan, H. Galindo, X. Tang, J.J. Lee, N. Kalhor, Wistuba, II and C.A. Moran, Histologic patterns and molecular characteristics of lung adenocarcinoma associated with clinical outcome, *Cancer* **118** (2012), 2889–2899.
- [6] H. Kadara, L. Shen, J. Fujimoto, P. Saintigny, C.W. Chow, W. Lang, Z. Chu, M. Garcia, M. Kabbout, Y.H. Fan, C. Behrens, D.A. Liu, L. Mao, J.J. Lee, K.A. Gold, J. Wang, K.R. Coombes, E.S. Kim, W.K. Hong and Wistuba, II, Characterizing the molecular spatial and temporal field of injury in early-stage smoker non-small cell lung cancer patients after definitive surgery by expression profiling, *Cancer Prev Res (Phila)* **6** (2013), 8–17.
- [7] H.K. Wistuba II, Field cancerization in non-small cell lung cancer: implications in disease pathogenesis, *Proc Am Thorac Soc* **9** (2012), 38–42.
- [8] A. Spira, J.E. Beane, V. Shah, K. Steiling, G. Liu, F. Schembri, S. Gilman, Y.M. Dumas, P. Calner, P. Sebastiani, S. Sridhar, J. Beamis, C. Lamb, T. Anderson, N. Gerry, J. Keane, M.E. Lenburg and J.S. Brody, Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer, *Nat Med* **13** (2007), 361–366.
- [9] G.A. Silvestri, A. Vachani, D. Whitney, M. Elashoff, K. Porta Smith, J.S. Ferguson, E. Parsons, N. Mitra, J. Brody, M.E. Lenburg and A. Spira, A bronchial genomic classifier for the diagnostic evaluation of lung cancer, *N Engl J Med* **373** (2015), 243–251.
- [10] S.A. Belinsky, K.C. Liechty, F.D. Gentry, H.J. Wolf, J. Rogers, K. Vu, J. Haney, T.C. Kennedy, F.R. Hirsch, Y. Miller, W.A. Franklin, J.G. Herman, S.B. Baylin, P.A. Bunn and T. Byers, Promoter hypermethylation of multiple genes in sputum precedes lung cancer incidence in a high-risk cohort, *Cancer Res* **66** (2006), 3338–3344.
- [11] N. Fusco, C. Fumagalli and E. Guerini-Rocco, Looking for sputum biomarkers in lung cancer secondary prevention: where are we now? *J Thorac Dis* **9** (2017), 4277–4279.
- [12] A.J. Hubers, C.F. Prinsen, G. Sozzi, B.I. Witte and E. Thunnissen, Molecular sputum analysis for the diagnosis of lung cancer, *Br J Cancer* **109** (2013), 530–537.
- [13] L. Ma, V.B. Bajic and Z. Zhang, On the classification of long non-coding RNAs, *RNA Biol* **10** (2013), 925–933.
- [14] R. Spizzo, M.I. Almeida, A. Colombatti and G.A. Calin, Long non-coding RNAs and cancer: a new frontier of translational research? *Oncogene* **31** (2012), 4577–4587.
- [15] J. Yang, J. Lin, T. Liu, T. Chen, S. Pan, W. Huang and S. Li, Analysis of lncRNA expression profiles in non-small cell lung cancers (NSCLC) and their clinical subtypes, *Lung Cancer* **85** (2014), 110–115.
- [16] D. Meseure, K. Drak Alsibai, A. Nicolas, I. Bieche and A. Morillon, Long noncoding RNAs as new architects in cancer epigenetics, prognostic biomarkers, and potential therapeutic targets, *Biomed Res Int* **2015** (2015), 320214.
- [17] J.R. Prensner and A.M. Chinnaiyan, The emergence of lncRNAs in cancer biology, *Cancer Discov* **1** (2011), 391–407.
- [18] M. Zhou, M. Guo, D. He, X. Wang, Y. Cui, H. Yang, D. Hao and J. Sun, A potential signature of eight long non-coding RNAs predicts survival in patients with non-small cell lung cancer, *J Transl Med* **13** (2015), 231.
- [19] M. Li, M. Qiu, Y. Xu, Q. Mao, J. Wang, G. Dong, W. Xia, R. Yin and L. Xu, Differentially expressed protein-coding genes and long noncoding RNA in early-stage lung cancer, *Tumour Biol* **36** (2015), 9969–9978.
- [20] L.H. Schmidt, T. Spieker, S. Koschmieder, S. Schaffers, J. Humberg, D. Jungen, E. Bulk, A. Hascher, D. Wittmer, A. Marra, L. Hillejan, K. Wiebe, W.E. Berdel, R. Wiewrodt and C. Muller-Tidow, 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* **6** (2011), 1984–1992.
- [21] J. Chen, R. Wang, K. Zhang and L.B. Chen, Long non-coding RNAs in non-small cell lung cancer as biomarkers and therapeutic targets, *J Cell Mol Med* **18** (2014), 2425–2436.
- [22] H. Tao, J.J. Yang, X. Zhou, Z.Y. Deng, K.H. Shi and J. Li, Emerging role of long noncoding RNAs in lung cancer: Current status and future prospects, *Respir Med* **110** (2016), 12–19.

- [23] X.H. Liu, Z.L. Liu, M. Sun, J. Liu, Z.X. Wang and W. De, The long non-coding RNA HOTAIR indicates a poor prognosis and promotes metastasis in non-small cell lung cancer, *BMC Cancer* **13** (2013), 464.
- [24] Y. Lin, Q. Leng, M. Zhan and F. Jiang, A plasma long noncoding RNA signature for early detection of lung cancer, *Transl Oncol* **11** (2018), 1225–1231.
- [25] Q. Leng, Y. Lin, M. Zhan and F. Jiang, An integromic signature for lung cancer early detection, *Oncotarget* **9** (2018), 24684–24692.
- [26] J. Ma, K. Mannoor, L. Gao, A. Tan, M.A. Guarnera, M. Zhan, A. Shetty, S.A. Stass, L. Xing and F. Jiang, Characterization of microRNA transcriptome in lung cancer by next-generation deep sequencing, *Mol Oncol* **8** (2014), 1208–1219.
- [27] L. Gao, J. Ma, K. Mannoor, M.A. Guarnera, A. Shetty, M. Zhan, L. Xing, S.A. Stass and F. Jiang, Genome-wide small nucleolar RNA expression analysis of lung cancer by next-generation deep sequencing, *Int J Cancer* **136** (2015), E623–629.
- [28] M.S. Romeo, I.A. Sokolova, L.E. Morrison, C. Zeng, A.E. Baron, F.R. Hirsch, Y.E. Miller, W.A. Franklin and M. Varella-Garcia, Chromosomal abnormalities in non-small cell lung carcinomas and in bronchial epithelia of high-risk smokers detected by multi-target interphase fluorescence in situ hybridization, *J Mol Diagn* **5** (2003), 103–112.
- [29] M. Varella-Garcia, J. Kittelson, A.P. Schulte, K.O. Vu, H.J. Wolf, C. Zeng, F.R. Hirsch, T. Byers, T. Kennedy, Y.E. Miller, R.L. Keith and W.A. Franklin, Multi-target interphase fluorescence in situ hybridization assay increases sensitivity of sputum cytology as a predictor of lung cancer, *Cancer Detect Prev* **28** (2004), 244–251.
- [30] L. Yu, J. Shen, K. Mannoor, M. Guarnera and F. Jiang, Identification of ENO1 as a potential sputum biomarker for early-stage lung cancer by shotgun proteomics, *Clin Lung Cancer* (2014).
- [31] J. Shen, J. Liao, M.A. Guarnera, H. Fang, L. Cai, S.A. Stass and F. Jiang, Analysis of MicroRNAs in sputum to improve computed tomography for lung cancer diagnosis, *J Thorac Oncol* **9** (2014), 33–40.
- [32] N. Li, J. Ma, M.A. Guarnera, H. Fang, L. Cai and F. Jiang, Digital PCR quantification of miRNAs in sputum for diagnosis of lung cancer, *J Cancer Res Clin Oncol* **140** (2014), 145–150.
- [33] N. Anjuman, N. Li, M. Guarnera, S.A. Stass and F. Jiang, Evaluation of lung flute in sputum samples for molecular analysis of lung cancer, *Clin Transl Med* **2** (2013), 15.
- [34] L. Yu, N.W. Todd, L. Xing, Y. Xie, H. Zhang, Z. Liu, H. Fang, J. Zhang, R.L. Katz and F. Jiang, Early detection of lung adenocarcinoma in sputum by a panel of microRNA markers, *Int J Cancer* **127** (2010), 2870–2878.
- [35] F. Jiang, N.W. Todd, R. Li, H. Zhang, H. Fang and S.A. Stass, A panel of sputum-based genomic marker for early detection of lung cancer, *Cancer Prev Res (Phila)* **3** (2010), 1571–1578.
- [36] L. Xing, N.W. Todd, L. Yu, H. Fang and F. Jiang, Early detection of squamous cell lung cancer in sputum by a panel of microRNA markers, *Mod Pathol* **23** (2010), 1157–1164.
- [37] Y. Xie, N.W. Todd, Z. Liu, M. Zhan, H. Fang, H. Peng, M. Alattar, J. Deepak, S.A. Stass and F. Jiang, Altered miRNA expression in sputum for diagnosis of non-small cell lung cancer, *Lung Cancer* **67** (2010), 170–176.
- [38] F. Jiang, N.W. Todd, Q. Qiu, Z. Liu, R.L. Katz and S.A. Stass, Combined genetic analysis of sputum and computed tomography for noninvasive diagnosis of non-small-cell lung cancer, *Lung Cancer* **66** (2009), 58–63.
- [39] R.L. Katz, T.M. Zaidi, R.L. Fernandez, J. Zhang, W. He, C. Acosta, M. Daniely, L. Madi, M.A. Vargas, Q. Dong, X. Gao, F. Jiang, N.P. Caraway, A.A. Vaporciyan, J.A. Roth and M.R. Spitz, Automated detection of genetic abnormalities combined with cytology in sputum is a sensitive predictor of lung cancer, *Mod Pathol* **21** (2008), 950–960.
- [40] Q. Qiu, N.W. Todd, R. Li, H. Peng, Z. Liu, H.G. Yfantis, R.L. Katz, S.A. Stass and F. Jiang, Magnetic enrichment of bronchial epithelial cells from sputum for lung cancer diagnosis, *Cancer* **114** (2008), 275–283.
- [41] R. Li, N.W. Todd, Q. Qiu, T. Fan, R.Y. Zhao, W.H. Rodgers, H.B. Fang, R.L. Katz, S.A. Stass and F. Jiang, Genetic deletions in sputum as diagnostic markers for early detection of stage I non-small cell lung cancer, *Clin Cancer Res* **13** (2007), 482–487.
- [42] W.H. Roa, J.O. Kim, R. Razzak, H. Du, L. Guo, R. Singh, S. Gazala, S. Ghosh, E. Wong, A.A. Joy, J.Z. Xing and E.L. Bedard, Sputum microRNA profiling: a novel approach for the early detection of non-small cell lung cancer, *Clin Invest Med* **35** (2012), E271.
- [43] E.O. Machida, M.V. Brock, C.M. Hooker, J. Nakayama, A. Ishida, J. Amano, M.A. Picchi, S.A. Belinsky, J.G. Herman, S. Taniguchi and S.B. Baylin, Hypermethylation of ASC/TMS1 is a sputum marker for late-stage lung cancer, *Cancer Res* **66** (2006), 6210–6218.
- [44] G. Saccomanno, R.P. Saunders, V.E. Archer, O. Auerbach, M. Kuschner and P.A. Beckler, Cancer of the lung: the cytology of sputum prior to the development of carcinoma, *Acta Cytol* **9** (1965), 413–423.
- [45] Y. Su, M.A. Guarnera, H. Fang and F. Jiang, Small non-coding RNA biomarkers in sputum for lung cancer diagnosis, *Mol Cancer* **15** (2016), 36.
- [46] S.J. Swensen, M.D. Silverstein, D.M. Ilstrup, C.D. Schleck and E.S. Edell, The probability of malignancy in solitary pulmonary nodules. Application to small radiologically indeterminate nodules, *Arch Intern Med* **157** (1997), 849–855.
- [47] A. McWilliams, M.C. Tammemagi, J.R. Mayo, H. Roberts, G. Liu, K. Soghrati, K. Yasufuku, S. Martel, F. Laberge, M. Gingras, S. Atkar-Khattra, C.D. Berg, K. Evans, R. Finley, J. Yee, J. English, P. Nasute, J. Goffin, S. Puksa, L. Stewart, S. Tsai, M.R. Johnston, D. Manos, G. Nicholas, G.D. Goss, J.M. Seely, K. Amjadi, A. Tremblay, P. Burrowes, P. MacEachern, R. Bhatia, M.S. Tsao and S. Lam, Probability of cancer in pulmonary nodules detected on first screening CT, *N Engl J Med* **369** (2013), 910–919.
- [48] J.A. Hanley and B.J. McNeil, A method of comparing the areas under receiver operating characteristic curves derived from the same cases, *Radiology* **148** (1983), 839–843.
- [49] J. Luo and C. Xiong, Youden index and associated cut-points for three ordinal diagnostic groups, *Commun Stat Simul Comput* **42** (2013), 1213–1234.
- [50] E.F. Schisterman, D. Faraggi, B. Reiser and J. Hu, Youden index and the optimal threshold for markers with mass at zero, *Stat Med* **27** (2008), 297–315.
- [51] S. Baek, C.A. Tsai and J.J. Chen, Development of biomarker classifiers from high-dimensional data, *Brief Bioinform* **10** (2009), 537–546.
- [52] J. Shen, Z. Liu, N.W. Todd, H. Zhang, J. Liao, L. Yu, M.A. Guarnera, R. Li, L. Cai, M. Zhan and F. Jiang, Diagnosis of lung cancer in individuals with solitary pulmonary nodules by plasma microRNA biomarkers, *BMC Cancer* **11** (2011), 374.
- [53] Y. Cui, F. Zhang, C. Zhu, L. Geng, T. Tian and H. Liu, Up-regulated lncRNA SNHG1 contributes to progression of non-small cell lung cancer through inhibition of miR-101-3p and

- activation of Wnt/beta-catenin signaling pathway, *Oncotarget* **8** (2017), 17785–17794.
- [54] Q. Lu, S. Shan, Y. Li, D. Zhu, W. Jin and T. Ren, Long non-coding RNA SNHG1 promotes non-small cell lung cancer progression by up-regulating MTDH via sponging miR-145-5p, *FASEB J* (2018), fj201701237RR.
- [55] H.Y. Zhang, W. Yang, F.S. Zheng, Y.B. Wang and J.B. Lu, Long non-coding RNA SNHG1 regulates zinc finger E-box binding homeobox 1 expression by interacting with TAp63 and promotes cell metastasis and invasion in Lung squamous cell carcinoma, *Biomed Pharmacother* **90** (2017), 650–658.
- [56] I. Matouk, E. Raveh, P. Ohana, R.A. Lail, E. Gershtain, M. Gilon, N. De Groot, A. Czerniak and A. Hochberg, The increasing complexity of the oncofetal h19 gene locus: functional dissection and therapeutic intervention, *Int J Mol Sci* **14** (2013), 4298–4316.
- [57] L. Liu and S. Lu, lncRNA H19 promotes viability and epithelial-mesenchymal transition of lung adenocarcinoma cells by targeting miR-29b-3p and modifying STAT3, *Int J Oncol* **54** (2019), 929–941.
- [58] E. Raveh, I.J. Matouk, M. Gilon and A. Hochberg, The H19 Long non-coding RNA in cancer initiation, progression and metastasis – a proposed unifying theory, *Mol Cancer* **14** (2015), 184.
- [59] Y. Lei, W. Guo, B. Chen, L. Chen, J. Gong and W. Li, Tumor-released lncRNA H19 promotes gefitinib resistance via packaging into exosomes in non-small cell lung cancer, *Oncol Rep* **40** (2018), 3438–3446.
- [60] G. Loewen, J. Jayawickramarajah, Y. Zhuo and B. Shan, Functions of lncRNA HOTAIR in lung cancer, *J Hematol Oncol* **7** (2014), 90.
- [61] S.S. Chen, M. Peng, G.Z. Zhou, Y.C. Pu, M.C. Yi, Y. Zhu and B. Jiang, Long non-coding RNA HOTAIR regulates the development of non-small cell lung cancer through miR-217/DACH1 signaling pathway, *Eur Rev Med Pharmacol Sci* **23** (2019), 670–678.
- [62] M.M. Wei, Y.C. Zhou, Z.S. Wen, B. Zhou, Y.C. Huang, G.Z. Wang, X.C. Zhao, H.L. Pan, L.W. Qu, J. Zhang, C. Zhang, X. Cheng and G.B. Zhou, Long non-coding RNA stabilizes the Y-box-binding protein 1 and regulates the epidermal growth factor receptor to promote lung carcinogenesis, *Oncotarget* **7** (2016), 59556–59571.
- [63] L. Shen, L. Chen, Y. Wang, X. Jiang, H. Xia and Z. Zhuang, Long noncoding RNA MALAT1 promotes brain metastasis by inducing epithelial-mesenchymal transition in lung cancer, *J Neurooncol* **121** (2015), 101–108.
- [64] P. Li, J. Li, R. Yang, F. Zhang, H. Wang, H. Chu, Y. Lu, S. Dun, Y. Wang, W. Zang, Y. Du, X. Chen, G. Zhao and G. Zhang, Study on expression of lncRNA RGMB-AS1 and repulsive guidance molecule b in non-small cell lung cancer, *Diagn Pathol* **10** (2015), 63.
- [65] Y.R. Yang, S.Z. Zang, C.L. Zhong, Y.X. Li, S.S. Zhao and X.J. Feng, Increased expression of the lncRNA PVT1 promotes tumorigenesis in non-small cell lung cancer, *Int J Clin Exp Pathol* **7** (2014), 6929–6935.
- [66] E.J. Whiteside, I. Seim, J.P. Pauli, A.J. O’Keeffe, P.B. Thomas, S.L. Carter, C.M. Walpole, J.N. Fung, P. Josh, A.C. Herington and L.K. Chopin, Identification of a long non-coding RNA gene, growth hormone secretagogue receptor opposite strand, which stimulates cell migration in non-small cell lung cancer cell lines, *Int J Oncol* **43** (2013), 566–574.
- [67] T. Hu and Y.R. Lu, BCYRN1, a c-MYC-activated long non-coding RNA, regulates cell metastasis of non-small-cell lung cancer, *Cancer Cell Int* **15** (2015), 36.
- [68] J. Li, P. Li, W. Zhao, R. Yang, S. Chen, Y. Bai, S. Dun, X. Chen, Y. Du, Y. Wang, W. Zang, G. Zhao and G. Zhang, Expression of long non-coding RNA DLX6-AS1 in lung adenocarcinoma, *Cancer Cell Int* **15** (2015), 48.
- [69] Z. Zeng, H. Bo, Z. Gong, Y. Lian, X. Li, W. Zhang, H. Deng, M. Zhou, S. Peng, G. Li and W. Xiong, AFAP1-AS1, a long noncoding RNA upregulated in lung cancer and promotes invasion and metastasis, *Tumour Biol* **37** (2016), 729–737.
- [70] Z. Hou, W. Zhao, J. Zhou, L. Shen, P. Zhan, C. Xu, C. Chang, H. Bi, J. Zou, X. Yao, R. Huang, L. Yu and J. Yan, A long non-coding RNA Sox2ot regulates lung cancer cell proliferation and is a prognostic indicator of poor survival, *Int J Biochem Cell Biol* **53** (2014), 380–388.
- [71] J. Luo, L. Tang, J. Zhang, J. Ni, H.P. Zhang, L. Zhang, J.F. Xu and D. Zheng, Long non-coding RNA CARLo-5 is a negative prognostic factor and exhibits tumor pro-oncogenic activity in non-small cell lung cancer, *Tumour Biol* **35** (2014), 11541–11549.
- [72] H. Luo, Y. Sun, G. Wei, J. Luo, X. Yang, W. Liu, M. Guo and R. Chen, Functional characterization of long noncoding RNA lnc_bc060912 in human lung carcinoma cells, *Biochemistry* **54** (2015), 2895–2902.
- [73] Y. Wu, H. Liu, X. Shi, Y. Yao, W. Yang and Y. Song, The long non-coding RNA HNF1A-AS1 regulates proliferation and metastasis in lung adenocarcinoma, *Oncotarget* **6** (2015), 9160–9172.
- [74] M. Qiu, Y. Xu, X. Yang, J. Wang, J. Hu, L. Xu and R. Yin, CCAT2 is a lung adenocarcinoma-specific long non-coding RNA and promotes invasion of non-small cell lung cancer, *Tumour Biol* **35** (2014), 5375–5380.
- [75] M. Qiu, Y. Xu, J. Wang, E. Zhang, M. Sun, Y. Zheng, M. Li, W. Xia, D. Feng, R. Yin and L. Xu, A novel lncRNA, LUADT1, promotes lung adenocarcinoma proliferation via the epigenetic suppression of p27, *Cell Death Dis* **6** (2015), e1858.
- [76] L. Zhang, X.F. Zhou, G.F. Pan and J.P. Zhao, Enhanced expression of long non-coding RNA ZXF1 promoted the invasion and metastasis in lung adenocarcinoma, *Biomed Pharmacother* **68** (2014), 401–407.
- [77] F.Q. Nie, M. Sun, J.S. Yang, M. Xie, T.P. Xu, R. Xia, Y.W. Liu, X.H. Liu, E.B. Zhang, K.H. Lu and Y.Q. Shu, Long non-coding RNA ANRIL promotes non-small cell lung cancer cell proliferation and inhibits apoptosis by silencing KLF2 and P21 expression, *Mol Cancer Ther* **14** (2015), 268–277.
- [78] R. Yang, P. Li, G. Zhang, C. Lu, H. Wang and G. Zhao, Long non-coding RNA XLOC_008466 functions as an oncogene in human non-small cell lung cancer by targeting miR-874, *Cell Physiol Biochem* **42** (2017), 126–136.
- [79] H. Sang, H. Liu, P. Xiong and M. Zhu, Long non-coding RNA functions in lung cancer, *Tumour Biol* **36** (2015), 4027–4037.
- [80] X. Shi, M. Sun, H. Liu, Y. Yao, R. Kong, F. Chen and Y. Song, A critical role for the long non-coding RNA GAS5 in proliferation and apoptosis in non-small-cell lung cancer, *Mol Carcinog* **54**(Suppl 1) (2015), E1–E12.
- [81] L. Han, R. Kong, D.D. Yin, E.B. Zhang, T.P. Xu, W. De and Y.Q. Shu, Low expression of long noncoding RNA GAS6-AS1 predicts a poor prognosis in patients with NSCLC, *Med Oncol* **30** (2013), 694.
- [82] L. Han, E.B. Zhang, D.D. Yin, R. Kong, T.P. Xu, W.M. Chen, R. Xia, Y.Q. Shu and W. De, Low expression of long noncoding RNA PANDAR predicts a poor prognosis of non-small cell lung cancer and affects cell apoptosis by regulating Bcl-2, *Cell Death Dis* **6** (2015), e1665.

- [83] X. Xie, H.T. Liu, J. Mei, F.B. Ding, H.B. Xiao, F.Q. Hu, R. Hu and M.S. Wang, LncRNA HMLncRNA717 is down-regulated in non-small cell lung cancer and associated with poor prognosis, *Int J Clin Exp Pathol* **7** (2014), 8881–8886.
- [84] J. Liu, L. Wan, K. Lu, M. Sun, X. Pan, P. Zhang, B. Lu, G. Liu and Z. Wang, The long noncoding RNA MEG3 contributes to cisplatin resistance of human lung adenocarcinoma, *PLoS One* **10** (2015), e0114586.
- [85] M. Sun, X.H. Liu, K.M. Wang, F.Q. Nie, R. Kong, J.S. Yang, R. Xia, T.P. Xu, F.Y. Jin, Z.J. Liu, J.F. Chen, E.B. Zhang, W. De and Z.X. Wang, Downregulation of BRAF activated non-coding RNA is associated with poor prognosis for non-small cell lung cancer and promotes metastasis by affecting epithelial-mesenchymal transition, *Mol Cancer* **13** (2014), 68.
- [86] Y. Yang, H. Li, S. Hou, B. Hu, J. Liu and J. Wang, The noncoding RNA expression profile and the effect of lncRNA AK126698 on cisplatin resistance in non-small-cell lung cancer cell, *PLoS One* **8** (2013), e65309.

Supplementary



Supplementary Fig. 1. The receiver-operator characteristic (ROC) curve of sputum lncRNA biomarkers for lung cancer diagnosis. Three lncRNAs (SNHG1, H19, and HOTAIR) were selected as the best ones and incorporated into an algorithm: Probability $P = \exp(U) / [1 + \exp(U)]$, where $U = 2.93 + 2.12 * \log(\text{SNHG1}) - 1.26 * \log(\text{H19}) - 1.1795 * \log(\text{HOTAIR})$. The three-lncRNA-algorithm produced the area under the ROC (AUC) of 0.90 for distinguishing lung cancer cases from cancer-free individuals.

Copyright of Cancer Biomarkers is the property of IOS Press and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.

Fucosylation genes as circulating biomarkers for lung cancer

Qixin Leng¹, Jen-Hui Tsou¹, and Min Zhan², Feng Jiang^{1*}

1. Department of Pathology, University of Maryland School of Medicine, 10 S. Pine St.
Baltimore, MD 21201, USA

2. Departments of Epidemiology & Public Health, University of Maryland School of Medicine,
660 W. Redwood St. Baltimore, MD 21201, USA

*Correspondence to Feng Jiang, Department of Pathology, The University of Maryland
School of Medicine, 10 South Pine Street, MSTF 7th floor, Baltimore, MD 21201-1192, USA.

E-mail: fjiang@som.umaryland.edu.

Short running title: Fucosylation genes as biomarkers for lung cancer

Abstract

Purpose

Fucosyltransferases (FUTs) catalyze fucosylation, which plays a central role in biological processes. Aberrant fucosylation is associated with malignant transformation. Here we investigated whether transcriptional levels of genes coding the FUTs in plasma could provide cell-free circulating biomarkers for lung cancer.

Methods

mRNA expression of all 13 *Futs* (*Fut1-11*, *Pofut1*, and *Pofut2*) was evaluated by PCR assay in 48 lung tumor tissues and the 48 matched noncancerous lung tissues, and plasma of 64 lung cancer patients and 32 cancer-free individuals to develop plasma *Fut* biomarkers. The developed plasma *Fut* biomarkers were validated in an independent cohort of 40 lung cancer patients and 20 controls for their diagnostic performance.

Results

Four of the 13 *Futs* showed a different transcriptional level in 48 lung tumor tissues compared with the 48 matched nonconscious tissues (All <0.05). Two (*Fut8*, and *Pofut1*) of the four *Futs* had a higher plasma level in 64 lung cancer patients compared with 32 control subjects, and consistent with that in lung tissue specimens. Combined analysis of the two *Futs* produced 81% sensitivity and 86% specificity for diagnosis of lung cancer, and was independent of stage and histology of lung tumors. The diagnostic performance of the two plasma biomarkers was successfully validated in the different cohort of 40 lung cancer patients and 20 control individuals.

Conclusion

The fucosylation genes may provide new circulating biomarkers for the early detection of lung cancer.

Keywords: diagnosis; lung cancer; plasma; fucosyltransferases; biomarkers.

1. Introduction

Lung cancer is the leading cause of cancer death among both men and women. More people die of lung cancer than of colon, breast, and prostate cancers combined [1; 2]. Non-small cell lung cancer (NSCLC) accounts for more than 85% of all lung cancers. NSCLC consists of adenocarcinoma (AC), squamous cell carcinoma (SCC), and large cell carcinoma (LC). Tobacco smoking is the major cause of NSCLC. Results of the National Lung Screening Trial (NLST) shows that the early detection of lung cancer using low-dose CT (LDCT) can reduce the mortality[3] [4]. However, LDCT has a specificity of only 61%, producing a high false positive rate or over-diagnosis [5]. As a result, many smokers with benign diseases will have harmful diagnostic and therapeutic procedures. Therefore, the lack of noninvasive approaches for accurately diagnosing early stage lung cancer remains a top clinical challenge.

During tumor development, cancer cells undergo apoptosis and necrosis, and release tumor-associated molecules that can circulate in bloodstream. The tumors-derived molecules in plasma may provide cell-free circulating cancer biomarkers. Although numerous plasma biomarkers have been developed by detecting the circulating cell-free DNA, gene methylated products, proteins, and metabolites for lung cancer early detection, none of them has been well accepted in the clinics [6]. For instance, a blood test (Cancer-SEEK) was recently developed that could detect eight common cancer types by determining circulating proteins and mutations of cell-free DNA [7]. However, the test had about 60% sensitivity for diagnosis of all stages of lung cancer and only 40% sensitivity for the stage I disease[7].

Glycosylation is one of the most abundant protein modifications, and involved in major physiological events, including cell differentiation, proliferation, trafficking, migration and intercellular signaling. Gathering evidences have demonstrated that aberrant glycosylation is the result of alterations in glycosyltransferases that play crucial in the development and

progression of carcinogenesis [8]. Fucosylation is the major type of glycosylation and regulated by fucosyltransferases (FUTs), which catalyze the transfer of the fucose residue from GDP-fucose donor substrate to acceptor substrates present on oligosaccharides, glycoproteins and glycolipids [9]. There are 13 different FUTs, including FUT1 to 11, protein O-fucosyltransferase 1 (POFUT1), and POFUT2. [10; 11]. Abnormal protein expression of FUTs has been proven to associate the development and progression of malignancies, including lung cancer [10]. Furthermore, studying glycans and glycan-binding proteins has shown that changes in fucosylation of glycoproteins were potential cancer biomarkers. For example, increased fucosylated alpha-fetoprotein (AFP) level is observed in sera of patients with hepatocellular carcinoma. AFP is one of the most representative types of glycan-related cancer biomarkers [12; 13]. However, AFP is also elevated in sera of patients with various benign diseases such as accurate and hepatitis [14]. Therefore, the fucosylated glycoproteins-based circulating biomarkers exhibit an insufficient value in the diagnosis of malignancies. In addition, the current proteomic techniques for analysis of the plasma glycobiological biomarkers are labor intensive, expensive, and time consuming.

Since the elevated activity of the 13 FUTs is regulated by their own increased gene expression, we hypothesize that mRNA expression levels of the *Fut* genes in plasma may provide circulating biomarkers for lung cancer. In this study, we first evaluated transcriptional levels of the 13 *Futs* in 48 lung tumor and 48 matched nonconscious lung tissues. We determined the mRNA expression levels in plasma of 64 lung cancer and 32 control subjects and developed a panel of plasma biomarkers, which were validated in additional 40 lung cancer and 20 control subjects. A panel of plasma biomarkers for lung cancer early detection was successfully developed.

2. Materials and Methods

2.1. Patients and clinical specimens. This study was approved by the Institutional Review Boards of University of Maryland Baltimore and Veterans Affairs Maryland Health Care System. Surgically resected tissue specimens were obtained from 48 lung cancer patients who had either a lobectomy or a pneumonectomy. Tumor tissues were intraoperatively dissected from the surrounding lung parenchyma. The paired 48 normal lung tissues were also obtained from the same patients at an area distant from their tumors. Serial cryostat sections from the specimens were prepared and used to confirm the diagnosis based on the WHO classification of tumors of the lung ¹⁰. All 48 cases were diagnosed with histologically confirmed stage I NSCLC, including 25 AC and 23 SCC.

To collect blood samples, we enrolled lung cancer patients and cancer-free smokers by using the inclusion and/or exclusion criteria recommended by U.S. Preventive Services Task Force for lung cancer screening in heavy smokers[15]. We collected blood and prepared plasma using the standard operating protocols (SOPs) developed by The NCI-Early Detection Research Network (EDRN)[16; 17]. Briefly, 10 ml whole blood was collected by venous puncture with butterfly needles into BD Vacutainer spray-coated K2EDTA Tubes (BD, Franklin Lakes, NJ). All the blood samples were processed within 2 hours of collection. To obtain plasma, whole blood samples were centrifuged at 3,000 rpm at room temperature for 15 minutes. The supernatant was immediately removed from the buffy coat, and frozen in 1.8 ml RNase-free cryo-tubes (Thermo Scientific, Rochester, NY), and stored at -80°C until further processing. A total of 104 NSCLC patients and 52 cancer-free smokers were recruited. Among the cancer patients, 26 patients were female and 78 were male. Twenty-five had stage I NSCLC, 17 with stage II, 28 with stage III, 29 with stage IV, and 5 with unknown stage. Of the cancer-free smokers, 13 patients were female and 39 were male. There were no significant differences of age, gender and smoking status between the NSCLC patients and cancer-free

smokers. In this study, the cases and controls were randomly grouped into two cohorts: a development cohort and a validation cohort. The development cohort consisted of 64 lung cancer patients and 32 cancer-free smokers, while the validation cohort comprised 40 lung cancer patients and 20 cancer-free smokers. The demographic and clinical variables of the two sets are shown in Tables 1-2.

2.2. RNA isolation and quantitative reverse transcriptase PCR (qRT-PCR).

RNA was extracted from the specimens by using Trizol LS reagent (Invitrogen Carlsbad, CA) and RNeasy Mini Kit (Qiagen, Hilden, Germany) [18; 19]. RT was carried out to generate cDNA by using a RT Kit (Applied Biosystems, Foster City, CA) as described in our published works [18]. PCR was performed to measure expressions of target genes by using a PCR kit (Applied Biosystems) on a Bio-Red IQ5 Multicolor RT-PCR Detection System (Bio-Red, Hercules, CA). Primers and probes of the targeted 13 *FUTs* genes are shown in the supplementary table 1. Expression levels of the genes were determined using comparative cycle threshold (Ct) method with the equation $2^{-\Delta\Delta Ct}$ by using miR-1228 as an internal control [20]. The targeted genes with Ct values > 35 were considered to be below the detection level of qRT-PCR.

2.3. Droplet Digital PCR (ddPCR).

ddPCR for analysis of expression level of the genes was performed as described in our previous work [21; 22]. Briefly, TaqMan™ reaction mix (Applied Biosystems) containing sample cDNA was partitioned into aqueous droplets in oil via the QX100 Droplet Generator (Bio-Rad), and then transferred to a 96-well PCR plate. A two-step thermocycling protocol (95°C × 10 min; 40 cycles of [94°C × 30 s, 60°C × 60 s], 98°C × 10 min) was undertaken in a Bio-Rad C1000 (Bio-Rad). The PCR plate was loaded on Droplet Reader (Bio-Rad), by which copy number of each gene per µl PCR reaction was directly determined. We used QuantaSoft 1.7.4 analysis software (Bio-Rad) and Poisson

statistics to compute droplet concentrations (copies/ μ L). All assays were done in triplicates, and one no-template control and two interplate controls were carried along in each experiment.

2.4. Statistical analysis. Pearson's correlation analysis was applied to assess relationship between gene expressions and demographic and clinical characteristics of the lung cancer patients and control individuals. The area under receiver operating characteristic (ROC) curve (AUC) analyses were used to determine sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and corresponding cut-off value of each gene [23]. All P values shown were two sided, and a P value of <0.05 was considered statistically significant.

3. Results

3.1. Identifying *Futs* whose abnormal transcriptional levels were associated with lung cancer. The transcription levels of all 13 *Futs* were examined by using qRT-PCR in 48 stage I NSCLC tissues and the matched noncancerous lung tissues. Three (*Futs-7 and 8*, and *pofut1*) displayed a higher, whereas one gen (*Fut-4*) exhibited a lower mRNA expression levels in lung cancer tissues compared with the matched noncancerous lung tissues (All $p<0.05$) (Fig.1). There was no significant difference of the expression of the *Futs* between histological types of lung cancer (All $p>0.05$).

3.2. Developing a panel of plasma *Fut* biomarkers for lung cancer early detection. Transcriptional levels of the 4 *Futs* were first determined in plasma by qRT-PCR in a training set of 64 cases and 32 controls. The amplification curves of the RT-PCR analysis for the genes were not reliably generated by qRT-PCR. In contrast, when ddPCR was used to determine the level of the 4 *Futs* (*Futs-7 and 8*, *pofut1*, and *Fut-4*) in the plasma samples,

each well of the samples contained at least 10,000 droplets. Thus, ddPCR rather than qRT-PCR could quantify the *Fut* genes in plasma.

Of the 4 genes, *Fut8* and *Pofut1* had a high expression level in plasma of lung cancer patients vs. cancer-free controls (All $P < 0.05$) (Fig. 2A). There were significant correlations between the expression levels of the 2 genes in plasma and those in the surgical tissue specimens (All $r \geq 0.92$, all $P \leq 0.05$). Other 2 genes (*Fut4* and *Fut7*) did not show changes in plasma of lung cancer patients vs. controls. Furthermore, *Fut8* and *Pofut1* exhibited AUC values of 0.86 and 0.81, respectively, in distinguishing NSCLC patients from the healthy individuals (Fig. 2B). When optimal cutoffs for the two *FUT*s were set at 1.56 and 1.64, respectively, the individual genes produced 62.50-73.44% sensitivities, 90.63-90.88% specificities, 93.02-94.00% PPVs, and 63.04-71.88% NPVs (Table 3). Combined use of the two genes based on at least one positive result in either *Fut8* or *Pofut1* produced the highest classification accuracy (0.833) compared to any one used alone (all $p < 0.05$) (Table 3). The 2 genes used in combination created 81.76% sensitivity, 86.26% specificity, 92.86% PPV, and 70% NPV for diagnosis of lung cancer, thus significantly improving the cancer detection rate by a single gene with only a 4% decrease in specificity (Table 3). Furthermore, the estimated correlation among the 2 genes was very low ($r=0.187$, $p=0.06$), further supporting that the combined analysis of the 2 *FUT* genes outperformed a single one for diagnosis of NSCLC. In addition, combined analysis of the 2 plasma biomarkers did not show special association with stage and histological type of lung cancer, and patients' age, gender, and smoking status (All $P > 0.05$).

3.3. Validating the sputum markers in an independent set of lung cancer patients and controls. To evaluate the diagnostic performance of the plasma biomarker panel, the 2 genes (*Fut8* and *Pofut1*) were assessed in plasma samples of 40 NSCLC patients

and 20 healthy controls. The 2 genes used in combination could differentiate the NSCLC patients from healthy controls with 82.50% sensitivity and 85.00% specificity. Furthermore, no statistically significant difference was found in the sensitivity and specificity of the markers for stages and histological types of NSCLC (All $p > 0.05$). Moreover, there was no association of expressions of the 2 genes with the age, gender, or smoking status of the lung cancer patients and normal individuals (All $p > 0.05$). Taken together, the results confirm that the panel of 2 plasma biomarkers could be used for the early detection of lung cancer.

4. Discussion

In this present study, we have for the first time demonstrated that mRNA expression of *Futs* could provide cell-free circulating biomarkers for lung cancer. Furthermore, using ddPCR assay, a sensitive and robust technique, we successfully developed 2 *Futs* as small panel biomarkers for effective diagnosis of lung cancer. In addition, the diagnostic performance is independent of stage and histological type of the NSCLC, and age and gender of subjects. Therefore, the plasma biomarkers have an important characteristic if it is employed for more precisely and easily identifying early stage lung cancer.

Chen et al showed that *Fut8* was up-regulated in lung cancer and tissues [24]. A high protein expression of FUT8 was correlated with tumor metastasis, disease recurrence, and poor survival in patients with NSCLC [24; 25]. Downregulation of *Fut8* significantly inhibited the malignant behaviors of lung cancer cells. *Fut8* could globally modify surface antigens, receptors, and adhesion molecules [24]. Dysregulation of *Fut8* was involved in the regulation of dozens of genes associated with the malignancy through multiple mechanisms [24]. The observations in the present study provide further evidence that dysregulation of FUT8 has important role in lung tumorigenesis. Importantly, we demonstrate that a high mRNA

expression level of *Fut8* in plasma may provide useful biomarker for lung cancer early detection. POFUT1 is an important O-glycosyltransferase and has an essential role for extracellular Fringe to function [26; 27]. POFUT1 can active Notch1 in breast cancer [28]. Analysis of POFUT1 may have diagnostic or prognostic value in the patients with cancers [28; 29]. Ma et al. found that POFUT1 overexpression could prompt the binding of Notch ligand Dll1 to Notch1 receptor, and thus activated Notch1 signaling pathway in hepatocellular carcinoma cells [30]. Here we report an elevated mRNA level of *Pofut1* in both lung tumor tissues and plasma specimens of lung cancer patients, suggesting that the gene play an important biological function in lung carcinogenesis. Nevertheless, the possible biological role of aberrant expression of *Fut8* and *Pofut1* in lung cancer development and progression is warranted to be investigated.

There are some weaknesses in this study. i), we evaluate and develop the *Futs* as plasma biomarkers using a retrospective cases and controls, which may produce selection bias and overfitting. Our ongoing project is to prospectively validate the biomarkers in a large cohort of cases and controls. ii), based on the FDA criteria [31], a test directed at a disease with a prevalence of $\leq 5\%$ must have a sensitivity exceeding 95% when the specificity is less than or equal to 95%, and vice versa [31]. The overall sensitivity and specificity of the biomarkers are 81% and 86%, and thus not high enough for clinical diagnosis of NSCLC. The integration of the biomarkers with other classes of biomarkers, such as microRNAs [32; 33; 34; 35; 36; 37; 38; 39; 40; 41] or DNA mutations [42] or methylation [43], is one path to improve the early detection of lung cancer.

In summary, we developed a panel of plasma *Fut* biomarkers for lung cancer early detection. Future use of the biomarkers may provide a potential diagnostic test for the early

detection of lung cancer. Nevertheless, further exploring its value for routine use for the early detection of lung cancer would be required.

Legends

Figure 1. mRNA expression levels of 13 *Futs* in surgical lung tumor tissues and the paired noncancerous lung tissues. Four genes (*Futs-4*, *7* and *8*, and *pofut1*) shows a different level in lung tumor tissues as compared with the paired noncancerous lung tissues (All $p < 0.05$).

Figure 2. mRNA expression levels of *Fut8* and *Pofut1* in plasma samples of 64 lung cancer patients and 32 cancer-free controls. A, *Fut8* and *Pofut1* displayed a higher plasma level in lung cancer patients vs. cancer-free controls (all $p < 0.0001$). B, the receiver operating characteristic (ROC) curves of *Fut8* and *Pofut1* produced an area under the ROC curve (AUC) of 0.86 and 0.81, respectively, in diagnosis of lung cancer.

Table 1. Characteristics of 64 NSCLC patients and 32 cancer-free smokers in a training set.

Table 2. Characteristics of 40 NSCLC patients and 20 cancer-free smokers in a testing set.

Table 3. The comparison of diagnostic performance of one-gene and the two genes used in combination for the detection of lung cancer in a training set.

Supplementary table 1. Primers and probes of 13 *Futs*.

Compliance with Ethical Standards:

Disclosure of potential conflicts of interest: The authors declare no conflict of interest.

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical

standards. This article does not contain any studies with animals performed by any of the authors.

Funding: This work was supported in part by NCI R21CA205746, VA Merit Award Io1 CX000512, Award from the Geaton and JoAnn DeCesaris Family Foundation, and Maryland Innovation Initiative (MII) Commercialization Program- Phase 1/2 Grant (F.J.)

References:

- [1]American Cancer Society., Cancer facts and figures, in, American Cancer Society., New York,, pp. v.
- [2]X. Yao, Q. Mao, Y. Li, C. Hao, L. Bian, P. Chen, F. Gao, X. Wu, W. Lu, Q. Gao, X. Li, Z. Liang, Poorly neutralizing polyclonal antibody in vitro against coxsackievirus A16 circulating strains can prevent a lethal challenge in vivo. *Hum Vaccin Immunother* (2018) 0.
- [3]Cancer Facts & Figures 2012. American Cancer Society (ACS). *Journal of Consumer Health on the Internet* 16 (2012) 366-367.
- [4]D.R. Aberle, A.M. Adams, C.D. Berg, W.C. Black, J.D. Clapp, R.M. Fagerstrom, I.F. Gareen, C. Gatsonis, P.M. Marcus, J.D. Sicks, Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 365 (2011) 395-409.
- [5]E.F. Patz, Jr., P. Pinsky, C. Gatsonis, J.D. Sicks, B.S. Kramer, M.C. Tammemagi, C. Chiles, W.C. Black, D.R. Aberle, Overdiagnosis in low-dose computed tomography screening for lung cancer. *JAMA Intern Med* 174 (2014) 269-274.
- [6]A.J. Hubers, C.F. Prinsen, G. Sozzi, B.I. Witte, E. Thunnissen, Molecular sputum analysis for the diagnosis of lung cancer. *Br J Cancer* 109 (2013) 530-537.
- [7]J.D. Cohen, L. Li, Y. Wang, C. Thoburn, B. Afsari, L. Danilova, C. Douville, A.A. Javed, F. Wong, A. Mattox, R.H. Hruban, C.L. Wolfgang, M.G. Goggins, M. Dal Molin, T.L. Wang, R. Roden, A.P. Klein, J. Ptak, L. Dobbyn, J. Schaefer, N. Silliman, M. Popoli, J.T. Vogelstein, J.D. Browne, R.E. Schoen, R.E. Brand, J. Tie,

- P. Gibbs, H.L. Wong, A.S. Mansfield, J. Jen, S.M. Hanash, M. Falconi, P.J. Allen, S. Zhou, C. Bettegowda, L. Diaz, C. Tomasetti, K.W. Kinzler, B. Vogelstein, A.M. Lennon, N. Papadopoulos, Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* (2018).
- [8] V.B. Chachadi, G. Bhat, P.W. Cheng, Glycosyltransferases involved in the synthesis of MUC-associated metastasis-promoting selectin ligands. *Glycobiology* 25 (2015) 963-975.
- [9] L.H. Wu, B.Z. Shi, Q.L. Zhao, X.Z. Wu, Fucosylated glycan inhibition of human hepatocellular carcinoma cell migration through binding to chemokine receptors. *Glycobiology* 20 (2010) 215-223.
- [10] W. Zhou, H. Ma, G. Deng, L. Tang, J. Lu, X. Chen, Clinical significance and biological function of fucosyltransferase 2 in lung adenocarcinoma. *Oncotarget* 8 (2017) 97246-97259.
- [11] F.X. Sullivan, R. Kumar, R. Kriz, M. Stahl, G.Y. Xu, J. Rouse, X.J. Chang, A. Boodhoo, B. Potvin, D.A. Cumming, Molecular cloning of human GDP-mannose 4,6-dehydratase and reconstitution of GDP-fucose biosynthesis in vitro. *J Biol Chem* 273 (1998) 8193-8202.
- [12] J. Breborowicz, A. Mackiewicz, D. Breborowicz, Microheterogeneity of alpha-fetoprotein in patient serum as demonstrated by lectin affino-electrophoresis. *Scand J Immunol* 14 (1981) 15-20.
- [13] Y. Aoyagi, Y. Suzuki, K. Igarashi, A. Saitoh, M. Oguro, T. Yokota, S. Mori, M. Nomoto, M. Isemura, H. Asakura, The usefulness of simultaneous determinations of glucosaminylation and fucosylation indices of alpha-fetoprotein in the differential diagnosis of neoplastic diseases of the liver. *Cancer* 67 (1991) 2390-2394.
- [14] N. Taniguchi, Y. Kizuka, Glycans and cancer: role of N-glycans in cancer biomarker, progression and metastasis, and therapeutics. *Adv Cancer Res* 126 (2015) 11-51.
- [15] L.L. Humphrey, M. Deffebach, M. Pappas, C. Baumann, K. Artis, J.P. Mitchell, B. Zakher, R. Fu, C.G. Slatore, Screening for lung cancer with low-dose computed tomography: a systematic review to update the US Preventive services task force recommendation. *Ann Intern Med* 159 (2013) 411-420.
- [16] J.R. Marks, K.S. Anderson, P. Engstrom, A.K. Godwin, L.J. Esserman, G. Longton, E.S. Iversen, A. Mathew, C. Patriotis, M.S. Pepe, Construction and analysis of the NCI-EDRN breast cancer reference set for circulating markers of disease. *Cancer Epidemiol Biomarkers Prev* 24 (2015) 435-441.

- [17] M.K. Tuck, D.W. Chan, D. Chia, A.K. Godwin, W.E. Grizzle, K.E. Krueger, W. Rom, M. Sanda, L. Sorbara, S. Stass, W. Wang, D.E. Brenner, Standard operating procedures for serum and plasma collection: early detection research network consensus statement standard operating procedure integration working group. *J Proteome Res* 8 (2009) 113-117.
- [18] J. Ma, A. Jemal, R. Smith, Reply to lung cancer deaths averted by screening should be considered in the context of tobacco control policies. *Cancer* 119 (2013) 3420-3421.
- [19] J. Shen, Z. Liu, N.W. Todd, H. Zhang, J. Liao, L. Yu, M.A. Guarnera, R. Li, L. Cai, M. Zhan, F. Jiang, Diagnosis of lung cancer in individuals with solitary pulmonary nodules by plasma microRNA biomarkers. *BMC Cancer* 11 (2011) 374.
- [20] J. Ma, K. Mannoer, L. Gao, A. Tan, M.A. Guarnera, M. Zhan, A. Shetty, S.A. Stass, L. Xing, F. Jiang, Characterization of microRNA transcriptome in lung cancer by next-generation deep sequencing. *Mol Oncol* 8 (2014) 1208-1219.
- [21] N. Li, J. Ma, M.A. Guarnera, H. Fang, L. Cai, F. Jiang, Digital PCR quantification of miRNAs in sputum for diagnosis of lung cancer. *J Cancer Res Clin Oncol* 140 (2014) 145-150.
- [22] J. Ma, N. Li, M. Guarnera, F. Jiang, Quantification of Plasma miRNAs by Digital PCR for Cancer Diagnosis. *Biomark Insights* 8 (2013) 127-136.
- [23] L.E. Dodd, M.S. Pepe, Partial AUC estimation and regression. *Biometrics* 59 (2003) 614-623.
- [24] C.Y. Chen, Y.H. Jan, Y.H. Juan, C.J. Yang, M.S. Huang, C.J. Yu, P.C. Yang, M. Hsiao, T.L. Hsu, C.H. Wong, Fucosyltransferase 8 as a functional regulator of nonsmall cell lung cancer. *Proc Natl Acad Sci U S A* 110 (2013) 630-635.
- [25] R. Honma, I. Kinoshita, E. Miyoshi, U. Tomaru, Y. Matsuno, Y. Shimizu, S. Takeuchi, Y. Kobayashi, K. Kaga, N. Taniguchi, H. Dosaka-Akita, Expression of fucosyltransferase 8 is associated with an unfavorable clinical outcome in non-small cell lung cancers. *Oncology* 88 (2015) 298-308.
- [26] M. Stahl, K. Uemura, C. Ge, S. Shi, Y. Tashima, P. Stanley, Roles of Pofut1 and O-fucose in mammalian Notch signaling. *J Biol Chem* 283 (2008) 13638-13651.

- [27]C. Lorient, A. Audfray, F. Dupuy, A. Germot, A. Maftah, The two N-glycans present on bovine Pofut1 are differently involved in its solubility and activity. *FEBS J* 274 (2007) 1202-1211.
- [28]G. Wan, L. Tian, Y. Yu, F. Li, X. Wang, C. Li, S. Deng, X. Yu, X. Cai, Z. Zuo, F. Cao, Overexpression of Pofut1 and activated Notch1 may be associated with poor prognosis in breast cancer. *Biochem Biophys Res Commun* 491 (2017) 104-111.
- [29]S. Dong, Z. Wang, B. Huang, J. Zhang, Y. Ge, Q. Fan, Bioinformatics insight into glycosyltransferase gene expression in gastric cancer: POFUT1 is a potential biomarker. *Biochem Biophys Res Commun* 483 (2017) 171-177.
- [30]L. Ma, P. Dong, L. Liu, Q. Gao, M. Duan, S. Zhang, S. Chen, R. Xue, X. Wang, Overexpression of protein O-fucosyltransferase 1 accelerates hepatocellular carcinoma progression via the Notch signaling pathway. *Biochem Biophys Res Commun* 473 (2016) 503-510.
- [31]B.M. Nolen, A. Lomakin, A. Marrangoni, L. Velikokhatnaya, D. Prosser, A.E. Lokshin, Urinary protein biomarkers in the early detection of lung cancer. *Cancer Prev Res (Phila)* 8 (2015) 111-119.
- [32]J. Shen, J. Liao, M.A. Guarnera, H. Fang, L. Cai, S.A. Stass, F. Jiang, Analysis of MicroRNAs in sputum to improve computed tomography for lung cancer diagnosis. *J Thorac Oncol* 9 (2014) 33-40.
- [33]J. Su, J. Liao, L. Gao, J. Shen, M.A. Guarnera, M. Zhan, H. Fang, S.A. Stass, F. Jiang, Analysis of small nucleolar RNAs in sputum for lung cancer diagnosis. *Oncotarget* 7 (2016) 5131-5142.
- [34]L. Xing, J. Su, M.A. Guarnera, H. Zhang, L. Cai, R. Zhou, S.A. Stass, F. Jiang, Sputum microRNA biomarkers for identifying lung cancer in indeterminate solitary pulmonary nodules. *Clin Cancer Res* 21 (2015) 484-489.
- [35]N. Li, J. Ma, M.A. Guarnera, H. Fang, L. Cai, F. Jiang, Digital PCR quantification of miRNAs in sputum for diagnosis of lung cancer. *J Cancer Res Clin Oncol* 140 (2014) 145-150.
- [36]N. Anjuman, N. Li, M. Guarnera, S.A. Stass, F. Jiang, Evaluation of lung flut in sputum samples for molecular analysis of lung cancer. *Clin Transl Med* 2 (2013) 15.
- [37]L. Yu, N.W. Todd, L. Xing, Y. Xie, H. Zhang, Z. Liu, H. Fang, J. Zhang, R.L. Katz, F. Jiang, Early detection of lung adenocarcinoma in sputum by a panel of microRNA markers. *Int J Cancer* 127 (2010) 2870-2878.

- [38]F. Jiang, N.W. Todd, R. Li, H. Zhang, H. Fang, S.A. Stass, A panel of sputum-based genomic marker for early detection of lung cancer. *Cancer Prev Res (Phila)* 3 (2010) 1571-1578.
- [39]L. Xing, N.W. Todd, L. Yu, H. Fang, F. Jiang, Early detection of squamous cell lung cancer in sputum by a panel of microRNA markers. *Mod Pathol* 23 (2010) 1157-1164.
- [40]Y. Xie, N.W. Todd, Z. Liu, M. Zhan, H. Fang, H. Peng, M. Alattar, J. Deepak, S.A. Stass, F. Jiang, Altered miRNA expression in sputum for diagnosis of non-small cell lung cancer. *Lung Cancer* 67 (2010) 170-176.
- [41]F. Jiang, N.W. Todd, Q. Qiu, Z. Liu, R.L. Katz, S.A. Stass, Combined genetic analysis of sputum and computed tomography for noninvasive diagnosis of non-small-cell lung cancer. *Lung Cancer* 66 (2009) 58-63.
- [42]A.J. Hubers, D.A. Heideman, Y. Yatabe, M.D. Wood, J. Tull, M. Taron, M.A. Molina, C. Mayo, J. Bertran-Alamillo, G.J. Herder, R. Koning, D. Sie, B. Ylstra, G.A. Meijer, P.J. Snijders, B.I. Witte, P.E. Postmus, E.F. Smit, E. Thunnissen, EGFR mutation analysis in sputum of lung cancer patients: a multitechnique study. *Lung Cancer* 82 (2013) 38-43.
- [43]Y. Su, H. Fang, F. Jiang, Integrating DNA methylation and microRNA biomarkers in sputum for lung cancer detection. *Clin Epigenetics* 8 (2016) 109.

A Plasma Long Noncoding RNA Signature for Early Detection of Lung Cancer



Yanli Lin^{*,†,1}, Qixin Leng^{*,1}, Min Zhan[‡] and Feng Jiang[†]

^{*}Department of Cell Engineering, Beijing Institute of Biotechnology, No. 20 Dongdajie Street, Fengtai District, Beijing 100071, China; [†]Department of Pathology, University of Maryland School of Medicine, 10 S. Pine St. Baltimore, MD 21201, USA; [‡]Departments of Epidemiology & Public Health, University of Maryland School of Medicine, 660 W. Redwood St. Baltimore, MD 21201, USA

Abstract

The early detection of lung cancer is a major clinical challenge. Long noncoding RNAs (lncRNAs) have important functions in tumorigenesis. Plasma lncRNAs directly released from primary tumors or the circulating cancer cells might provide cell-free cancer biomarkers. The objective of this study was to investigate whether the lncRNAs could be used as plasma biomarkers for early-stage lung cancer. By using droplet digital polymerase chain reaction, we determined the diagnostic performance of 26 lung cancer-associated lncRNAs in plasma of a development cohort of 63 lung cancer patients and 33 cancer-free individuals, and a validation cohort of 39 lung cancer patients and 28 controls. In the development cohort, 7 of the 26 lncRNAs were reliably measured in plasma. Two (SNHG1 and RMRP) displayed a considerably high plasma level in lung cancer patients vs. cancer-free controls (all $P < .001$). Combined use of the plasma lncRNAs as a biomarker signature produced 84.13% sensitivity and 87.88% specificity for diagnosis of lung cancer, independent of stage and histological type of lung tumor, and patients' age and sex (all $P > .05$). The diagnostic value of the plasma lncRNA signature for lung cancer early detection was confirmed in the validation cohort. The plasma lncRNA signature may provide a potential blood-based assay for diagnosing lung cancer at the early stage. Nevertheless, a prospective study is warranted to validate its clinical value.

Translational Oncology (2018) 11, 1225–1231

Introduction

Approximately 155,870 Americans will die from lung cancer each year, more than the other 3 leading cancers combined (breast, prostate, and colorectal cancers). Over 85% lung cancers are non-small cell lung cancers (NSCLCs). NSCLC mainly consists of adenocarcinoma (AC) and squamous cell carcinoma (SCC). Tobacco smoking is the major cause of NSCLC. The early detection of lung cancer in a large randomized trial using low-dose CT (LDCT) has revealed a 20% reduction in mortality as compared to chest X-rays [1]. Therefore, LDCT is recently recommended to be used for lung cancer early detection among smokers [2,3]. However, LDCT is associated with overdiagnosis, excessive cost, and radiation exposure [2,3]. The development of noninvasive or circulating biomarkers that can accurately and cost-effectively diagnose early-stage lung cancer is required [4].

Long noncoding RNAs (lncRNAs) have minimum transcript length of 200 bp and play vital roles in various biological processes [5,6]. lncRNAs can regulate different molecular signaling pathways

via changing gene expression, and therefore, their dysregulations are implicated in numerous mechanisms of carcinogenesis [7,8]. Dysregulation of some lncRNAs has been found in relation to oncogenesis and metastasis of lung tumor [9–11]. Importantly, plasma lncRNAs directly released from primary tumors or the circulating cancer cells might provide biomarkers for human

Address all Correspondence to: Feng Jiang, Department of Pathology, The University of Maryland School of Medicine, 10 South Pine Street, MSTF 7th Floor, Baltimore, MD 21201-1192, USA. E-mail: fjiang@som.umaryland.edu

¹ The authors contributed equally to this work.

Received 14 May 2018; Revised 19 July 2018; Accepted 24 July 2018

© 2018 The Authors. Published by Elsevier Inc. on behalf of Neoplasia Press, Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>). 1936-5233/18

<https://doi.org/10.1016/j.tranon.2018.07.016>

malignancies [12]. To date, several plasma lncRNAs have been identified that show the potential to distinguish lung cancer patients from noncancer subjects [12–20]. Yet none of them has been accepted in the clinical settings for lung cancer diagnosis, mainly due to the low sensitivity and specificity.

Recent studies have characterized 21 lncRNAs whose aberrations are associated with lung cancer [10,21–46]. Furthermore, using whole-genomic next-generation sequencing (NGS) to analyze ncRNA profile of primary lung tumor tissues, we recently identified additional five lung cancer–related lncRNAs [47,48]. These lung tumor–associated lncRNAs may provide a comprehensive list of biomarker candidates for developing circulating lung cancer biomarkers. The objective of this study was to investigate whether the lung cancer–associated lncRNAs could be used as plasma biomarkers for lung cancer.

Materials and Methods

Patients and Clinical Specimens

This study was approved by the Institutional Review Boards of University of Maryland Baltimore and Veterans Affairs Maryland Health Care System. We recruited lung cancer patients and cancer-free smokers by using the inclusion and/or exclusion criteria recommended by the US Preventive Services Task Force for lung cancer screening in heavy smokers [49]. We collected blood in BD Vacutainer spray-coated K2EDTA Tubes (BD, Franklin Lakes, NJ) and prepared plasma using the standard operating protocols developed by The NCI-Early Detection Research Network [50,51]. The specimens were processed within 2 hours of collection by centrifugation at 1300×g for 10 minutes at 4°C. A total of 102 NSCLC patients and 55 cancer-free smokers were recruited. Among the cancer patients, 24 patients were female and 78 were male. Twenty-three had stage I NSCLC, 18 with stage II, 28 with stage III, 28 with stage IV, and 5 with unknown stage. Of the cancer-free smokers, 14 patients were female and 41 were male. There were no significant differences of age, gender, and smoking status between the NSCLC patients and cancer-free smokers. The cases and controls were randomly grouped into two cohorts: a development cohort and a validation cohort. The development cohort consisted of 63 lung cancer patients and 33 cancer-free smokers, while the validation cohort comprised 39 lung cancer patients and 28 cancer-free smokers. The demographic and clinical variables of the two cohorts are shown in Tables 1 and 2.

Table 1. Characteristics of NSCLC Patients and Cancer-Free Smokers in a Development Cohort

	NSCLC Cases (n = 63)	Controls (n = 33)	P Value
Age	67.93 (SD 9.16)	63.79 (SD 16.12)	.18
Sex			.36
Female	15	8	
Male	48	25	
Smoking pack-years (median)	32.1	31.76	.19
Stage			
Stage I	14		
Stage II	10		
Stage III	17		
Stage IV	18		
Unknown	4		
Histological type			
Adenocarcinoma	32		
Squamous cell carcinoma	31		

RNA Isolation and Quantitative Reverse Transcriptase Polymerase Chain Reaction (qRT-PCR)

RNA was extracted from the specimens by using Trizol LS reagent (Invitrogen, Carlsbad, CA) and RNeasy Mini Kit (Qiagen, Hilden, Germany) [52–54]. RT was carried out to generate cDNA by using an RT Kit (Applied Biosystems, Foster City, CA) as described in our published works [52–54]. PCR was performed to measure expressions of target genes by using a PCR kit (Applied Biosystems) on a Bio-Rad IQ5 Multicolor RT-PCR Detection System (Bio-Rad, Hercules, CA). Expression levels of the genes were determined using comparative cycle threshold (C_T) method with miR-1228 as an internal control. The targeted genes with C_T values >35 were considered to be below the detection level of qRT-PCR [55].

Droplet Digital PCR (ddPCR)

ddPCR for analysis of the genes was performed as described in our previous work. Briefly, TaqMan reaction mix (Applied Biosystems) containing sample cDNA was partitioned into aqueous droplets in oil via the QX100 Droplet Generator (Bio-Rad) and then transferred to a 96-well PCR plate. A two-step thermocycling protocol (95°C × 10 minutes; 40 cycles of 94°C × 30 seconds and 60°C × 60 seconds, and 98°C × 10 minutes) was undertaken in a Bio-Rad C1000 (Bio-Rad). The PCR plate was loaded on Droplet Reader (Bio-Rad), by which copy number of each gene per µl PCR reaction was directly determined. We used QuantaSoft 1.7.4 analysis software (Bio-Rad) and Poisson statistics to compute droplet concentrations (copies/µl). Only genes that had at least 10,000 droplets were considered to be robustly detectable by ddPCR in plasma and subsequently underwent further analysis [56,57]. All assays were done in triplicates, and one no-template control and two interplate controls were carried along in each experiment.

Statistical Analysis

Pearson correlation analysis was applied to assess relationship between gene expressions and demographic and clinical characteristics of the lung cancer patients and control individuals. The area under receiver operating characteristic curve (AUC) analyses were used to determine sensitivity, specificity, and corresponding cutoff value of each gene [58]. All P values shown were two sided, and a P value of < .05 was considered statistically significant.

Table 2. Characteristics of NSCLC Patients and Cancer-Free Smokers in a Validation Cohort

	NSCLC Cases (n = 39)	Controls (n = 28)	P Value
Age	66.58 (SD 9.93)	63.68 (SD 13.27)	.25
Sex			.45
Female	9	6	
Male	30	22	
Smoking pack-years (median)	33.39	29.64	.26
Stage			
Stage I	9		
Stage II	8		
Stage III	11		
Stage IV	10		
Unknown	1		
Histological type			
Adenocarcinoma	22		
Squamous cell carcinoma	17		

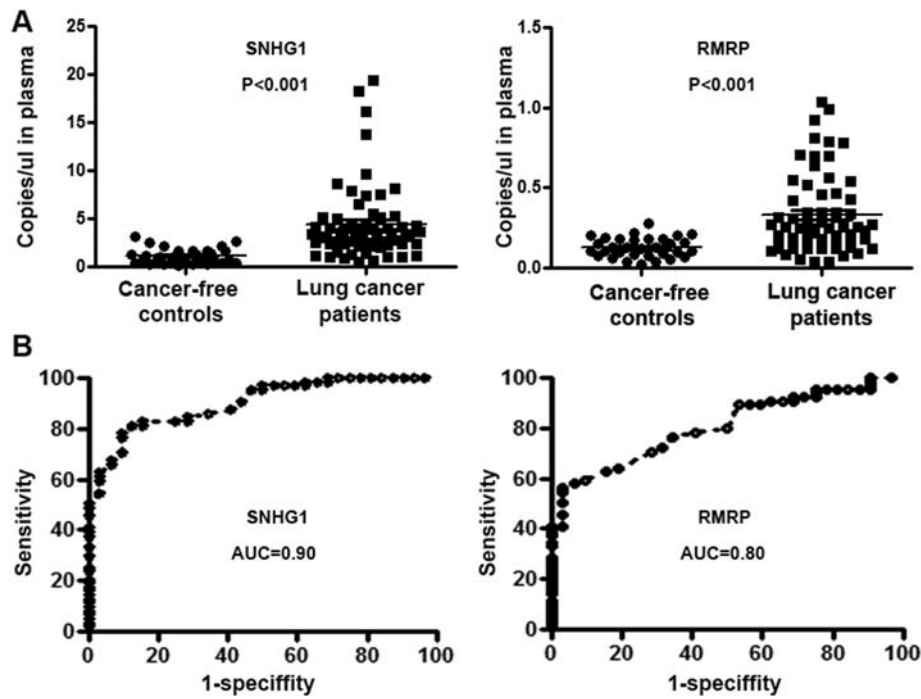


Figure 1. Expression levels of SNHG1 and RMRP in plasma samples of 63 lung cancer patients and 33 cancer-free controls. (A) SNHG1 and RMRP displayed a higher plasma level in lung cancer patients vs. cancer-free controls (all $P < .001$). (B) The receiver operating characteristic curves of SNHG1 and RMRP produced an AUC of 0.90 and 0.80, respectively, in diagnosis of lung cancer.

Results

Developing a Plasma lncRNA Signature for Lung Cancer Early Detection

We first measured expression levels of the 26 lncRNAs (Supplementary Table 1) in plasma by using qRT-PCR in a discovery cohort of 63 cases and 33 controls. The lncRNAs had a C_T value of ≥ 35 in 75% plasma samples. However, the internal control gene, miR-1228, stably displayed a C_T value of 20-22 across the plasma samples. The results suggested that the amplification curves for the lncRNAs were not reliably generated, and their expression levels in plasma were too low to be detectable by qRT-PCR. We have proven that ddPCR is a direct method for absolutely and quantitatively measuring ncRNAs with a higher sensitivity compared with qRT-PCR [56,57,59]. Therefore, we used ddPCR to determine expression level of the lncRNAs in the plasma samples. Seven (26.9%) of the 26 lncRNAs could generated at least 10,000 droplets in each well of the plasma samples. Therefore, the seven lncRNAs could be successfully “read” by ddPCR for the absolute quantification in the plasma samples. The seven genes are SNHG1, MALAT1, HOTAIR, H19, MEG3, MEG8, and RMRP.

Among the seven genes, SNHG1 and RMRP had a higher expression level in plasma of lung cancer patients versus cancer-free

controls (all $P \leq .05$) (Figure 1A). Conversely, other five lncRNAs did not display a different plasma level in lung cancer cases versus cancer-free controls (all $P \geq .05$). Furthermore, the plasma expression levels of the SNHG1 and RMRP were independent of stages and histological types of lung cancer. In addition, SNHG1 and RMRP exhibited AUC values of 0.90 and 0.80, respectively, in distinguishing NSCLC patients from healthy individuals (Figure 1B). Using Youden's index [60], we set up optimal cutoff for the two genes at 1.11 and 0.12, respectively. As a result, the use of the individual genes alone produced 61.00%-78.78% sensitivities and 87.88%-90.91% (Table 3). Combined use of the two genes based on at least one positive result in either SNHG1 or RMRP produced the highest classification accuracy (85.42%) compared to any one used alone (all $P < .05$) (Table 3). The two genes used in combination produced a sensitivity of 84.13% and a specificity of 87.88% for diagnosis of lung cancer, thus considerably improving the detection rate by a single gene with only a 2% decline in specificity (Table 3). Furthermore, the estimated correlation determined by Pearson correlation analysis among levels of the two lncRNAs was very low ($R^2 = 0.011$, $P = .53$), further supporting that the combined analysis of the two genes outperformed a single one. In addition, combined analysis of the two plasma biomarkers did not show special association with stage and histological type of lung cancer and patients' age, gender, and smoking status (all $P > .05$).

Table 3. Diagnostic Performance of One-Gene vs. a Plasma lncRNA Signature for Lung Cancer Diagnosis in a Development Cohort

	Accuracy	Sensitivity (95% CI)	Specificity (95% CI)
SNHG1	81.25% (72.00%-88.49%)	77.78% (65.54%-87.28%)	87.88% (71.80%-96.60%)
RMRP	71.88% (61.78%-80.58%)	61.90% (48.80%-73.85%)	90.91% (75.67%-98.08%)
A plasma lncRNA signature	85.42% (76.74%-91.79%)	84.13% (72.74%-92.12%)	87.50% (71.80%-96.60%)

Abbreviation: CI, confidence interval.

Table 4. Diagnostic Performance of One-Gene vs. a Plasma lncRNA Signature for Lung Cancer Diagnosis in a Validation Set

	Accuracy	Sensitivity (95% CI)	Specificity (95% CI)
<i>SNHG1</i>	80.33% (68.16%-89.40%)	76.92% (60.67%-88.87)	86.36% (65.09%-97.09%)
<i>RMRP</i>	72.13% (59.17%-82.85%)	61.54% (44.62%-76.64%)	90.91% (70.84%-98.88%)
A plasma lncRNA signature	83.62% (71.91%-91.85%)	82.05% (66.47%-92.46%)	86.36% (65.09%-97.09%)

Validating the Plasma lncRNA Marker Signature in an Independent Set of Lung Cancer Patients and Controls

To evaluate the diagnostic performance of the biomarker signature, the two lncRNAs (*SNHG1* and *RMRP*) were assessed by using ddPCR in plasma of additional 39 NSCLC patients and 28 healthy controls. The two genes used in combination could differentiate the NSCLC patients from healthy controls with 82.05% sensitivity and 83.36% specificity (Table 4). Furthermore, no statistically significant difference was found in the sensitivity and specificity of the biomarker signature for stages and histological types of NSCLC (all $P > .05$). Moreover, there was no association of expressions of the two genes with the age, gender, or smoking status of the lung cancer patients and normal individuals (all $P > .05$). Taken together, the results confirm the potential of combined use of the two lncRNAs as a plasma biomarker signature for the early detection of lung cancer.

Discussion

Circulating cell-free lncRNAs biomarkers show promise as biomarkers for cancer diagnosis. However, unlike other ncRNA (e.g., microRNAs), lncRNAs have the lowest levels in plasma among several different RNA species [61], presenting a major challenge in the development of cell-free lncRNAs as cancer biomarkers. For instance, Schlosser et al. recently demonstrated that expressions of lncRNAs were robustly detectable in tissues yet undetectable or sporadically measurable in the matched plasma by using qRT-PCR, a routine platform used for nucleic acid detection [61]. Regular qPCR has some limitations in determining expression of ncRNAs: i) It is an indirect and labor-consuming approach. ii) It requires an internal control gene for normalization. Yet none of the investigated genes has been accepted as a standard control. iii) Its sensitivity for a low copy number of genes is very low. Our current observations are consistent with Schlosser's finding [61]: of the 26 lung cancer-associated lncRNAs, none is reliably measurable in plasma using qRT-PCR. Therefore, conventional qPCR might not be an appropriate tool for the development of lncRNAs as circulating biomarkers given that circulating lncRNAs in body fluids are present in low abundance. We have shown that ddPCR is a direct method for sensitively measuring ncRNAs [56,59] since it depends on limiting partition of the PCR volume, where a positive result of a large number of microreactions indicates the presence of a single molecule in a given reaction [62]. The number of positive reactions, together with Poisson's distribution, produces a straight and high-confidence measurement of the original target concentration. Importantly, ddPCR does not require a reliance on rate-based measurements (C_T values), endogenous controls, and calibration curves and therefore overcome the obstacles linked to the regular qPCR in quantification of genes in plasma. Here we demonstrate that 7 of the 26 lung cancer-associated lncRNAs that are not detectable by qRT-PCR are robustly measurable by ddPCR in plasma. Therefore, ddPCR may address the limitations of the qPCR in quantification of lncRNAs in plasma and hence help develop the genes as cell-free cancer biomarkers.

The previous plasma lncRNA-based assays were mostly developed from the limited number of lung cancer-associated lncRNAs and

only consisted of a single lncRNA gene [12–20]. Since lung tumor is a heterogeneous group of neoplasms and develops from a multitude of molecular changes, a single lncRNA-based assay may not achieve the performance required to move forward for clinically detecting lung cancer. The development of a panel of multiple biomarkers by integrating analysis of multifaceted and diverse lncRNAs would provide a synergistic test for lung cancer diagnosis. By searching published data, we found 21 lncRNAs whose malfunction was well characterized in lung tumorigenesis [10,21–46]. Furthermore, by systematically and comprehensively defining ncRNA changes of NSCLC in surgical lung tumor tissues using whole-transcriptome NGS [47,48], we recently identified additional five lung cancer-associated lncRNAs [47]. Both the published and our NGS-defined lncRNAs of lung tumors may provide a comprehensive set of biomarker candidates for lung cancer. From the 26 lncRNAs, our present study identified and optimized a plasma signature consisting of two lncRNAs that created a higher diagnostic value for lung cancer detection than did individual lncRNAs used alone [12–20]. In addition, the diagnostic performance of the biomarkers was further blindly validated in a different cohort, suggesting that the plasma signature might be a robust assay for lung cancer diagnosis. Moreover, the performance of this plasma lncRNA signature for lung cancer diagnosis was independent of tumor stage and histology. This might be an important characteristic if the plasma lncRNA signature is employed for identifying early-stage lung cancer.

The two lncRNAs (*SNHG1* and *RMRP*) have diverse and important functions in lung tumorigenesis through regulating different molecular pathways. Elevated expression of *SNHG1* was frequently observed in lung cancer tissues and significantly correlated with larger tumor size, advanced stage, lymph node metastasis, and poor overall survival of the patients [63]. Furthermore, *SNHG1* could promote NSCLC progression of lung cancer via miR-101-3p/SOX9/Wnt/ β -catenin regulatory network and miR-145-5p/MTDH axis [63,64]. In addition, *SNHG1* plays an oncogenic role in lung squamous cell carcinoma through ZEB1 signaling pathway by inhibiting TAP63 [65]. *RMRP* is best known for being a component of the nuclear RNase MRP complex, which participates in the processing of ribosomal RNA to generate the short mature 5.8S rRNA [66] and cleaves B-cyclin mRNA, lowering B-cyclin levels during mitosis [67]. In addition, *RMRP* interacts with telomerase to form a complex with RNA-dependent RNA polymerase activity capable of synthesizing dsRNA precursors processed by DICER1 into siRNAs [68]. Moreover, *RMRP* is important for mitochondrial DNA replication and RNA processing [69]. Upregulation of *RMRP* is found in lung adenocarcinoma tissues [70]. *RMRP* might act as an oncogenic lncRNA to promote the expression of KRAS, FMNL2, and SOX9 by inhibiting miR-206 expression in lung cancer [70]. Our current study extends the previous findings by developing them as a biomarker signature that might be clinically useful in the early detection of lung cancer.

There are some limitations in this present study. 1) The size of the cohorts is small. Furthermore, the plasma samples were obtained from the hospital-based patients with clinical diagnosis. The participants might not be representative of high-risk populations (e.g., heavy

smokers) in screening setting for lung cancer. We will perform a prospective and multisite lung cancer screening trial to validate the diagnostic value of the plasma lncRNA signature. 2) The early diagnosis of lung cancer by using LDCT could reduce the mortality [3]. However, LDCT has a low specificity for the early detection of lung cancer, presenting a major clinical challenge [71]. We are evaluating whether the plasma lncRNA signature could improve the specificity of LDCT for the early detection of NSCLC by specifically distinguishing malignant from benign pulmonary growths. 3) The cell-free circulating tumor DNA, microRNAs, or DNA methylation provides other important types of biomarkers for lung cancer [54,72–75,76–78]. Our ongoing efforts are to investigate whether integrating the lncRNA signature with other types of biomarkers might improve the early detection of lung cancer. 4) In the current study, SNHG1 or RMRP expression level in plasma is independent of histological types of lung cancer. Furthermore, the combined use of the two genes is also not associated with histological types of lung cancer. The findings in plasma samples are inconsistent with the two previous studies in tissue specimens that showed that SNHG1 or RMRP was associated with SCC or AC, respectively [65,70]. A new study using paired plasma and tumor tissue specimens from the same lung cancer patients is needed to validate the discrepancy and further understand the mechanism underlying the divergence.

Conclusions

A plasma lncRNA signature was developed that could differentiate early-stage NSCLC patients from cancer-free smokers. Nevertheless, undertaking a prospective study to further validate this plasma biomarker signature in a large cohort is required.

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.tranon.2018.07.016>.

Declarations

Ethical Approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors.

The Institutional Review Boards of University of Maryland Baltimore and Veterans Affairs Maryland Health Care System approved this study. Reference number is H40666.

Competing Interests

The authors declare no conflict of interest.

Funding

Grant support: This work was supported in part by NCI R21CA205746, VA Merit Award I01 CX000512, Award from the Geaton and JoAnn DeCesaris Family Foundation, UMD-UMB Research and Innovation Seed Grant, DoD-Idea Development Award, and Maryland Innovation Initiative (MII) Commercialization Program-Phase 1/2 Grant (F.J.).

Authors' Contributions

Y. L., Q. L., M. Z., and F. J. conducted the experiments and participated in study design, coordination, data interpretation, and

preparing the manuscript. All authors read and approved the final manuscript.

Availability of Data and Materials

Not applicable.

Disclosure of Potential Conflicts of Interest

The authors declare no conflict of interest.

References

- [1] Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, Gareen IF, Gatsonis C, Marcus PM, and Sicks JD (2011). Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* **365**, 395–409.
- [2] Patz Jr EF, Pinsky P, Gatsonis C, Sicks JD, Kramer BS, Tammemagi MC, Chiles C, Black WC, and Aberle DR (2014). Overdiagnosis in low-dose computed tomography screening for lung cancer. *JAMA Intern Med* **174**(2), 269–274.
- [3] Aberle DR, Berg CD, Black WC, Church TR, Fagerstrom RM, Galen B, Gareen IF, Gatsonis C, Goldin J, and Gohagan JK, et al (2011). The National Lung Screening Trial: overview and study design. *Radiology* **258**, 243–253.
- [4] Hubers AJ, Prinsen CF, Sozzi G, Witte BI, and Thunnissen E (2013). Molecular sputum analysis for the diagnosis of lung cancer. *Br J Cancer* **109**(3), 530–537.
- [5] Ma L, Bajic VB, and Zhang Z (2013). On the classification of long non-coding RNAs. *RNA Biol* **10**, 925–933.
- [6] Spizzo R, Almeida MI, Colombatti A, and Calin GA (2012). Long non-coding RNAs and cancer: a new frontier of translational research? *Oncogene* **31**, 4577–4587.
- [7] Meseure D, Drak Alsbai K, Nicolas A, Bieche I, and Morillon A (2015). Long noncoding RNAs as new architects in cancer epigenetics, prognostic biomarkers, and potential therapeutic targets. *Biomed Res Int* **2015**, 320214.
- [8] Prensner JR and Chinnaiyan AM (2011). The emergence of lncRNAs in cancer biology. *Cancer Discov* **1**, 391–407.
- [9] Zhou M, Guo M, He D, Wang X, Cui Y, Yang H, Hao D, and Sun J (2015). A potential signature of eight long non-coding RNAs predicts survival in patients with non-small cell lung cancer. *J Transl Med* **13**, 231.
- [10] Li M, Qiu M, Xu Y, Mao Q, Wang J, Dong G, Xia W, Yin R, and Xu L (2015). Differentially expressed protein-coding genes and long noncoding RNA in early-stage lung cancer. *Tumour Biol* **36**, 9969–9978.
- [11] Schmidt LH, Spieker T, Koschmieder S, Schaffers S, Humberg J, Jungen D, Bulk E, Hascher A, Wittmer D, and Marra A, et al (2011). 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* **6**, 1984–1992.
- [12] Liang W, Lv T, Shi X, Liu H, Zhu Q, Zeng J, Yang W, Yin J, and Song Y (2016). Circulating long noncoding RNA GAS5 is a novel biomarker for the diagnosis of nonsmall cell lung cancer. *Medicine (Baltimore)* **95**(37)e4608.
- [13] Tantai J, Hu D, Yang Y, and Geng J (2015). Combined identification of long non-coding RNA XIST and HIF1A-AS1 in serum as an effective screening for non-small cell lung cancer. *Int J Clin Exp Pathol* **8**, 7887–7895.
- [14] Li N, Feng XB, Tan Q, Luo P, Jing W, Zhu M, Liang C, Tu J, and Ning Y (2017). Identification of circulating long noncoding RNA Linc00152 as a novel biomarker for diagnosis and monitoring of non-small-cell lung cancer. *Dis Markers* **17**, 743–749.
- [15] Li N, Wang Y, Liu X, Luo P, Jing W, Zhu M, and Tu J (2017). Identification of circulating long noncoding RNA HOTAIR as a novel biomarker for diagnosis and monitoring of non-small cell lung cancer. *Technol Cancer Res Treat* **17**, 1533–1539.
- [16] Wan L, Zhang L, Fan K, and Wang JJ (2017). Diagnostic significance of circulating long noncoding RNA PCAT6 in patients with non-small cell lung cancer. *Onco Targets Ther* **10**, 5695–5702.
- [17] Tan Q, Zuo J, Qiu S, Yu Y, Zhou H, Li N, Wang H, Liang C, Yu M, and Tu J (2017). Identification of circulating long non-coding RNA GAS5 as a potential biomarker for non-small cell lung cancer diagnosis, long non-coding RNA, plasma, GAS5, biomarker. *Int J Oncol* **50**, 1729–1738.
- [18] Zhu Q, Lv T, Wu Y, Shi X, Liu H, and Song Y (2017). Long non-coding RNA 00312 regulated by HOXA5 inhibits tumour proliferation and promotes apoptosis in non-small cell lung cancer. *J Cell Mol Med* **21**, 2184–2198.
- [19] Zhu H, Zhang L, Yan S, Li W, Cui J, Zhu M, Xia N, Yang Y, Yuan J, and Chen X, et al (2017). LncRNA16 is a potential biomarker for diagnosis of early-stage

- lung cancer that promotes cell proliferation by regulating the cell cycle. *Oncotarget* **8**, 7867–7877.
- [20] Wang HM, Lu JH, Chen WY, and Gu AQ (2015). Upregulated lncRNA-UCA1 contributes to progression of lung cancer and is closely related to clinical diagnosis as a predictive biomarker in plasma. *Int J Clin Exp Med* **8**, 11824–11830.
- [21] Wei MM, Zhou YC, Wen ZS, Zhou B, Huang YC, Wang GZ, Zhao XC, Pan HL, Qu LW, and Zhang J, et al (2016). Long non-coding RNA stabilizes the Y-box-binding protein 1 and regulates the epidermal growth factor receptor to promote lung carcinogenesis. *Oncotarget* **7**, 59556–59571.
- [22] Shen L, Chen L, Wang Y, Jiang X, Xia H, and Zhuang Z (2015). Long noncoding RNA MALAT1 promotes brain metastasis by inducing epithelial-mesenchymal transition in lung cancer. *J Neurooncol* **121**, 101–108.
- [23] Loewen G, Jayawickramarajah J, Zhuo Y, and Shan B (2014). Functions of lncRNA HOTAIR in lung cancer. *J Hematol Oncol* **7**, 90–96.
- [24] Li P, Li J, Yang R, Zhang F, Wang H, Chu H, Lu Y, Dun S, Wang Y, and Zang W, et al (2015). Study on expression of lncRNA RGMB-AS1 and repulsive guidance molecule b in non-small cell lung cancer. *Diagn Pathol* **10**, 63–68.
- [25] Yang YR, Zang SZ, Zhong CL, Li YX, Zhao SS, and Feng XJ (2014). Increased expression of the lncRNA PVT1 promotes tumorigenesis in non-small cell lung cancer. *Int J Clin Exp Pathol* **7**, 6929–6935.
- [26] Whiteside EJ, Seim I, Pauli JP, O'Keeffe AJ, Thomas PB, Carter SL, Walpole CM, Fung JN, Josh P, and Herington AC, et al (2013). Identification of a long non-coding RNA gene, growth hormone secretagogue receptor opposite strand, which stimulates cell migration in non-small cell lung cancer cell lines. *Int J Oncol* **43**, 566–574.
- [27] Hu T and Lu YR (2015). BCYRN1, a c-MYC-activated long non-coding RNA, regulates cell metastasis of non-small-cell lung cancer. *Cancer Cell Int* **15**, 36–40.
- [28] Li J, Li P, Zhao W, Yang R, Chen S, Bai Y, Dun S, Chen X, Du Y, and Wang Y, et al (2015). Expression of long non-coding RNA DLX6-AS1 in lung adenocarcinoma. *Cancer Cell Int* **15**, 48–53.
- [29] Zeng Z, Bo H, Gong Z, Lian Y, Li X, Zhang W, Deng H, Zhou M, Peng S, and Li G, et al (2016). AFAP1-AS1, a long noncoding RNA upregulated in lung cancer and promotes invasion and metastasis. *Tumour Biol* **37**, 729–737.
- [30] Hou Z, Zhao W, Zhou J, Shen L, Zhan P, Xu C, Chang C, Bi H, Zou J, and Yao X, et al (2014). A long noncoding RNA Sox2ot regulates lung cancer cell proliferation and is a prognostic indicator of poor survival. *Int J Biochem Cell Biol* **53**, 380–388.
- [31] Luo J, Tang L, Zhang J, Ni J, Zhang HP, Zhang L, Xu JF, and Zheng D (2014). Long non-coding RNA CARLo-5 is a negative prognostic factor and exhibits tumor pro-oncogenic activity in non-small cell lung cancer. *Tumour Biol* **35**, 11541–11549.
- [32] Luo H, Sun Y, Wei G, Luo J, Yang X, Liu W, Guo M, and Chen R (2015). Functional characterization of long noncoding RNA Lnc_bc060912 in human lung carcinoma cells. *Biochemistry* **54**, 2895–2902.
- [33] Wu Y, Liu H, Shi X, Yao Y, Yang W, and Song Y (2015). The long non-coding RNA HNF1A-AS1 regulates proliferation and metastasis in lung adenocarcinoma. *Oncotarget* **6**, 9160–9172.
- [34] Qiu M, Xu Y, Yang X, Wang J, Hu J, Xu L, and Yin R (2014). CCAT2 is a lung adenocarcinoma-specific long non-coding RNA and promotes invasion of non-small cell lung cancer. *Tumour Biol* **35**, 5375–5380.
- [35] Qiu M, Xu Y, Wang J, Zhang E, Sun M, Zheng Y, Li M, Xia W, Feng D, and Yin R, et al (2015). A novel lncRNA, LUADT1, promotes lung adenocarcinoma proliferation via the epigenetic suppression of p27. *Cell Death Dis* **6**, 18–22.
- [36] Zhang L, Zhou XF, Pan GF, and Zhao JP (2014). Enhanced expression of long non-coding RNA ZXF1 promoted the invasion and metastasis in lung adenocarcinoma. *Biomed Pharmacother* **68**, 401–407.
- [37] Nie FQ, Sun M, Yang JS, Xie M, Xu TP, Xia R, Liu YW, Liu XH, Zhang EB, and Lu KH, et al (2015). Long noncoding RNA ANRIL promotes non-small cell lung cancer cell proliferation and inhibits apoptosis by silencing KLF2 and P21 expression. *Mol Cancer Ther* **14**, 268–277.
- [38] Yang R, Li P, Zhang G, Lu C, Wang H, and Zhao G (2017). Long non-coding RNA XLOC_008466 functions as an oncogene in human non-small cell lung cancer by targeting miR-874. *Cell Physiol Biochem* **42**, 126–136.
- [39] Sang H, Liu H, Xiong P, and Zhu M (2015). Long non-coding RNA functions in lung cancer. *Tumour Biol* **36**, 4027–4037.
- [40] Shi X, Sun M, Liu H, Yao Y, Kong R, Chen F, and Song Y (2015). A critical role for the long non-coding RNA GAS5 in proliferation and apoptosis in non-small-cell lung cancer. *Mol Carcinog* **54**, 1–12.
- [41] Han L, Kong R, Yin DD, Zhang EB, Xu TP, De W, and Shu YQ (2013). Low expression of long noncoding RNA GAS6-AS1 predicts a poor prognosis in patients with NSCLC. *Med Oncol* **30**, 694–699.
- [42] Han L, Zhang EB, Yin DD, Kong R, Xu TP, Chen WM, Xia R, Shu YQ, and De W (2015). Low expression of long noncoding RNA PANDAR predicts a poor prognosis of non-small cell lung cancer and affects cell apoptosis by regulating Bcl-2. *Cell Death Dis* **6**, 16–22.
- [43] Xie X, Liu HT, Mei J, Ding FB, Xiao HB, Hu FQ, Hu R, and Wang MS (2014). lncRNA HMLincRNA717 is down-regulated in non-small cell lung cancer and associated with poor prognosis. *Int J Clin Exp Pathol* **7**, 8881–8886.
- [44] Liu J, Wan L, Lu K, Sun M, Pan X, Zhang P, Lu B, Liu G, and Wang Z (2015). The long noncoding RNA MEG3 contributes to cisplatin resistance of human lung adenocarcinoma. *PLoS One* **10**, 11–17.
- [45] Sun M, Liu XH, Wang KM, Nie FQ, Kong R, Yang JS, Xia R, Xu TP, Jin FY, and Liu ZJ, et al (2014). Downregulation of BRAF activated non-coding RNA is associated with poor prognosis for non-small cell lung cancer and promotes metastasis by affecting epithelial-mesenchymal transition. *Mol Cancer* **13**, 68–72.
- [46] Yang Y, Li H, Hou S, Hu B, Liu J, and Wang J (2013). The noncoding RNA expression profile and the effect of lncRNA AK126698 on cisplatin resistance in non-small-cell lung cancer cell. *PLoS One* **8**, 65–60.
- [47] Ma J, Mannoor K, Gao L, Tan A, Guarnera MA, Zhan M, Shetty A, Stass SA, Xing L, and Jiang F (2014). Characterization of microRNA transcriptome in lung cancer by next-generation deep sequencing. *Mol Oncol* **8**, 1208–1219.
- [48] Gao L, Ma J, Mannoor K, Guarnera MA, Shetty A, Zhan M, Xing L, Stass SA, and Jiang F (2015). Genome-wide small nucleolar RNA expression analysis of lung cancer by next-generation deep sequencing. *Int J Cancer* **136**, 623–629.
- [49] Humphrey LL, Deffebach M, Pappas M, Baumann C, Artis K, Mitchell JP, Zakher B, Fu R, and Slatore CG (2013). Screening for lung cancer with low-dose computed tomography: a systematic review to update the US Preventive services task force recommendation. *Ann Intern Med* **159**, 411–420.
- [50] Marks JR, Anderson KS, Engstrom P, Godwin AK, Esserman LJ, Longton G, Iversen ES, Mathew A, Patriotis C, and Pepe MS (2015). Construction and analysis of the NCI-EDRN breast cancer reference set for circulating markers of disease. *Cancer Epidemiol Biomarkers Prev* **24**, 435–441.
- [51] Tuck MK, Chan DW, Chia D, Godwin AK, Grizzle WE, Krueger KE, Rom W, Sanda M, Sorbara L, and Stass S, et al (2009). Standard operating procedures for serum and plasma collection: early detection research network consensus statement standard operating procedure integration working group. *J Proteome Res* **8**, 113–117.
- [52] Ma J, Jemal A, and Smith R (2013). Reply to lung cancer deaths averted by screening should be considered in the context of tobacco control policies. *Cancer* **119**, 3420–3421.
- [53] Shen J, Liu Z, Todd NW, Zhang H, Liao J, Yu L, Guarnera MA, Li R, Cai L, and Zhan M, et al (2011). Diagnosis of lung cancer in individuals with solitary pulmonary nodules by plasma microRNA biomarkers. *BMC Cancer* **11**, 374–379.
- [54] Shen J, Todd NW, Zhang H, Yu L, Lingxiao X, Mei Y, Guarnera M, Liao J, Chou A, and Lu CL, et al (2011). Plasma microRNAs as potential biomarkers for non-small-cell lung cancer. *Lab Invest* **91**(4), 579–587.
- [55] Guthrie JL, Seah C, Brown S, Tang P, Jamieson F, and Drews SJ (2008). Use of Bordetella pertussis BP3385 to establish a cutoff value for an IS481-targeted real-time PCR assay. *J Clin Microbiol* **46**, 3798–3799.
- [56] Li N, Ma J, Guarnera MA, Fang H, Cai L, and Jiang F (2014). Digital PCR quantification of miRNAs in sputum for diagnosis of lung cancer. *J Cancer Res Clin Oncol* **140**, 145–150.
- [57] Ma J, Li N, Guarnera M, and Jiang F (2013). Quantification of plasma miRNAs by digital PCR for cancer diagnosis. *Biomark Insights* **8**, 127–136.
- [58] Dodd LE and Pepe MS (2003). Partial AUC estimation and regression. *Biometrics* **59**, 614–623.
- [59] Li H, Jiang Z, Leng Q, Bai F, Wang J, Ding X, Li Y, Zhang X, Fang H, and Yfantis HG, et al (2017). A prediction model for distinguishing lung squamous cell carcinoma from adenocarcinoma. *Oncotarget* **8**, 50704–50714.
- [60] Schisterman EF, Perkins NJ, Liu A, and Bondell H (2005). Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. *Epidemiology* **16**, 73–81.
- [61] Schlosser K, Hanson J, Villeneuve PJ, Dimitrakoulas J, McIntyre L, Pilote L, and Stewart DJ (2016). Assessment of circulating lncRNAs under physiologic and pathologic conditions in humans reveals potential limitations as biomarkers. *Sci Rep* **6**, 36–39.
- [62] Day E, Dear PH, and McCaughan F (2013). Digital PCR strategies in the development and analysis of molecular biomarkers for personalized medicine. *Methods* **59**, 101–107.
- [63] Cui Y, Zhang F, Zhu C, Geng L, Tian T, and Liu H (2017). Upregulated lncRNA SNHG1 contributes to progression of non-small cell lung cancer

- through inhibition of miR-101-3p and activation of Wnt/beta-catenin signaling pathway. *Oncotarget* **8**, 17785–17794.
- [64] Lu Q, Shan S, Li Y, Zhu D, Jin W, and Ren T (2018). Long noncoding RNA SNHG1 promotes non-small cell lung cancer progression by up-regulating MTDH via sponging miR-145-5p. *FASEB J* **20**, 17–23.
- [65] Zhang HY, Yang W, Zheng FS, Wang YB, and Lu JB (2017). Long non-coding RNA SNHG1 regulates zinc finger E-box binding homeobox 1 expression by interacting with TAp63 and promotes cell metastasis and invasion in lung squamous cell carcinoma. *Biomed Pharmacother* **90**, 650–658.
- [66] Schmitt ME and Clayton DA (1993). Nuclear RNase MRP is required for correct processing of pre-5.8S rRNA in *Saccharomyces cerevisiae*. *Mol Cell Biol* **13**, 7935–7941.
- [67] Noh JH, Kim KM, Abdelmohsen K, Yoon JH, Panda AC, Munk R, Kim J, Curtis J, Moad CA, and Wohler CM, et al (2016). HuR and GRSF1 modulate the nuclear export and mitochondrial localization of the lncRNA RMRP. *Genes Dev* **30**, 1224–1239.
- [68] Maida Y, Yasukawa M, Furuuchi M, Lassmann T, Possemato R, Okamoto N, Kasim V, Hayashizaki Y, Hahn WC, and Masutomi K (2009). An RNA-dependent RNA polymerase formed by TERT and the RMRP RNA. *Nature* **461**, 230–235.
- [69] Chang DD and Clayton DA (1987). A mammalian mitochondrial RNA processing activity contains nucleus-encoded RNA. *Science* **235**, 1178–1184.
- [70] Meng Q, Ren M, Li Y, and Song X (2016). LncRNA-RMRP acts as an oncogene in lung cancer. *PLoS One* **11**, 164–184.
- [71] Marcus PM (2015). Lung cancer screening with low dose computed tomography (LDCT): looking back and moving forward. *Ann Transl Med* **3**, 41–46.
- [72] Leng Q, Lin Y, Jiang F, Lee CJ, Zhan M, Fang H, and Wang Y (2017). A plasma miRNA signature for lung cancer early detection. *Oncotarget* **8**, 111902–111911.
- [73] Lin Y, Leng Q, Jiang Z, Guarnera MA, Zhou Y, Chen X, Wang H, Zhou W, Cai L, and Fang H, et al (2017). A classifier integrating plasma biomarkers and radiological characteristics for distinguishing malignant from benign pulmonary nodules. *Int J Cancer* **141**, 1240–1248.
- [74] Ma J, Li N, Lin Y, Gupta C, and Jiang F (2016). Circulating neutrophil microRNAs as biomarkers for the detection of lung cancer. *Biomark Cancer* **8**, 1–7.
- [75] Ma J, Lin Y, Zhan M, Mann DL, Stass SA, and Jiang F (2015). Differential miRNA expressions in peripheral blood mononuclear cells for diagnosis of lung cancer. *Lab Invest* **95**(10), 1197–1206.
- [76] Shen J and Jiang F (2012). Applications of microRNAs in the diagnosis and prognosis of lung cancer. *Expert Opin Med Diagn* **6**, 197–207.
- [77] Normanno N, Denis MG, Thress KS, Ratcliffe M, and Reck M (2017). Guide to detecting epidermal growth factor receptor (EGFR) mutations in ctDNA of patients with advanced non-small-cell lung cancer. *Oncotarget* **8**, 12501–12516.
- [78] Balgkouranidou I, Chimonidou M, Milaki G, Tsaroucha E, Kakolyris S, Georgoulas V, and Lianidou E (2016). SOX17 promoter methylation in plasma circulating tumor DNA of patients with non-small cell lung cancer. *Clin Chem Lab Med* **54**, 1385–1393.

An integromic signature for lung cancer early detection

Qixin Leng^{1,*}, Yanli Lin^{1,*}, Min Zhan² and Feng Jiang¹

¹Department of Pathology, University of Maryland School of Medicine, Baltimore, MD 21201, USA

²Department of Epidemiology & Public Health, University of Maryland School of Medicine, Baltimore, MD 21201, USA

*These authors have contributed equally to this work

Correspondence to: Feng Jiang, **email:** fjiang@som.umaryland.edu

Keywords: diagnosis; early stage; lung cancer; plasma; biomarkers

Received: March 24, 2018

Accepted: April 07, 2018

Published: May 15, 2018

Copyright: Leng et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License 3.0 (CC BY 3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

ABSTRACT

We previously developed three microRNAs (miRs-21, 210, and 486-5p), two long noncoding RNAs (lncRNAs) (SNHG1 and RMRP), and two fucosyltransferase (FUT) genes (FUT8 and POFUT1) as potential plasma biomarkers for lung cancer. However, the diagnostic performance of the individual panels is not sufficient to be used in the clinics. Given the heterogeneity of lung tumors developed from multifactorial molecular aberrations, we determine whether integrating the different classes of molecular biomarkers can improve diagnosis of lung cancer. By using droplet digital PCR, we analyze expression of the seven genes in plasma of a development cohort of 64 lung cancer patients and 33 cancer-free individuals. The panels of three miRNAs (miRs-21, 210, and 486-5p), two lncRNAs (SNHG1 and RMRP), and two FUTs (FUT8 and POFUT1) have a sensitivity of 81-86% and a specificity of 84-87% for diagnosis of lung cancer. From the seven genes, an integromic plasma signature comprising miR-210, SNHG1, and FUT8 is developed that produces higher sensitivity (95.45%) and specificity (96.97%) compared with the individual biomarker panels (all $p < 0.05$). The diagnostic value of the signature was confirmed in a validation cohort of 40 lung cancer patients and 29 controls, independent of stage and histological type of lung tumor, and patients' age, sex, and smoking status (all $p > 0.05$). The integration of the different categories of biomarkers might improve diagnosis of lung cancer.

INTRODUCTION

Over 85% lung cancers are non-small cell lung cancers (NSCLC). NSCLC mainly consists of adenocarcinoma (AC) and squamous cell carcinoma (SCC). Tobacco smoking is the major cause of NSCLC. Since the prognosis for patients with lung cancer is strongly correlated to the tumor stage, diagnosing lung cancer at a curable stage can reduce the mortality [1]. The early detection of lung cancer in a large randomized trial using low-dose CT (LDCT) has revealed a 20% reduction in mortality as compared to chest X-rays [1]. However, LDCT is associated with over-diagnosis, excessive cost, and radiation exposure [2, 3]. The development

of circulating biomarkers that can accurately and cost-effectively identify early stage lung cancer is required [4].

During tumor development, cancer cells undergo apoptosis and necrosis, and release tumor-associated molecules that can circulate in bloodstream. The tumors-derived molecules in plasma provide cell-free circulating cancer biomarkers. Regulatory non-coding RNAs (ncRNAs) can be classified into two major classes based on the transcript size: small ncRNAs (<200 bp) including microRNAs (miRNAs) and long ncRNAs (lncRNAs) (>200 bp) [5, 6]. Through different molecular mechanisms or pathways, the two types of ncRNAs have diverse and critical functions in tumorigenesis [7-10]. Furthermore, plasma miRNAs and lncRNAs directly released from primary lung tumors or the circulating lung cancer cells

might provide cell-free biomarkers for lung cancer [8]. For instance, we recently developed a panel of three plasma miRNA biomarkers with 86% sensitivity and 87% specificity and a panel of two plasma lncRNA biomarkers with 83% sensitivity and 84% specificity for lung cancer early detection [7, 11-15].

Emerging evidences have demonstrated that aberrant glycosylation leads to cancer development and progression [16]. Fucosylation is the major type of glycosylation, and regulated by fucosyltransferases (FUTs) [16, 17]. We recently found that combined use of two plasma FUTs (FUT8 and POFUT1) had 81% sensitivity and 84% specificity for diagnosis of lung cancer, thus providing a new category of cell-free circulating biomarkers for lung cancer.

Since NSCLC is a heterogeneous disease and develops from multifactorial molecular aberrations [18], the analysis of one type of molecular changes may not achieve the performance required to move forward for clinical application. Indeed, although our individual panels of plasma biomarkers show promise for lung cancer diagnosis, their sensitivities (81-86%) and specificities (84-87%) are not sufficient to be used in the laboratory settings.

Because miRNAs, lncRNAs, and FUTs have highly diverse roles that drive the development of lung cancer [7-10, 16], we hypothesize that integrating the different classes of biomarkers may improve the early detection of lung cancer. Here we evaluate the individual and combined applications of the three categories of plasma molecular biomarkers for lung cancer.

RESULTS

The three individual panels of plasma biomarkers displayed a different level in NSCLC patients vs. smokers

Droplet Digital PCR (ddPCR) was used for quantification of the genes (miRs-21, 210, 486-5p, SNHG1, RMRP, FUT8, and POFUT1) in plasma of a development cohort of 64 lung cancer patients and 33 cancer-free individuals. All the seven genes generated at least 10,000 droplets in each well of the plasma samples. Therefore, the seven genes could be successfully “read” by ddPCR for their absolute quantification in plasma. These genes had a significantly different expression level in plasma of the NSCLC patients compared with the control individuals (all $P < 0.05$). As a result, the individual genes resulted in 50.09 to 75.76% sensitivities and 63.64 to 90.91% specificities for detection of NSCLC (Table 1). Furthermore, the panel of three microRNA biomarkers (miRs-21, 210, and 486-5p) had an area under receiver operating characteristic curve (AUC) of 0.92 with 86.36% sensitivity and 87.88% specificity, the panel of two plasma lncRNA biomarkers (SNHG1 and RMRP) displayed 0.89

AUC with 83.33% sensitivity and 84.85% specificity, and the panel of two FUTs (FUT8 and POFUT1) exhibited an AUC of 0.85 with 81.82% sensitivity and 84.85% specificity for diagnosis of lung cancer (Table 2). The individual panels of the genes didn't show special association with stage and histology of the NSCLC, age, gender, and smoking status of the participants (All $p > 0.05$). The seven genes would be potential plasma biomarkers for lung cancer.

An integromic plasma signature for lung cancer early detection

We used logistic regression models with constrained parameters as in least absolute shrinkage and selection operator (LASSO) and AUCs to determine performance of different patterns of combining the genes. From the seven genes, one miRNA (miR-210), one lncRNA (SNHG1), and one FUT (FUT8) were selected as the best biomarkers (all $P < 0.001$). A logisitic regression model with each of the different types of genes was developed as an integromic signature for diagnosing lung cancer: $U = -7.29 + 2.8 \times \log(\text{SNHG1}) + 3.83 \times \log(\text{FUT8}) + 3.36 \times \log(\text{miR-210})$. Combined analysis of the 3 biomarkers by using the logisitic regression model produced a higher AUC (0.97) (Figure 1) than did the individual panels of biomarkers ($p < 0.05$). We used the highest Youden's J index to set up corresponding cut-off value [19]. The optimal cut-off for the integromic signature was $U = 0.79$. Any subject with $U \geq 0.79$ was classified as a lung cancer case. As a result, the integromic plasma signature yielded significantly higher sensitivity (95.45%), specificity (96.97%), and accuracy (95.96%) compared with the individual panels of biomarkers (all $p < 0.05$) (Table 2). Furthermore, combined use of all the seven genes did not produce higher sensitivity and specificity compared with the integromic plasma signature ($p > 0.05$). In addition, Pearson's correlation analysis showed that the relationships among levels of the three genes were very low (All $p > 0.05$), implying that the integration of the different classes of molecular biomarkers has complementary classification. Moreover, the integromic plasma signature had no special association with histological type of the NSCLC, age, gender, and smoking status of the participants (All $p > 0.05$). The integromic signature did not show statistical difference of sensitivity and specificity for different stages of NSCLC (Supplementary Figure 1).

Validating the integromic plasma signature for lung cancer detection

The plasma expression levels of the three genes (miR-210, SNHG1, and FUT8) were assessed by using ddPCR in a validation cohort of additional 40 NSCLC patients and 29 healthy controls. Combined analysis of

Table 1: Diagnostic performance of individual genes for lung cancer in a development cohort

	Sensitivity (95% CI)	Specificity (95% CI)
miRs-21	75.76% (63.64% to 85.46%)	63.64% (45.12% to 79.60%)
miR-210	50.09% (46.29% to 71.05%)	72.73% (54.48% to 86.70%)
miR-486-5p	72.73% (60.36% to 82.97%)	63.64% (45.12% to 79.60%)
SNHG1	75.76% (63.64% to 85.46%)	81.82% (64.54% to 93.02%)
RMRP	62.12% (49.34% to 73.78%)	90.91% (75.67% to 98.08%)
FUT8	71.21 (58.75% to 81.70%)	87.88% (71.80% to 96.60%)
POFUT1	60.61% (47.81% to 72.42%)	90.91% (75.67% to 98.08%)

Abbreviations: CI, confidence interval.

Table 2: The area under receiver operating characteristic curves (AUCs) of the individual panels of biomarkers and the plasma integromic signature in a development cohort

	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy
A panel of 3 miRNAs	0.92 (0.87 to 0.97)	86.36% (75.69% to 93.57%)	87.88% (71.80% to 96.60%)	86.87% (78.59% to 92.82%)
A panel of 2 lncRNAs	0.89 (0.85 to 0.93)	83.33% (72.13% to 91.38%)	84.85% (68.10% to 94.89%)	83.84% (75.09% to 90.47%)
A panel of 2 FUTs	0.85 (0.78 to 0.91)	81.82% (70.39% to 90.24%)	84.85% (68.10% to 94.89%)	82.83% (73.94% to 89.67%)
An integromic signature	0.95 (0.91 to 0.99)	95.45% (87.29% to 99.05%)	96.97% (84.24% to 99.92%)	95.96% (89.98% to 98.89%)

Abbreviations: CI, confidence interval.

the three genes by using the logistic regression model created 0.94 AUC for lung cancer diagnosis. There was no significant difference between the develop cohort and validation cohort with regarding the signature's AUCs (0.95 vs. 0.94, $p=0.46$) (Figure 2). In the validation cohort, the three genes used in combination could differentiate the NSCLC patients from healthy controls with a sensitivity of 95.00% (82.08% to 99.12%) and a specificity of 96.55% (80.37% to 99.82%). In line with the findings in the development cohort, the integromic plasma signature did not show statistical difference of sensitivity and specificity across different stages and subtypes of NSCLC (all $p>0.05$). Moreover, there was no association of expressions of the genes with the age, gender, or smoking status of the individuals (All $p>0.05$).

DISCUSSION

Although showing promise, the use of the individual miRNA, lncRNA, or FUT biomarker panels alone has moderate sensitivities (81-86%) and specificities (84-87%). The miRNAs, lncRNAs, and FUTs have highly different functions in carcinogenesis [20-22]. Given the heterogeneous nature of lung cancer and the numerous

cellular pathways involved, we hypothesize that integrating the different classes of molecular biomarkers may improve the early detection of lung cancer. Intriguingly, the integrated analysis of only one of each type of biomarkers by using a single platform (ddPCR) yields a significantly higher diagnostic performance compared with any panel of one type of genes. Furthermore, the correlations among the changes of the miRNAs, lncRNA, and FUT are very low, supporting that the diagnostic vales of the three classes of molecular alterations could be complementary to each other. Therefore, the observation confirms our hypothesis. Moreover, since the integromic plasma signature shows similar sensitivity and specificity in the early vs. advanced stages of NSCLC, it might be a useful approach for the early detection of lung cancer, a clinically challenging.

miR-210 stimulates a hypoxic phenotype and upsurges radioresistance in NSCLCs [21]. Hypoxia-induced miR-210 can regulate tumor cell susceptibility to cytolytic T-lymphocyte-mediated lysis by a mechanism involving its downstream targets PTPN1, HOXA1, and TP53I11 [23]. SNHG1 could promote NSCLC progression of lung cancer via miR-101-3p/SOX9/Wnt/ β -catenin regulatory network and miR-145-5p/MTDH axis [14, 22]. In addition, SNHG1 plays

an oncogenic role in lung squamous cell carcinoma through ZEB1 signaling pathway by inhibiting TAp63 [24]. FUT8 inhibits the malignant behaviors of lung cancer cells and is involved in the regulation of dozens of genes associated with the malignancy through multiple mechanisms [20]. Upregulation of FUT8 could contribute to epithelial-mesenchymal transition via the transactivation of β -catenin/lymphoid enhancer-binding factor-1 (LEF-1) [20].

The study does have some limitations. 1), the sample size of the cohorts is small. We will perform a new study to prospectively validate the integromic signature for lung cancer early detection using a large population. 2), it is well known that lung cancer-associated molecular genetic changes are also related to chronic obstructive pulmonary disease (COPD) [25]. Many lung cancer patients who are smokers and cancer-free heavy smokers have COPD [25]. COPD could impact molecular genetic profiles in plasma

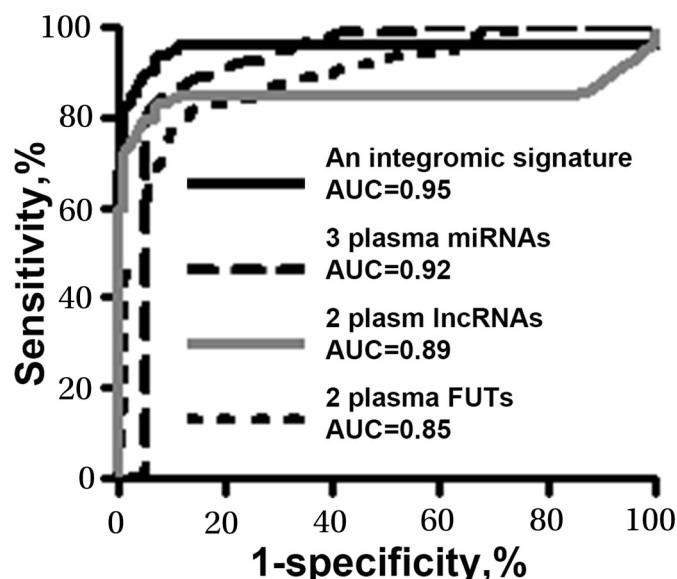


Figure 1: Diagnostic values of the individual panels of biomarkers and integromic plasma signature in a development cohort. The integromic plasma biomarker signature yields a higher area under receiver operating characteristic curve (AUC) than does individual panels of biomarkers (All $p < 0.05$).

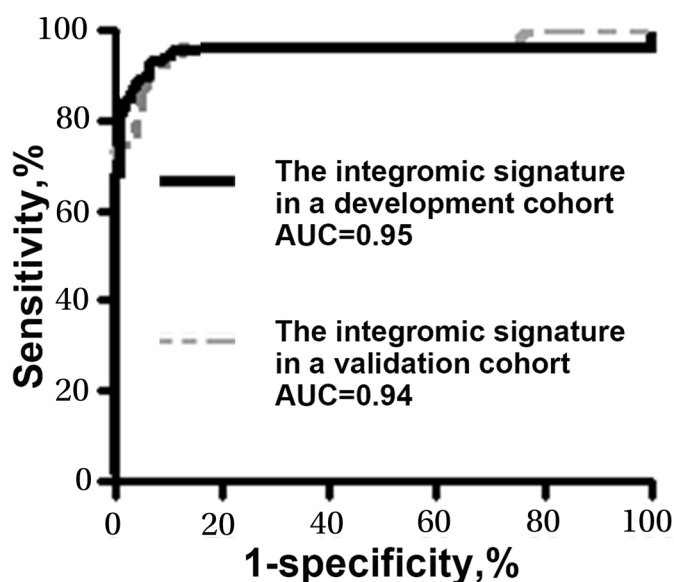


Figure 2: Comparison of AUCs of the integromic plasma signature for diagnosis of lung cancer in different cohorts. The integromic plasma signature has no significant difference of AUCs in the development cohort (black line) vs. validation cohort (gray line) (0.95 vs. 0.94, $p = 0.46$).

Table 3: Characteristics of a development cohort and a validation cohort

	A development cohort		P-value
	NSCLC cases (n = 66)	Controls (n = 33)	
Age	67.68 (SD 9.23)	62.70 (SD 15.33)	0.16
Sex			0.39
Female	17	8	
Male	49	25	
Smoking pack-years (median)	33.6	32.69	0.17
Stage			
Stage I	15		
Stage II	11		
Stage III	17		
Stage IV	19		
Unknown	4		
Histological type			
Adenocarcinoma	34		
Squamous cell carcinoma	32		

	A validation cohort		P-value
	NSCLC cases (n = 40)	Controls (n = 29)	
Age	64.57 (SD 9.28)	63.76 (SD 13.49)	0.28
Sex			0.46
Female	10	8	
Male	30	21	
Smoking pack-years (median)	33.68	31.63	0.23
Stage			
Stage I	9		
Stage II	8		
Stage III	11		
Stage IV	11		
Unknown	1		
Histological type			
Adenocarcinoma	22		
Squamous cell carcinoma	18		

Abbreviations: NSCLC, non-small cell lung cancer.

of both lung cancer patients and cancer-free heavy smokers. In this present project, there is no COPD information of the cases and controls enrolled. Therefore, we are not able to evaluate if the biomarkers identified in the study are associated with COPD. We will recruit lung cancer patients

and cancer-free smokers who have COPD, and determine if COPD is the confounding effect on the molecular changes. 3), the early detection of NSCLC using LDCT followed by appropriate treatments can significantly reduce lung cancer mortality in smokers [1]. LDCT is now recommended for

lung cancer screening in smokers. Yet LDCT has a low specificity for the early detection of lung cancer, presenting a major clinical challenge [1]. The participants enrolled in this project are not representative of the smokers in LDCT screening setting for lung cancer. We will perform a prospective trial to determine if the integromic signature could improve the spesitivity of LDCT for the early deteciton of lung cancer in smokers.

MATERIALS AND METHODS

Patients and clinical specimens

Using a protocol approved by the local Institutional Review Boards Institutional Review Boards, we recruited lung cancer patients and cancer-free smokers according to the inclusion and/or exclusion criteria recommended by U.S. Preventive Services Task Force [26]. Briefly, we enrolled smokers between the ages of 55-80 who had at least a 30 pack-year smoking history and were former smokers (quit within 15 years). Exclusion criteria included pregnancy, current pulmonary infection, surgery within 6 months, radiotherapy within 1 year, and life expectancy of < 1 year. We collected blood in BD Vacutainer spray-coated K2EDTA Tubes (BD, Franklin Lakes, NJ) and prepared plasma using the standard operating protocols developed by The NCI-Early Detection Research Network [27]. The specimens were processed within 2 hours of collection by centrifugation at 1,300 X g for 10 minutes at 4°C. The surgical-pathologic staging of NSCLC was used as the ground truth according to the TNM classification of the International Union Against Cancer (UICC) with the American Joint Committee on Cancer (AJCC) and the International Staging System for Lung Cancer [28, 29]. A total of 106 NSCLC patients and 62 cancer-free smokers were recruited. Among the cancer patients, 27 patients were female and 79 were male. Twenty-four had stage I NSCLC, 19 with stage II, 28 with stage III, 30 with stage IV, and 5 with unknown stage. Fifty-six lung cancer patients were diagnosed with AC, while 40 with SCC. Of the cancer-free smokers, 16 patients were female and 46 were male. There were no significant differences of age, gender and smoking status between the NSCLC patients and cancer-free smokers. The cases and controls were randomly grouped into two cohorts: a development cohort and a validation cohort. The development cohort consisted of 66 lung cancer patients and 33 cancer-free smokers, while the validation cohort comprised 40 lung cancer patients and 29 cancer-free smokers. The demographic and clinical variables of the two cohorts are shown in Table 3.

ddPCR

RNA was extracted from plasma by using Trizol LS reagent (Invitrogen Carlsbad, CA) and RNeasy Mini Kit (Qiagen, Hilden, Germany) [11, 12]. The

qualification and quantification of RNA were assessed by using Biospectrometer (Hutchinson Technology Inc, Hutchinson, MN) and Electrophoresis Bioanalyzer (Agilent Technologies, Foster City, CA). Reverse Transcriptase (RT) was carried out to generate cDNA by using a RT Kit (Applied Biosystems, Foster City, CA) [11, 12]. ddPCR for analysis of expression level of the genes was performed as described in our published works by using a QX200™ Droplet Digital™ PCR System (Bio-Rad, Hercules, CA) [11-15, 30-50]. Briefly, PCR reaction mix containing cDNA was partitioned into aqueous droplets in oil via the QX100 Droplet Generator, and then transferred to a 96-well PCR plate. A two-step thermocycling protocol (95°C ×10min; 40 cycles of [94°C ×30s, 60°C ×60s], 98°C ×10 min) was undertaken in a Bio-Rad C1000 (Bio-Rad, Pleasanton, CA). The PCR plate was then transferred to the QX100 Droplet Reader for automatic reading of samples in all wells. Copy number of each gene per µl PCR reaction was directly determined. Primers and probes of the targeted genes are shown in Supplementary Table 1. We used QuantaSoft 1.7.4 analysis software (Bio-Rad) and Poisson statistics to compute droplet concentrations (copies/µL). Only genes that had at least 10,000 droplets were considered to be robustly detectable by ddPCR in plasma and subsequently underwent further analysis [31]. All assays were done in triplicates, and one no-template control and two interplate controls were carried along in each experiment.

Statistical analysis

To estimate sample size, we set AUC of H0 (the null hypothesis) at 0.5. H1 represented the alternative hypothesis. To have a high reproducibility with adequate precision, we required ≥28 subjects per group. With this sample size, we would have 85% power to detect an AUC of 0.75 at the 2% significance level. Therefore, the sample size in the two cohorts could have enough statistical power. Pearson's correlation analysis was applied to assess relationship between gene expressions and demographic and clinical characteristics of the patients and control individuals. AUCs were used to determine accuracy, sensitivity, and specificity of each gene. We used the highest Youden's J index (sum of sensitivity and specificity—1) to set up corresponding cut-off value [19]. Logistic regression models with constrained parameters as in LASSO were used to eliminate the irrelevant genes, develop composite panels of biomarkers, and optimize a signature with the highest sensitivity and specificity. To compare the signature and our previously developed plasma biomarker panels, we compared their AUCs to determine the sensitivity and specificity as previously described [15].

CONCLUSIONS

Given the heterogeneous nature of NSCLC developed from multifactorial molecular aberrations, we

have for the first time demonstrated that the integration of miRNA, lncRNA, and FUT biomarkers could provide an efficient approach for diagnosis of lung cancer. Nonetheless, a large multi-center clinical project to prospectively validate the full utility of the integromic signature is required.

Author contributions

QL, YL, MZ, and FJ conducted the experiments and participated in study design, coordination, and data interpretation, and preparing the manuscript. All authors read and approved the final manuscript.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

Ethical approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors.

FUNDING

This work was supported in part by NCI R21CA205746, VA Merit Award I01 CX000512, Award from the Geaton and JoAnn DeCesaris Family Foundation, UMD-UMB Research and Innovation Seed Grant, DoD-Idea Development Award, and Maryland Innovation Initiative (MII) Commercialization Program- Phase 1/2 Grant (F.J.).

REFERENCES

1. Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, Gareen IF, Gatsonis C, Marcus PM, Sicks JD. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med*. 2011; 365:395-409.
2. Patz EF Jr, Pinsky P, Gatsonis C, Sicks JD, Kramer BS, Tammemagi MC, Chiles C, Black WC, Aberle DR. Overdiagnosis in low-dose computed tomography screening for lung cancer. *JAMA Intern Med*. 2014; 174:269-274.
3. Aberle DR, Berg CD, Black WC, Church TR, Fagerstrom RM, Galen B, Gareen IF, Gatsonis C, Goldin J, Gohagan JK, Hillman B, Jaffe C, Kramer BS, et al. The National Lung Screening Trial: overview and study design. *Radiology*. 2011; 258:243-253.
4. Hubers AJ, Prinsen CF, Sozzi G, Witte BI, Thunnissen E. Molecular sputum analysis for the diagnosis of lung cancer. *Br J Cancer*. 2013; 109:530-537.
5. Carthew RW, Sontheimer EJ. Origins and Mechanisms of miRNAs and siRNAs. *Cell*. 2009; 136:642-655.
6. Orom UA, Shiekhattar R. Long noncoding RNAs usher in a new era in the biology of enhancers. *Cell*. 2013; 154:1190-1193.
7. Shen J, Jiang F. Applications of MicroRNAs in the Diagnosis and Prognosis of Lung Cancer. *Expert Opin Med Diagn*. 2012; 6:197-207.
8. Mitchell PS, Parkin RK, Kroh EM, Fritz BR, Wyman SK, Pogosova-Agadjanyan EL, Peterson A, Noteboom J, O'Brian KC, Allen A, Lin DW, Urban N, Drescher CW, et al. Circulating microRNAs as stable blood-based markers for cancer detection. *Proc Natl Acad Sci U S A*. 2008; 105:10513-10518.
9. Huarte M. The emerging role of lncRNAs in cancer. *Nat Med*. 2015; 21:1253-1261.
10. Yanaihara N, Caplen N, Bowman E, Seike M, Kumamoto K, Yi M, Stephens RM, Okamoto A, Yokota J, Tanaka T, Calin GA, Liu CG, Croce CM, et al. Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. *Cancer Cell*. 2006; 9:189-198.
11. Shen J, Todd NW, Zhang H, Yu L, Lingxiao X, Mei Y, Guarnera M, Liao J, Chou A, Lu CL, Jiang Z, Fang H, Katz RL, Jiang F. Plasma microRNAs as potential biomarkers for non-small-cell lung cancer. *Lab Invest*. 2011; 91:579-587.
12. Shen J, Liu Z, Todd NW, Zhang H, Liao J, Yu L, Guarnera MA, Li R, Cai L, Zhan M, Jiang F. Diagnosis of lung cancer in individuals with solitary pulmonary nodules by plasma microRNA biomarkers. *BMC Cancer*. 2011; 11:374.
13. Shen J, Liao J, Guarnera MA, Fang H, Cai L, Stass SA, Jiang F. Analysis of MicroRNAs in sputum to improve computed tomography for lung cancer diagnosis. *J Thorac Oncol*. 2014; 9:33-40.
14. Leng Q, Lin Y, Jiang F, Lee CJ, Zhan M, Fang H, Wang Y. A plasma miRNA signature for lung cancer early detection. *Oncotarget*. 2017; 8:111902-111911. <https://doi.org/10.18632/oncotarget.22950>.
15. Lin Y, Leng Q, Jiang Z, Guarnera MA, Zhou Y, Chen X, Wang H, Zhou W, Cai L, Fang H, Li J, Jin H, Wang L, et al. A classifier integrating plasma biomarkers and radiological characteristics for distinguishing malignant from benign pulmonary nodules. *Int J Cancer*. 2017; 141:1240-1248.
16. Pinho SS, Reis CA. Glycosylation in cancer: mechanisms and clinical implications. *Nat Rev Cancer*. 2015; 15:540-555.
17. Kirwan A, Utratna M, O'Dwyer ME, Joshi L, Kilcoyne M. Glycosylation-Based Serum Biomarkers for Cancer Diagnostics and Prognostics. *Biomed Res Int*. 2015; 2015:490531.

18. Kadara H, Wistuba II. Field cancerization in non-small cell lung cancer: implications in disease pathogenesis. *Proc Am Thorac Soc.* 2012; 9:38-42.
19. Youden WJ. Index for rating diagnostic tests. *Cancer.* 1950; 3:32-35.
20. Chen CY, Jan YH, Juan YH, Yang CJ, Huang MS, Yu CJ, Yang PC, Hsiao M, Hsu TL, Wong CH. Fucosyltransferase 8 as a functional regulator of nonsmall cell lung cancer. *Proc Natl Acad Sci U S A.* 2013; 110:630-635.
21. Grosso S, Doyen J, Parks SK, Bertero T, Paye A, Cardinaud B, Gounon P, Lacas-Gervais S, Noel A, Pouyssegur J, Barbry P, Mazure NM, Mari B. MiR-210 promotes a hypoxic phenotype and increases radioresistance in human lung cancer cell lines. *Cell Death Dis.* 2013; 4:e544.
22. Cui Y, Zhang F, Zhu C, Geng L, Tian T, Liu H. Upregulated lncRNA SNHG1 contributes to progression of non-small cell lung cancer through inhibition of miR-101-3p and activation of Wnt/beta-catenin signaling pathway. *Oncotarget.* 2017; 8:17785-17794. <https://doi.org/10.18632/oncotarget.14854>.
23. Noman MZ, Buat S, Romero P, Ketari S, Janji B, Mari B, Mami-Chouaib F, Chouaib S. Hypoxia-inducible miR-210 regulates the susceptibility of tumor cells to lysis by cytotoxic T cells. *Cancer Res.* 2012; 72:4629-4641.
24. Zhang HY, Yang W, Zheng FS, Wang YB, Lu JB. Long non-coding RNA SNHG1 regulates zinc finger E-box binding homeobox 1 expression by interacting with TAp63 and promotes cell metastasis and invasion in Lung squamous cell carcinoma. *Biomed Pharmacother.* 2017; 90:650-658.
25. Carr LL, Jacobson S, Lynch DA, Foreman MG, Flanagan EL, Hersh CP, Sciurba FC, Wilson DO, Sieren JC, Mulhall P, Kim V, Kinsey CM, Bowler RP. Features of COPD as Predictors of Lung Cancer. *Chest.* 2018; 55:286-287.
26. Humphrey LL, Deffebach M, Pappas M, Baumann C, Artis K, Mitchell JP, Zakher B, Fu R, Slatore CG. Screening for lung cancer with low-dose computed tomography: a systematic review to update the US Preventive services task force recommendation. *Ann Intern Med.* 2013; 159:411-420.
27. Tuck MK, Chan DW, Chia D, Godwin AK, Grizzle WE, Krueger KE, Rom W, Sanda M, Sorbara L, Stass S, Wang W, Brenner DE. Standard operating procedures for serum and plasma collection: early detection research network consensus statement standard operating procedure integration working group. *J Proteome Res.* 2009; 8:113-117.
28. Leong SS, Rocha Lima CM, Sherman CA, Green MR. The 1997 International Staging System for non-small cell lung cancer: have all the issues been addressed? *Chest.* 1999; 115: 242-248.
29. Bulzebruck H, Bopp R, Drings P, Bauer E, Krysa S, Probst G, van Kaick G, Muller KM, Vogt-Moykopf I. New aspects in the staging of lung cancer. Prospective validation of the International Union Against Cancer TNM classification. *Cancer.* 1992; 70:1102-1110.
30. Li N, Ma J, Guarnera MA, Fang H, Cai L, Jiang F. Digital PCR quantification of miRNAs in sputum for diagnosis of lung cancer. *J Cancer Res Clin Oncol.* 2014; 140:145-150.
31. Ma J, Li N, Guarnera M, Jiang F. Quantification of Plasma miRNAs by Digital PCR for Cancer Diagnosis. *Biomark Insights.* 2013; 8:127-136.
32. Su J, Leng Q, Lin Y, Ma J, Jiang F, Lee CJ, Fang H. Integrating Circulating Immunological and Sputum Biomarkers for the Early Detection of Lung Cancer. *Biomark Cancer.* 2018; 10: 299-307.
33. Su Y, Fang H, Jiang F. Integrating DNA methylation and microRNA biomarkers in sputum for lung cancer detection. *Clin Epigenetic.* 2016; 8:109.
34. Su Y, Guarnera MA, Fang H, Jiang F. Small non-coding RNA biomarkers in sputum for lung cancer diagnosis. *Clin Epigenetics.* 2016; 15:36.
35. Su J, Anjuman N, Guarnera MA, Zhang H, Stass SA, Jiang F. Analysis of Lung Flute-collected Sputum for Lung Cancer Diagnosis. *Biomark Insights.* 2015; 10:55-61.
36. Su J, Liao J, Gao L, Shen J, Guarnera MA, Zhan M, Fang H, Stass SA, Jiang F. Analysis of small nucleolar RNAs in sputum for lung cancer diagnosis. *Oncotarget.* 2016; 7:5131-5142. <https://doi.org/10.18632/oncotarget.4219>.
37. Xing L, Su J, Guarnera MA, Zhang H, Cai L, Zhou R, Stass SA, Jiang F. Sputum microRNA biomarkers for identifying lung cancer in indeterminate solitary pulmonary nodules. *Clin Cancer Res.* 2015; 21:484-489.
38. Pine PS, Lund SP, Stass SA, Kukuruga D, Jiang F, Sorbara L, Srivastava S, Salit M. Cell-based reference samples designed with specific differences in microRNA biomarkers. *BMC Biotechnol.* 2018; 18:17-26.
39. Anjuman N, Li N, Guarnera M, Stass SA, Jiang F. Evaluation of lung flute in sputum samples for molecular analysis of lung cancer. *Clin Transl Med.* 2013; 2:15.
40. Yu L, Todd NW, Xing L, Xie Y, Zhang H, Liu Z, Fang H, Zhang J, Katz RL, Jiang F. Early detection of lung adenocarcinoma in sputum by a panel of microRNA markers. *Int J Cancer.* 2010; 127:2870-2878.
41. Jiang F, Todd NW, Li R, Zhang H, Fang H, Stass SA. A panel of sputum-based genomic marker for early detection of lung cancer. *Cancer Prev Res (Phila).* 2010; 3:1571-1578.
42. Xing L, Todd NW, Yu L, Fang H, Jiang F. Early detection of squamous cell lung cancer in sputum by a panel of microRNA markers. *Mod Pathol.* 2010; 23:1157-1164.
43. Xie Y, Todd NW, Liu Z, Zhan M, Fang H, Peng H, Alattar M, Deepak J, Stass SA, Jiang F. Altered miRNA expression

- in sputum for diagnosis of non-small cell lung cancer. *Lung Cancer*. 2010; 67:170-176.
44. Jiang F, Todd NW, Qiu Q, Liu Z, Katz RL, Stass SA. Combined genetic analysis of sputum and computed tomography for noninvasive diagnosis of non-small-cell lung cancer. *Lung Cancer*. 2009; 66:58-63.
 45. Li H, Jiang Z, Leng Q, Bai F, Wang J, Ding X, Li Y, Zhang X, Fang H, Yfantis HG, Xing L, Jiang F. A prediction model for distinguishing lung squamous cell carcinoma from adenocarcinoma. *Oncotarget*. 2017; 8:50704-50714. <https://doi.org/10.18632/oncotarget.17038>.
 46. Ma J, Guarnera MA, Zhou W, Fang H, Jiang F. A Prediction Model Based on Biomarkers and Clinical Characteristics for Detection of Lung Cancer in Pulmonary Nodules. *Transl Oncol*. 2017; 10:40-45.
 47. Li R, Todd NW, Qiu Q, Fan T, Zhao RY, Rodgers WH, Fang HB, Katz RL, Stass SA, Jiang F. Genetic deletions in sputum as diagnostic markers for early detection of stage I non-small cell lung cancer. *Clin Cancer Res*. 2007; 13:482-487.
 48. Ma J, Lin Y, Zhan M, Mann DL, Stass SA, Jiang F. Differential miRNA expressions in peripheral blood mononuclear cells for diagnosis of lung cancer. *Lab Invest*. 2015; 95:1197-1206.
 49. Gao L, Ma J, Mannoor K, Guarnera MA, Shetty A, Zhan M, Xing L, Stass SA, Jiang F. Genome-wide small nucleolar RNA expression analysis of lung cancer by next-generation deep sequencing. *Int J Cancer*. 2015; 136:E623-629.
 50. Ma J, Mannoor K, Gao L, Tan A, Guarnera MA, Zhan M, Shetty A, Stass SA, Xing L, Jiang F. Characterization of microRNA transcriptome in lung cancer by next-generation deep sequencing. *Mol Oncol*. 2014; 8:1208-1219.

RESEARCH

Open Access



An epigenetic classifier for early stage lung cancer

Yun Su^{1*}, Hong Bin Fang² and Feng Jiang³

Abstract

Background: Methylated genes detected in sputum are promise biomarkers for lung cancer. Yet the current PCR technologies for quantification of DNA methylation and diagnostic value of the sputum biomarkers are not sufficient to be used for lung cancer early detection. The emerging droplet digital PCR (ddPCR) is a straightforward means for precise, direct, and absolute quantification of nucleic acids. Here, we investigate whether ddPCR can sensitively and robustly quantify DNA methylation in sputum for more precise diagnosis of lung cancer.

Results: First, the analytic performance of methylation-specific ddPCR (ddMSP) and quantitative methylation-specific PCR (qMSP) is determined in methylated and unmethylated DNA samples. Second, 29 genes, previously proposed as potential sputum biomarkers for lung cancer, are analyzed by using ddMSP in a training set of 127 lung cancer patients and 159 controls. ddMSP has higher sensitivity, precision, and reproducibility for quantification of methylation compared with qMSP (all $p < 0.05$). A classifier comprising four sputum methylation biomarkers for lung cancer is developed by using ddMSP, producing 86.6% sensitivity and 90.6% specificity, independent of stage and histology of lung cancer (all $p > 0.05$). The classifier has higher accuracy compared with sputum cytology (88.8 vs. 70.6%, $p < 0.01$). The diagnostic performance is confirmed in a testing set of 89 cases and 107 controls.

Conclusions: ddMSP is a robust tool for reliable quantification of DNA methylation in sputum, and the epigenetic classifier could help diagnose lung cancer at the early stage.

Keywords: ddPCR, DNA methylation, Sputum, Diagnosis, Lung cancer

Background

Lung cancer is the leading cause of cancer death among men and women [1]. More than 85% lung tumors are non-small cell lung cancers (NSCLCs), which consist of adenocarcinoma (AC), squamous cell carcinoma (SCC), and large cell carcinoma (LC). Cigarette smoking is the foremost cause of NSCLC [2]. People who smoke cigarettes are nearly 30 times more likely to get lung cancer or die from lung cancer than people who do not smoke. Even smoking a few cigarettes a day or smoking occasionally increases the risk of lung cancer. Individuals who quit smoking have a lower risk of lung cancer than if they had continued to smoke, but their risk is higher than the risk for people who never smoked. The National Lung Screening Trial (NLST) results show that using low-dose CT (LDCT) for the early detection of

lung cancer in smokers can reduce the mortality by 20% as compared to chest X-rays [1]. Therefore, LDCT is recently recommended to be used for lung cancer early detection among smokers [3, 4]. However, LDCT is associated with over-diagnosis, excessive cost, and radiation exposure, limiting its clinical applications [3–5]. The development of noninvasive approaches that can accurately and cost-effectively diagnose early stage lung cancer among smokers remains clinically important [6].

Lung cancer develops from a field defect characterized by an accumulation of molecular abnormalities resulted from repeated exposure of the airway of the smokers to the tobacco-related carcinogens [7–9]. Regardless of the anatomic location relative to the tumors, the molecular alterations observed in the large bronchial airway might reflect the altered changes existed in lung tumors [9–11]. Sputum is defined as secretions from the airways and contains bronchial epithelial cells exfoliated from the airways or lungs [12]. Therefore, the analysis of exfoliated bronchial

* Correspondence: yunsu326@yahoo.com

¹Department of Surgery, Jiangsu Province Hospital of Nanjing University of Chinese Medicine, 138 Xianlin Road, Nanjing 210023, China
Full list of author information is available at the end of the article



© The Author(s). 2018 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

epitheliums in sputum for the molecular changes may provide a useful tool for noninvasively and cost-effectively diagnosing lung cancer.

DNA methylations of tumor suppressor genes (TSGs) are early molecular events in lung carcinogenesis and thus show great promise as biomarkers for early stage lung cancer [6, 11, 13–34]. Conventional qPCR-based platforms, particularly, methylation-specific PCR (qMSP), have been used for detecting DNA methylation of TSGs in sputum [13]. However, qMSP has some weaknesses, limiting its use in the clinical settings. For example, qMSP is an indirect approach, which requires internal controls for data normalization [35]. Furthermore, qMSP's sensitivity for analyzing low copy number of genes is poor. This is particularly challenging for quantification of DNA methylation in bronchial epitheliums, as the large excess of non-epithelial cells in sputum could obscure detection of the relative scarcity of methylated DNA from the exfoliated bronchial epitheliums. A more sensitive, precise, and reproducibility method for quantification of methylated DNA in sputum would provide a useful means for noninvasive diagnosis of lung cancer.

Droplet digital PCR (ddPCR) is a direct method for quantitatively measuring nucleic acids [36–45], since it depends on limiting partition of the PCR volume, where a positive result of a large number of microreactions indicates the presence of a single molecule in a given reaction. The number of positive reactions, together with Poisson's distribution, can be used to produce a straight and high-confidence measurement of the original target concentration [43]. Furthermore, ddPCR does not require the reliance on rate-based measurements, endogenous controls, and the use of calibration curves. In addition, previous studies including our own research have demonstrated that ddPCR can quantify low-abundance nucleic acids and has higher sensitivity and precision than does conventional PCR [36, 37, 46]. The objective of this study is to investigate whether methylation-specific ddPCR (ddMSP) could sensitively and robustly quantify DNA methylations in sputum and hence develop a biomarker-based classifier for early stage lung cancer.

Methods

Study population

The study protocol was approved by the local Institutional Review Board. The participants in this study were recruited from the hospital at the point of their referral for suspected lung cancer between the ages of 55–80. Written informed consent was obtained from all enrolled subjects. Exclusion criteria included pregnancy, current pulmonary infection, surgery within 6 months, radiotherapy within 1 year, and life expectancy of < 1 year. Clinical diagnosis of lung cancer was made using

histopathologic examinations of specimens obtained by CT-guided transthoracic needle biopsy, transbronchial biopsy, videotape-assisted thoracoscopic surgery, or surgical resection. The surgical pathologic staging was determined according to the TNM classification of the International Union Against Cancer with the 8th American Joint Committee on Cancer and the International Staging System for Lung Cancer. Histopathological classification was determined according to the World Health Organization classification. A total of 482 subjects including 216 lung cancer patients and 266 cancer-free smokers were recruited. The 216 lung cancer patients were diagnosed with NSCLC consisting of 55 stage I cases, 55 stage II cases, 50 stage III cases, and 56 stage IV cases. One hundred and twelve cases were AC, 91 were SCC, and 13 were LC. The 266 cancer-free patients who were smokers and served as control subjects had granulomatous inflammation ($n = 117$), nonspecific inflammatory changes ($n = 105$) or lung infections ($n = 44$). The cancer-free smokers had been followed for at least 2 years, and none had any evidence of cancer. No difference of age, gender, and smoking status was observed in the lung cancer cases vs. controls (All $p > 0.05$). To refine the biomarkers whose changes specific to NSCLC, the cases were matched to the controls on gender, age, race, and smoking status as a nested case-control study. The cases and controls were then randomly split into a training set and a testing set by using a random number generator. The training set consisted of 127 lung cancer patients and 159 cancer-free controls. The testing set comprised 89 lung cancer patients and 107 cancer-free controls. The demographic and clinical characteristics of the two cohorts are presented in Tables 1 and 2.

Sample collection and sputum cytology

Sputum was collected from the participants as described in previous reports [47–54]. Briefly, to reduce the percentage of oral epithelial cells in the sputum, subjects were asked to blow their nose, rinse their mouth, and swallow water to minimize contamination of squamous cells from postnasal drip and saliva. Sputum samples were then coughed in a sterile container and processed within 2 h. To further minimize oral squamous cell contamination, opaque or dense portions that looked different from saliva under the inverted microscope were selected using blunt forceps from expectorate. The samples were processed on ice in 4 volumes of 0.1% dithiothreitol (Sigma-Aldrich, St. Louis, Mo) followed by 4 volumes of phosphate-buffered saline (PBS) (Sigma-Aldrich). The cell suspension was filtered through 45- μ m nylon gauzes (BNSH Thompson, Scarborough, ON, Canada). Absolute cell numbers and cell viability were quantitated by using a hemacytometer with trypan blue. Two cytocentrifuge slides

Table 1 Characteristics of NSCLC patients and cancer-free smokers in a training set

	NSCLC cases (n = 127)	Controls (n = 159)	p value
Age	65.48 (SD 12.32)	65.63 (SD 11.56)	0.32
Sex			0.35
Female	45	56	
Male	82	103	
Smoking pack-years (median)	35.23	33.79	0.34
Stage			
Stage I	33		
Stage II	32		
Stage III	29		
Stage IV	33		
Histological type			
Adenocarcinoma	63		
Squamous cell carcinoma	57		
Large cell carcinoma	7		

Abbreviations: NSCLC non-small cell lung cancer

were prepared from aliquots of cell suspension by using a cytospin machine (Shandon, Pittsburgh, PA) and were then stained with the Papanicolaou staining technique [12]. A sputum sample was considered adequate if lung macrophages or Curschmann spirals were present on the slides [11, 12]. Cytologic diagnosis was performed on the cytospin slides using the classification of Saccomanno et al. [12]. The remaining cells are stored at -80°C until used.

Table 2 Characteristics of NSCLC patients and cancer-free smokers in a testing set

	NSCLC cases (n = 89)	Controls (n = 107)	p value
Age	65.25 (SD 11.28)	65.36 (SD 11.48)	0.30
Sex			0.36
Female	31	37	
Male	58	70	
Smoking pack-years (median)	35.76	33.29	0.32
Stage			
Stage I	22		
Stage II	23		
Stage III	21		
Stage IV	23		
Histological type			
Adenocarcinoma	49		
Squamous cell carcinoma	34		
Large cell carcinoma	6		

Abbreviations: NSCLC non-small cell lung cancer

DNA isolation and bisulfite conversion

We extracted DNA from the specimens using DNeasy kit (Qiagen, Valencia, CA) as previously described [14]. We eluted DNA with 50 μL of elution buffer (10 mmol/L Tris-Cl, pH 8.5) (Sigma-Aldrich Corporation). DNA was quantified by using the Quantifiler Human DNA Quantification kit (Applied Biosystems, Foster City, CA). Bisulfite conversion was carried out on DNA by using the Zymo EZ DNA Methylation Kit (Zymo Research, Irvine, CA) according to the manufacturer's protocol.

Serially diluted methylated/unmethylated DNA specimens

We purchased 100% methylated and 100% unmethylated control human DNA samples (Zymo Research). We isolated DNA from sputum of a healthy nonsmoker whose sputum DNA did not harbor DNA methylation of TSGs, including *RASSF1A*, *3OST2*, and *PRDM14* [14]. To determine limit of quantification (LOQ) of an assay, we diluted methylated DNA into the sputum DNA sample in the following concentrations: 100, 25, 6.25, 1.56, 0.39, 0.1, 0.04, and 0% methylated DNA. To determine limits of detection (LOD) of an assay, we prepared serially diluted samples containing 5000, 2500, 1250, 625, 313, 156, and 0 pg methylated DNA in H_2O .

Quantification of DNA methylation in sputum by ddMSP

We added bisulfite-treated DNA (2 μL) to ddPCR mixture (18 μL) containing 2 $\times$ ddPCR Supermix for probes (no-dUTP), 750 nmol/L of each primer and 250 nmol/L of the corresponding probe in a final volume of 20 μL . Twenty-nine genes were selected for DNA methylation analysis, since the genes were previously reported as potential sputum methylation biomarkers for lung cancer [6, 13–34]. The 29 genes are *3OST2*, *APC*, *CDH1*, *CDO1*, *CXCL*, *CYGB*, *DAL-1*, *DAPK*, *DCR2*, *FAM19A4*, *FHIT*, *GATA*, *H-cadherin*, *HOXA9*, *JPH3*, *KIFLA*, *MAGE*, *p16*, *PAX5*, *PCDH20*, *PHACTR3*, *PRDM14*, *RAR β* , *RASSF1A*, *SOX17*, *SULF2*, *TAC1*, *TCF2L*, and *ZFP42* (Additional file 1: Table S1). Primers and probes of the targeted genes were designed in the studies [6, 13–34]. A thermocycling protocol ($95^{\circ}\text{C} \times 10 \text{ min}$; 40 cycles of [$94^{\circ}\text{C} \times 30 \text{ s}$, $60^{\circ}\text{C} \times 60 \text{ s}$], $98^{\circ}\text{C} \times 10 \text{ min}$) was undertaken in a Bio-Rad C1000 (Bio-Rad, Pleasanton, CA). The PCR plate was transferred to the QX100 Droplet Reader (Bio-Rad) for automatic reading of samples in all wells. We used QuantaSoft 1.7.4 analysis software (Bio-Rad) and Poisson statistics to compute droplet concentrations (copies/ μL ; PCR scale). Only tests that had at least 10,000 droplets were used for the ddMSP analysis [36, 37]. All assays were done in triplicates, and one no-template control and two interplate controls were carried along in each experiment.

Quantification of DNA methylation in sputum by qMSP

qMSP was done as previously described [13, 14]. The cycle threshold (Ct) values for each gene were determined. Ct values above 35 were censored according to previous recommendations [13, 14, 55–58]. To determine methylation level of target genes in a given sample, we normalized Ct values of the target genes in relation to that of *myoblast determination protein one* (*MYOD1*) [13, 32]. The percentage of methylated reference (PMR) was defined as target gene/*MYOD1* ratio of the sample divided by target gene/*MYOD1* ratio of the calibrator DNA (methylated control DNA) and multiplying by 100 [14].

Comparison of tolerance of ddMSP and qMSP to PCR inhibitors

To determine tolerance of ddMSP and qMSP to inhibitory substances of PCR, we directly introduced inhibitors, sodium dodecyl sulfate (SDS), and heparin (Sigma-Aldrich Corporation), into the PCR reactions [59, 60]. Differences in the resulting inhibition curves and the half-maximal inhibitory concentrations (IC₅₀) were assessed and compared as described previously [59, 60].

Statistical analysis

We used *t* test to determine significant differences of values of each gene between cases and controls. We used log transformation of the molecular results and applied Pearson's correlation analysis to assess relationship between DNA methylation and demographic characteristics of subjects. We calculated coefficient of variations (CV) to determine the variation between different measurements. We performed the linear regression between different measurements of the assays and the amount of input DNA. We used the receiver-operator characteristic (ROC) curve and area under the curve (AUC) to determine accuracy, sensitivity, and specificity of each gene or the tests. We employed logistic regression models with constrained parameters as in least absolute shrinkage and selection operator (LASSO) based on ROC criterion to eliminate the irrelevant genes and optimize a composite biomarker panel (classifier). The optimal panel of biomarkers was blindly applied to the testing data set to confirm the diagnostic value by comparing the AUC with the goodness-of-fit statistics [61].

Results

ddMSP has higher sensitivity, precision, and reproducibility for quantification of DNA methylation compared with qMSP

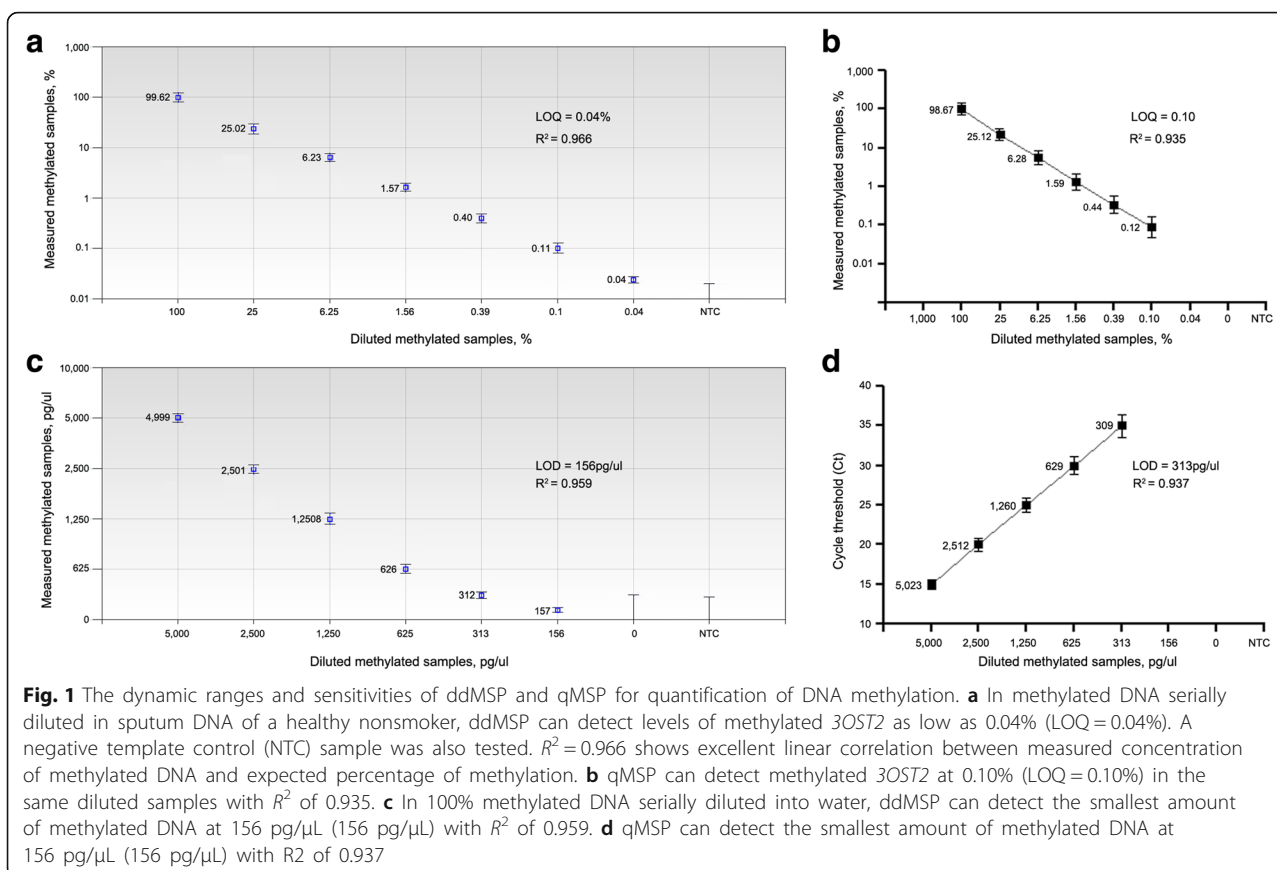
In methylated DNA serially diluted in sputum DNA of a healthy nonsmoker, ddMSP generated at least 10,000 droplets passing through a fluorescence detector. The results suggested that the specimens were successfully

“read” by ddPCR. ddMSP detected methylated genes (*RASSF1A*, *3OST2*, and *PRDM14*) at a concentration of 0.04% (LOQ = 0.04%) ($R^2 = 0.966$) (Fig. 1a), whereas qMSP detected the methylation at a concentration of 0.10% (LOQ = 0.10%) ($R^2 = 0.935$) ($p = 0.008$) (Fig. 1b). There was excellent linearity between the methylated DNA input and values measured by both qMSP and ddMSP (all $R^2 \geq 0.93$). Furthermore, the dispersion of values of the four analyses of the specimen was lower with ddMSP than with qMSP. The repeated measurements by ddMSP had a lower CV value compared with those determined by qMSP ($p = 0.03$) (Additional file 1: Table S2). Therefore, ddMSP had a higher precision for quantification of methylation compared to qMSP ($p = 0.03$) (Additional file 1: Table S2).

To determine the absolute LOD of the two platforms, 100% methylated DNA serially diluted into water and then tested by ddMSP and qMSP. The smallest amount of methylated DNA that can be reliably measured by ddMSP was 156 pg/μL (Fig. 1c), suggesting that ddMSP had a LOD of 156 pg/μL. qMSP produced more than 35 Ct values for the samples that had less than 313 pg methylated DNA per microliter, yielding a LOD of 313 pg/μL (Fig. 1d). Therefore, ddMSP had higher sensitivity as demonstrated by lower LOQ and LOD than did qMSP in the serial dilutions of DNA control samples (all $p < 0.001$).

To determine reproducibility of ddMSP and qMSP, the diluted samples were independently analyzed. The CVs of repeated measures by ddMSP on different days were more than twofold lower compared with those determined by qMSP (Additional file 1: Table S3). Furthermore, the CVs of repeated measures by different research staff using ddMSP were at least twofold lower than did those generated by qMSP (Additional file 1: Table S4). Therefore, ddMSP had a higher reproducibility than did qMSP for quantification of DNA methylation.

To evaluate analytic performance of ddMSP and qMSP in clinical sputum samples, sputum of 20 lung cancer patients and 20 cancer-free controls was tested for *RASSF1A*, whose aberrant methylation level was shown to be elevated in sputum of lung cancer patients [13, 14, 55–58]. Each well of the samples contained at least 10,000 droplets (Fig. 2a). Therefore, ddMSP analysis of DNA methylation could successfully be performed in clinical sputum specimens. *RASSF1A* analyzed by both the techniques displayed a high methylation level in lung cancer patients vs. controls (all $p < 0.05$). In the ddMSP assay, a specimen with $\geq$ one copy of DNA methylation of *RASSF1A*-per microliter was considered to be positive. When the criteria was used, of 20 sputum specimens of lung cancer patients, 11 (55%) had positive methylation of the gene detected by ddMSP. In the qMSP assay, a PMR $\geq 1\%$ was classified as positive for *RASSF1A* in a given sample [62].



When the criteria was used, 9 (45%) were positive for *RASSF1A* by qMSP. The same 4 sputum specimens of control subjects had positive methylation of the gene detected by both ddMSP and qMSP. Therefore, ddMSP analysis of DNA methylation of *RASSF1A* in sputum had a higher sensitivity (55%) than did qMSP (45%) ($p = 0.01$) for distinguishing lung cancer patients from control subjects, while maintaining the same specificity (80%) (Additional file 1: Figure S1). Furthermore, the CVs of repeated measures by ddMSP on different days by different researchers were approximately twofold lower compared with those generated by qMSP. Altogether, in clinical sputum specimens, ddMSP also exhibited higher sensitivity, accuracy, and reproducibility than did qMSP for quantification of DNA methylation.

To compare the tolerance of ddMSP and qMSP to PCR inhibitors, we added SDS and heparin directly into the PCR reactions and then calculated log IC50 values from the resulting inhibition curves. We found greater than a half log increase in IC50 of ddMSP over qMSP for both SDS and heparin (all $p < 0.05$), implying that ddMSP tolerated the presence of the inhibitors better than qMSP.

Diagnostic performance of ddMSP quantified-sputum methylation biomarkers for lung cancer

We first evaluated DNA methylation of 29 genes in the training cohort of 127 NSCLC patients and 159 controls. All the 29 genes displayed a higher level of methylation in patients vs. controls (all $p < 0.05$). ROC curve and AUC analysis showed that the genes had 29–88% sensitivities and 26–92% specificities in differentiating lung cancer patients from healthy controls (Additional file 1: Table S1). Since methylation levels of genes did not follow a normal distribution, we used the log transformation of ddPCR results. We then applied multivariate logistic regression models with stepwise regression based on ROC curve to develop a prediction classifier. Four genes (*HOXA9*, *RASSF1A*, *SOX17*, and *TAC1*) were identified as the best biomarkers (all $p < 0.001$) and incorporated into a logistic classifier: Probability of lung cancer = $e^x / (1 + e^x)$, where $x = 1.69 + 1.48 \times \log(\text{HOXA9}) - 1.25 \times \log(\text{RASSF1A}) + 0.27 \times \log(\text{SOX17}) + 0.16 \times \log(\text{TAC1})$. The logistic classifier produced 0.92 AUC for lung cancer detection (Fig. 2b). Furthermore, Pearson correlation among methylation levels of the four genes was low ($p > 0.05$), implying that their diagnostic values were complementary to each other. Using Youden's index, we set up

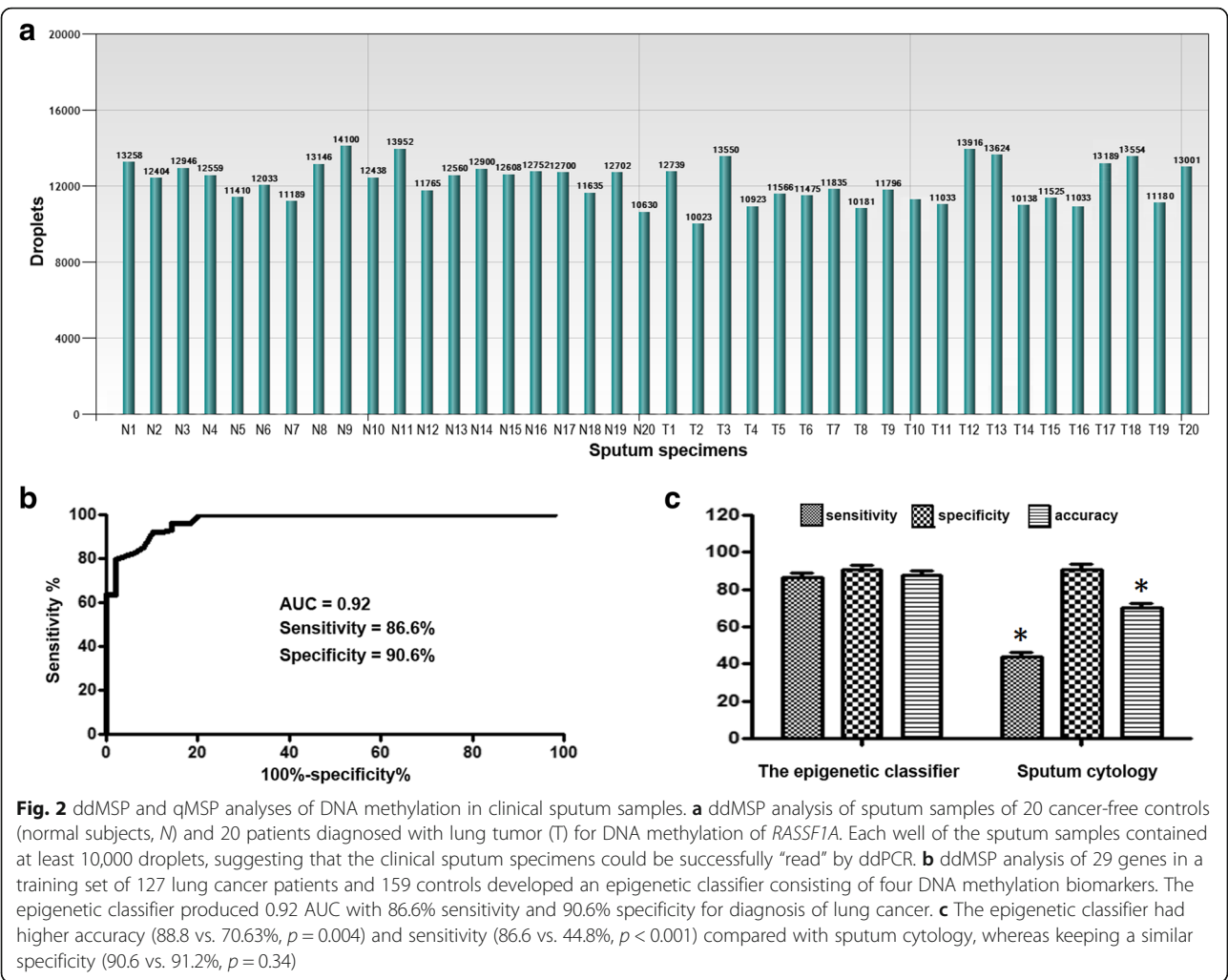


Fig. 2 ddMSP and qMSP analyses of DNA methylation in clinical sputum samples. **a** ddMSP analysis of sputum samples of 20 cancer-free controls (normal subjects, N) and 20 patients diagnosed with lung tumor (T) for DNA methylation of *RASSF1A*. Each well of the sputum samples contained at least 10,000 droplets, suggesting that the clinical sputum specimens could be successfully “read” by ddPCR. **b** ddMSP analysis of 29 genes in a training set of 127 lung cancer patients and 159 controls developed an epigenetic classifier consisting of four DNA methylation biomarkers. The epigenetic classifier produced 0.92 AUC with 86.6% sensitivity and 90.6% specificity for diagnosis of lung cancer. **c** The epigenetic classifier had higher accuracy (88.8 vs. 70.63%, $p = 0.004$) and sensitivity (86.6 vs. 44.8%, $p < 0.001$) compared with sputum cytology, whereas keeping a similar specificity (90.6 vs. 91.2%, $p = 0.34$)

optimal cutoff at 1.28 for the prediction classifier. Subsequently, combined use of the four genes by simply calculating the equation produced 86.6% sensitivity and 90.6% specificity. In addition, including other genes in the prediction classifier did not improve the accuracy for lung cancer diagnosis. The prevalence of the DNA methylation of the four genes was related with pack-years of smoking ($p = 0.03$). Since the cases and controls were matched 1:1 by age, gender, and smoking status as a nested case-control study, we adjusted the parameters during model building. The logistic classifier did not show special association with stage and histological type of lung cancer, and patients’ age, gender, and smoking status (all $p > 0.05$).

Moreover, the logistic classifier had higher accuracy (88.8 vs. 70.63%, $p = 0.004$) and sensitivity (86.6 vs. 44.8%, $p < 0.001$) than did sputum cytology, while maintaining a similar specificity (90.6 vs. 91.2%, $p = 0.34$) (Fig. 2c). The integrated use of the biomarkers and sputum cytology did not significantly increase the diagnostic value.

Validating the panel of ddMSP-quantified methylation biomarkers in a testing cohort

In the testing cohort, the panel of the four genes had 85.4% sensitivity and 91.6% specificity in differentiating lung cancer patients from controls (Table 3). In line with findings in the training set, the logistic classifier was not

Table 3 The diagnostic performance of the epigenetic classifier and sputum cytology in a testing set

	Accuracy (%) (95% CI)*	Sensitivity (%) (95% CI)*	Specificity (%) (95% CI)
The epigenetic classifier	88.78 (83.50 to 92.83)	85.39 (76.32 to 91.99)	91.59 (84.63 to 96.08)
Sputum cytology	71.43 (64.56 to 77.64)	46.07 (35.44 to 56.96)	92.52 (85.80 to 96.72)

Abbreviations: NSCLC non-small cell lung cancer, CI confidence interval
* $p < 0.05$

associated with patient's age, gender, and smoking status, as well as histological type and stage of NSCLC (all $p > 0.05$). Moreover, the logistic classifier yielded higher accuracy (88.8 vs. 71.4%, $p = 0.002$) and sensitivity (85.4 vs. 46.1%, $p < 0.001$), while keeping a similar specificity (91.6 vs. 92.5%, $p = 0.45$), as compared with sputum cytology (Table 3). Taken together, the validation data confirmed the potential of the ddMSP-quantified sputum biomarkers as a sensitive classifier for the early detection of lung cancer.

Discussion

This current study presents the earliest assessment of ddPCR, an emerging technique, for quantitative detection of DNA methylation in sputum. We find that ddMSP can absolutely and robustly quantify DNA methylation in sputum without requiring external references. Therefore, determination of DNA methylation in sputum by ddMSP is highly efficient, and data handling is forthright. Furthermore, our head-to-head comparison of ddMSP and qMSP reveals that ddMSP displays higher precision and reproducibility in measuring copy number of DNA methylation in both control DNA samples and clinical sputum specimens. The sensitivity of conventional qMSP for analyzing low-abundance methylated DNA is poor. This is particularly challenging for the determination of DNA methylation in bronchial epitheliums, since the large excess of normal cells in sputum could obscure detection of the relative scarcity of methylated DNA. We find that ddMSP has a higher sensitivity to quantify cancer-specific methylation and thus could overcome the obstacle of qMSP. In addition, the total time required for ddMSP is about twofold shorter than did qMSP and might be further reduced when an automated system is used [63, 64]. Moreover, ddMSP is not expensive and tolerates the PCR inhibitors better compared with conventional qMSP. Altogether, ddMSP is a straightforward and robust approach for accurate quantification of DNA methylation in sputum.

Importantly, using ddMSP, we develop and validate a DNA methylation-based classifier that has higher accuracy and sensitivity compared with sputum cytology, the clinical gold standard. Furthermore, ddMSP analysis of the four genes by simply calculating the equation would be a convenient tool to be used in the clinics. In addition, the diagnostic performance of the logistic classifier is independent of stage and histological type of the NSCLC, as well as age, gender, and smoking status of subjects. Therefore, the classifier has an important characteristic if it is employed for more precisely and easily identifying early stage lung cancer among smokers.

However, some limitations do exist in this present study: (i) we evaluate ddMSP for quantification of DNA methylation using a retrospective cases and controls,

which may produce selection bias and overfitting. Furthermore, the cases and controls are hospital-based patients, and not representative of the smokers in a screening setting for lung cancer early detection. We will perform a large trial to prospectively validate if the logistic classifier could help identify lung cancer at the early stage in a screening setting among smokers. (ii) the overall sensitivity and specificity of the DNA methylation-based classifier are 86.6 and 90.6%, which are not high enough for clinical diagnosis of NSCLC. The integration of the methylation biomarkers with other classes of biomarkers, such as microRNAs [36, 47–52, 65–67] or DNA mutations [33], is one path to improve the early detection of lung cancer [14]. (iii) since there is no sample left from the patients of the training and cohorts, we are not able to test the classification performance of qMSP in all the samples of the training and testing cohorts. We are consenting new cases and controls and collecting specimens and will then compare performance of qMSP and ddMSP in the same samples. (iv) like qMSP, ddMSP has the same limitations as do other PCR-based platforms. For instance, large experiments can become quite labor intensive to perform, when multiple genes are targeted. In this present study, we use 96-well PCR plates. Given the four genes to be analyzed with duplication, we only test 12 clinical samples at one time. In the future, we will use 484-well PCR plates, in which, we could simultaneously test 60 samples.

Conclusion

ddMSP could be a sensitive and robust tool for reliable quantitation of sputum methylation biomarkers. The ddMSP-quantified methylation classifier may provide a potential diagnostic test for the early detection of lung cancer and thus reduce the deaths and costs associated with the disease. Nevertheless, the continued development of this new technology, and further exploring its value for routine use for the early detection of lung cancer would be required.

Additional file

Additional file 1: Table S1. Twenty-nine genes displayed a higher level of methylation in sputum of lung cancer patients vs. controls. **Table S2.** CVs of the four analyses by ddMSP and qMSP on the same specimens. **Table S3.** CVs of repeated measures by ddMSP and qMSP on different times. **Table S4.** CVs of repeated measures by different researches using ddMSP and qMSP. **Figure S1.** Sensitivities and specificities of measuring DNA methylation of *RASSF1A* by ddMSP and qMSP for diagnosis of lung cancer in sputum samples of 20 cancer-free controls (normal subjects, N) and 20 patients diagnosed with lung tumor (T). ddMSP analysis of DNA methylation of *RASSF1A* had a higher sensitivity (55%) than did qMSP (45%) ($p = 0.01$) for distinguishing lung cancer cases from controls, while maintaining the same specificity (80%). *, $p = 0.01$. (PDF 235 kb)

Acknowledgements

This study was supported in part by a grant for cancer research from Jiangsu Province Hospital of Nanjing University of Chinese Medicine (Y. S.)

Authors' contributions

YS, HF, and FJ conducted the experiments and participated in study design, coordination, and data interpretation, and preparing the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

This study was supported in part by a grant for cancer research from the Jiangsu Province Hospital of Nanjing University of Chinese Medicine. The study was approved by the Nanjing University of Chinese Medicine Institutional Review Board and adhered to the Declaration of Helsinki, and all patients provided written informed consent.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Department of Surgery, Jiangsu Province Hospital of Nanjing University of Chinese Medicine, 138 Xianlin Road, Nanjing 210023, China. ²Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University Medical Center, 4000 Reservoir Road, N.W, Washington D.C. 20057, USA. ³Department of Pathology, University of Maryland School of Medicine, Baltimore, MD, USA.

Received: 24 February 2018 Accepted: 14 May 2018

Published online: 22 May 2018

References

- Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med*. 2011;365:395–409.
- Blumer G. Cigarette smoking and cancer of the lung. *Ill Med J*. 1951;100:98–9.
- Patz EF Jr, Pinsky P, Gatsonis C, Sicks JD, Kramer BS, Tammemagi MC. Overdiagnosis in low-dose computed tomography screening for lung cancer. *JAMA Intern Med*. 2014;174:269–74.
- Aberle DR, Berg CD, Black WC, Church TR, Fagerstrom RM, Galen B. The national lung screening trial: overview and study design. *Radiology*. 2011; 258:243–53.
- Sant M, Allemani C, Santaquilani M, Knijn A, Marchesi F, Capocaccia R. EURO-CARE-4. Survival of cancer patients diagnosed in 1995–1999. Results and commentary. *Eur J Cancer*. 2009;45:931–91.
- Hubers AJ, Prinsen CF, Sozzi G, Witte BJ, Thunnissen E. Molecular sputum analysis for the diagnosis of lung cancer. *Br J Cancer*. 2013;109:530–7.
- Belinsky SA. Gene-promoter hypermethylation as a biomarker in lung cancer. *Nat Rev Cancer*. 2004;4:707–17.
- Grossman DC, Curry SJ, Owens DK, Barry MJ, Davidson KW, Doubeni CA. Screening for adolescent idiopathic scoliosis: US preventive services task force recommendation statement. *JAMA*. 2018;319:165–72.
- Brody JS, Spira A. State of the art. Chronic obstructive pulmonary disease, inflammation, and lung cancer. *Proc Am Thorac Soc*. 2006;3:535–7.
- Kadara H, Wistuba II. Field cancerization in non-small cell lung cancer: implications in disease pathogenesis. *Proc Am Thorac Soc*. 2012;9:38–42.
- Belinsky SA, Liechty KC, Gentry FD, Wolf HJ, Rogers J, Vu K. Promoter hypermethylation of multiple genes in sputum precedes lung cancer incidence in a high-risk cohort. *Cancer Res*. 2006;66:3338–44.
- Saccomanno G, Saunders RP, Archer VE, Auerbach O, Kuschner M, Beckler PA. Cancer of the lung: the cytology of sputum prior to the development of carcinoma. *Acta Cytol*. 1965;9:413–23.
- Hubers AJ, Heideman DA, Burgers SA, Herder GJ, Sterk PJ, Rhodius RJ. DNA hypermethylation analysis in sputum for the diagnosis of lung cancer: training validation set approach. *Br J Cancer*. 2015;112:1105–13.
- Su Y, Fang H, Jiang F. Integrating DNA methylation and microRNA biomarkers in sputum for lung cancer detection. *Clin Epigenetics*. 2016;8:109.
- Liu D, Peng H, Sun Q, Zhao Z, Yu X, Ge S. The indirect efficacy comparison of DNA methylation in sputum for early screening and auxiliary detection of lung cancer: a meta-analysis. *Int J Environ Res Public Health*. 2017; 23:14–9.
- Hsu HS, Chen TP, Wen CK, Hung CH, Chen CY, Chen JT. Multiple genetic and epigenetic biomarkers for lung cancer detection in cytologically negative sputum and a nested case-control study for risk assessment. *J Pathol*. 2007;213:412–9.
- Guzman L, Depix MS, Salinas AM, Roldan R, Aguayo F, Silva A. Analysis of aberrant methylation on promoter sequences of tumor suppressor genes and total DNA in sputum samples: a promising tool for early detection of COPD and lung cancer in smokers. *Diagn Pathol*. 2012;7:87.
- Herman JG, Graff JR, Myohanen S, Nelkin BD, Baylin SB. Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. *Proc Natl Acad Sci U S A*. 1996;93:9821–6.
- Wang X, Cao A, Peng M, Hu C, Liu D, Gu T. The value of chest CT scan and tumor markers detection in sputum for early diagnosis of peripheral lung cancer. *Zhongguo Fei Ai Za Zhi*. 2004;7:58–63.
- Leng S, Wu G, Klinge DM, Thomas CL, Casas E, Picchi MA. Gene methylation biomarkers in sputum as a classifier for lung cancer risk. *Oncotarget*. 2017;8: 63978–85.
- Belinsky SA, Leng S, Wu G, Thomas CL, Picchi MA, Lee SJ. Gene methylation biomarkers in sputum and plasma as predictors for lung cancer recurrence. *Cancer Prev Res (Phila)*. 2017;10:635–40.
- Miyake M, Gomes Giacoia E, Aguilar Palacios D, Rosser CJ. Lung cancer risk assessment for smokers: gene promoter methylation signature in sputum. *Biomark Med*. 2012;6:512.
- Leng S, Do K, Yingling CM, Picchi MA, Wolf HJ, Kennedy TC. Defining a gene promoter methylation signature in sputum for lung cancer risk assessment. *Clin Cancer Res*. 2012;18:3387–95.
- Hwang SH, Kim KU, Kim JE, Kim HH, Lee MK, Lee CH. Detection of HOXA9 gene methylation in tumor tissues and induced sputum samples from primary lung cancer patients. *Clin Chem Lab Med*. 2011;49:699–704.
- Liu Y, Lan Q, Shen M, Jin J, Mumford J, Ren D. Aberrant gene promoter methylation in sputum from individuals exposed to smoky coal emissions. *Anticancer Res*. 2008;28:2061–6.
- Belinsky SA, Grimes MJ, Casas E, Stidley CA, Franklin WA, Bocklage TJ. Predicting gene promoter methylation in non-small-cell lung cancer by evaluating sputum and serum. *Br J Cancer*. 2007;96:1278–83.
- Su SB, Yang LJ, Zhang W, Jin YL, Nie JH, Tong J. p16 and MGMT gene methylation in sputum cells of uranium workers. *Zhonghua Lao Dong Wei Sheng Zhi Ye Bing Za Zhi*. 2006;24:92–5.
- Belinsky SA, Klinge DM, Dekker JD, Smith MW, Bocklage TJ, Gilliland FD. Gene promoter methylation in plasma and sputum increases with lung cancer risk. *Clin Cancer Res*. 2005;11:6505–11.
- Olaussen KA, Soria JC, Park YW, Kim HJ, Kim SH, Ro JY. Assessing abnormal gene promoter methylation in paraffin-embedded sputum from patients with NSCLC. *Eur J Cancer*. 2005;41:2112–9.
- Hubers AJ, Heideman DA, Duin S, Witte BJ, de Koning HJ, Groen HJ. DNA hypermethylation analysis in sputum of asymptomatic subjects at risk for lung cancer participating in the NELSON trial: argument for maximum screening interval of 2 years. *J Clin Pathol*. 2017;7:250–4.
- Hubers AJ, Brinkman P, Boksem RJ, Rhodius RJ, Witte BJ, Zwinderman AH. Combined sputum hypermethylation and eNose analysis for lung cancer diagnosis. *J Clin Pathol*. 2014;67:707–11.
- Hubers AJ, van der Drift MA, Prinsen CF, Witte BJ, Wang Y, Shivapurkar N. Methylation analysis in spontaneous sputum for lung cancer diagnosis. *Lung Cancer*. 2014;84:127–33.
- Hubers AJ, Heideman DA, Yatabe Y, Wood MD, Tull J, Taron M, et al. EGFR mutation analysis in sputum of lung cancer patients: a multitechnique study. *Lung Cancer*. 2013;82(1):38–43.
- Hubers AJ, Heideman DA, Herder GJ, Burgers SA, Sterk PJ, Kunst PW. Prolonged sampling of spontaneous sputum improves sensitivity of hypermethylation analysis for lung cancer. *J Clin Pathol*. 2012;65:541–5.
- Hindson CM, Chevillet JR, Briggs HA, Gallichotte EN, Ruf IK, Hindson BJ. Absolute quantification by droplet digital PCR versus analog real-time PCR. *Nat Methods*. 2013;10:1003–5.
- Li N, Ma J, Guarnera MA, Fang H, Cai L, Jiang F. Digital PCR quantification of miRNAs in sputum for diagnosis of lung cancer. *J Cancer Res Clin Oncol*. 2014;140:145–50.
- Ma J, Li N, Guarnera M, Jiang F. Quantification of plasma miRNAs by digital PCR for cancer diagnosis. *Biomark Insights*. 2013;8:127–36.
- Bhat S, Herrmann J, Armishaw P, Corbisier P, Emslie KR. Single molecule detection in nanofluidic digital array enables accurate measurement of DNA copy number. *Anal Bioanal Chem*. 2009;394:457–67.

39. Kiss MM, Ortoleva-Donnelly L, Beer NR, Warner J, Bailey CG, Colston BW. High-throughput quantitative polymerase chain reaction in picoliter droplets. *Anal Chem*. 2008;80:8975–81.
40. Kreutz JE, Munson T, Huynh T, Shen F, Du W, Ismagilov RF. Theoretical design and analysis of multivolume digital assays with wide dynamic range validated experimentally with microfluidic digital PCR. *Anal Chem*. 2011;83: 8158–68.
41. Pinheiro LB, Coleman VA, Hindson CM, Herrmann J, Hindson BJ, Bhat S. Evaluation of a droplet digital polymerase chain reaction format for DNA copy number quantification. *Anal Chem*. 2012;84:1003–11.
42. Pohl G, Shih IM. Principle and applications of digital PCR. *Expert Rev Mol Diagn*. 2004;4:41–7.
43. Vogelstein B, Kinzler KW. Digital PCR. *Proc Natl Acad Sci U S A*. 1999;96: 9236–41.
44. Hayden RT, Gu Z, Ingersoll J, Abdul-Ali D, Shi L, Pounds S, et al. Comparison of droplet digital PCR to real-time PCR for quantitative detection of cytomegalovirus. *J Clin Microbiol*. 2013;51(2):540–6.
45. Day E, Dear PH, McCaughan F. Digital PCR strategies in the development and analysis of molecular biomarkers for personalized medicine. *Methods*. 2013;59:101–7.
46. Li H, Jiang Z, Leng Q, Bai F, Wang J, Ding X. A prediction model for distinguishing lung squamous cell carcinoma from adenocarcinoma. *Oncotarget*. 2017;8:50704–14.
47. Yu L, Todd NW, Xing L, Xie Y, Zhang H, Liu Z. Early detection of lung adenocarcinoma in sputum by a panel of microRNA markers. *Int J Cancer*. 2010;127:2870–8.
48. Xing L, Todd NW, Yu L, Fang H, Jiang F. Early detection of squamous cell lung cancer in sputum by a panel of microRNA markers. *Mod Pathol*. 2010; 23:1157–64.
49. Xie Y, Todd NW, Liu Z, Zhan M, Fang H, Peng H. Altered miRNA expression in sputum for diagnosis of non-small cell lung cancer. *Lung Cancer*. 2010; 67:170–6.
50. Anjuman N, Li N, Guarnera M, Stass SA, Jiang F. Evaluation of lung flute in sputum samples for molecular analysis of lung cancer. *Clin Transl Med*. 2013;2:15.
51. Jiang F, Todd NW, Li R, Zhang H, Fang H, Stass SA. A panel of sputum-based genomic marker for early detection of lung cancer. *Cancer Prev Res (Phila)*. 2010;3:1571–8.
52. Jiang F, Todd NW, Qiu Q, Liu Z, Katz RL, Stass SA. Combined genetic analysis of sputum and computed tomography for noninvasive diagnosis of non-small-cell lung cancer. *Lung Cancer*. 2009;66:58–63.
53. Qiu Q, Todd NW, Li R, Peng H, Liu Z, Yfantis HG. Magnetic enrichment of bronchial epithelial cells from sputum for lung cancer diagnosis. *Cancer*. 2008;114:275–83.
54. Li R, Todd NW, Qiu Q, Fan T, Zhao RY, Rodgers WH. Genetic deletions in sputum as diagnostic markers for early detection of stage I non-small cell lung cancer. *Clin Cancer Res*. 2007;13:482–7.
55. Li W, Deng J, Jiang P, Zeng X, Hu S, Tang J. Methylation of the RASSF1A and RARB genes as a candidate biomarker for lung cancer. *Exp Ther Med*. 2012;3:1067–71.
56. Lee SM, Lee WK, Kim DS, Park JY. Quantitative promoter hypermethylation analysis of RASSF1A in lung cancer: comparison with methylation-specific PCR technique and clinical significance. *Mol Med Rep*. 2012;5:239–44.
57. Fischer JR, Ohnmacht U, Rieger N, Zemaitis M, Stoffregen C, Manegold C. Prognostic significance of RASSF1A promoter methylation on survival of non-small cell lung cancer patients treated with gemcitabine. *Lung Cancer*. 2007;56:115–23.
58. Chen H, Suzuki M, Nakamura Y, Ohira M, Ando S, Iida T. Aberrant methylation of RASGRF2 and RASSF1A in human non-small cell lung cancer. *Oncol Rep*. 2006;15:1281–5.
59. Dingle TC, Sedlak RH, Cook L, Jerome KR. Tolerance of droplet-digital PCR vs real-time quantitative PCR to inhibitory substances. *Clin Chem*. 2013;59: 1670–2.
60. Wilson IG. Inhibition and facilitation of nucleic acid amplification. *Appl Environ Microbiol*. 1997;63:3741–51.
61. Lemeshow S, Hosmer DW Jr. A review of goodness of fit statistics for use in the development of logistic regression models. *Am J Epidemiol*. 1982;115: 92–106.
62. Grote HJ, Schmiemann V, Gedert H, Bocking A, Kappes R, Gabbert HE. Methylation of RAS association domain family protein 1A as a biomarker of lung cancer. *Cancer*. 2006;108:129–34.
63. Kurimoto K, Hayashi M, Guerrero-Preston R, Koike M, Kanda M, Hirabayashi S. PAX5 gene as a novel methylation marker that predicts both clinical outcome and cisplatin sensitivity in esophageal squamous cell carcinoma. *Epigenetics*. 2017;12: 865–74.
64. Campomenosi P, Gini E, Noonan DM, Poli A, D'Antona P, Rotolo N. A comparison between quantitative PCR and droplet digital PCR technologies for circulating microRNA quantification in human lung cancer. *BMC Biotechnol*. 2016;16:60.
65. Shen J, Liao J, Guarnera MA, Fang H, Cai L, Stass SA. Analysis of MicroRNAs in sputum to improve computed tomography for lung cancer diagnosis. *J Thorac Oncol*. 2014;9:33–40.
66. Su J, Liao J, Gao L, Shen J, Guarnera MA, Zhan M. Analysis of small nucleolar RNAs in sputum for lung cancer diagnosis. *Oncotarget*. 2016;7:5131–42.
67. Xing L, Su J, Guarnera MA, Zhang H, Cai L, Zhou R. Sputum microRNA biomarkers for identifying lung cancer in indeterminate solitary pulmonary nodules. *Clin Cancer Res*. 2015;21:484–9.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



A Direct Plasma miRNA Assay for Early Detection and Histological Classification of Lung Cancer



Qixin Leng^{*}, Yue Wang[†] and Feng Jiang^{*}

^{*}Department of Pathology, University of Maryland School of Medicine, Baltimore, MD 21201, USA; [†]Customer Value Partners, Towson, MD 21286, USA

Abstract

Cell-free microRNAs in plasma provide circulating biomarkers for lung cancer. Most techniques for analysis of miRNAs require a large plasma volume to purify a sufficient RNA yield followed by complicated downstream processing. Small differences in the multiple procedures often cause large analytical variations and poor diagnostic values of the plasma biomarkers. Here we investigate whether directly quantifying plasma miRNAs without RNA purification could diagnose lung cancer. FirePlex assay was directly applied to 20 μ l plasma of 56 lung cancer patients and 28 cancer free controls for quantifying 11 lung tumor-associated miRNAs. FirePlex assay is easier, less expensive and time-consuming for quantification of plasma miRNAs compared with conventional reverse transcription PCR with an equivalent analytic performance. From the lung tumor-associated miRNAs, a prediction model based on two miRNAs (miRs-205-5p and -210-3p) was developed, producing 78.6% sensitivity and 89.3% specificity for identifying lung cancer. The diagnostic value was independent of stage of lung tumor, and patients' age and sex (all $P > 0.05$). Furthermore, based on the same two miRNAs, additional prediction models were developed with 75.0% sensitivity and 89.3% specificity for diagnosis of lung squamous cell carcinoma, and 82.2% sensitivity and 89.3% specificity for lung adenocarcinoma. The direct plasma assay can improve the efficacy of miRNA assessment in a small plasma volume by reducing multiple procedure-associated analytical variables. The developed plasma miRNA biomarkers might be useful for the early detection and histological classification of lung cancer.

Translational Oncology (2018) 11, 883–7

Introduction

Non-small-cell lung cancer (NSCLC) accounts for approximately 85% of lung cancer that is the number one cancer killer [1]. NSCLC mainly consists of two major histological subtypes: squamous cell carcinoma (SCC) and adenocarcinoma (AC) [2]. The 5-year survival rate for stage IV NSCLC is only 10%, whereas nearly 85% for stage IA NSCLC [1]. These statistics provide the primary rationale to improve the early detection of NSCLC. Furthermore, histological subtype is predictive of a differential response rate, overall survival, or toxicity profile from certain therapies. Precisely classifying subtypes of lung cancer is also clinically important for the personalized treatment of the malignancy [3–5]. MicroRNAs (miRNAs) are small molecules that have important functions in diverse biological processes, including cell proliferation, differentiation, and apoptosis [6–8]. Furthermore, dysfunction of miRNAs contributes to the development and progression of human malignancies, including lung cancer [9]. In addition, miRNA expression profiles represent molecular

signatures for the classification, diagnosis, and progression of tumor, and thus could be developed as cancer biomarkers [8,10]. Blood is one of the most easily and noninvasively accessible body fluids. The use of blood plasma for assessment of circulating miRNAs originating from primary tumors would provide a useful tool for the early detection and histological classification of lung cancer [11–13].

The most common techniques used in analysis of plasma miRNAs mainly include quantitative reverse transcription polymerase chain

Address all correspondence to: Feng Jiang, Department of Pathology, The University of Maryland School of Medicine, 10 South Pine Street, MSTF 7th Floor, Baltimore, MD 21201-1192, USA. E-mail: fjiang@som.umaryland.edu

Received 23 April 2018; Revised 30 April 2018; Accepted 1 May 2018

© 2018 The Authors. Published by Elsevier Inc. on behalf of Neoplasia Press, Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1936-5233/18

<https://doi.org/10.1016/j.tranon.2018.05.001>

reaction (qRT-PCR) and microarray [14]. Being sensitive and quantitative, qRT-PCR is relatively inexpensive and flexible, and thus the preferred choice for assessment of limited number of miRNAs in relatively small experiments. Compared with qRT-PCR, microarrays can target more miRNAs for larger studies, however, have a low detection sensitivity. Since the cancer-associated miRNAs are derived from primary tumor and could be ‘diluted’ in a background of normal miRNAs, the circulating miRNAs are present in very low abundance in plasma [11,15,16]. In order to profile the low-abundance miRNAs in plasma, these current techniques require a large plasma volume to purify a sufficient RNA yield followed by complex downstream workflow, including reverse transcription [17]. However, small differences in the complex processes often result in large analytical variations and poor diagnostic performances of the plasma biomarkers [14,17].

FirePlex technology is a new approach for quantification of miRNAs [18]. This technique is dependent on hydrogel particles that contain custom selected probes against target miRNAs. miRNAs bind to these probes and are then ligated to adaptor sequences for pre-detection amplification. Particles are optimized for use with common bench-top flow cytometers, allowing detection without specialized laboratory equipment. Furthermore, data analysis is relatively straightforward and does not require advanced bioinformatics skills. This technique can quantify multiple miRNAs across a range of samples without the labor-intensive workflow or large sample requirement for other techniques. Importantly, it can be used directly with crude body fluids without RNA extraction and reverse transcription. In addition, miRNA profiling by the FirePlex technology can be achieved from input of as little as 10 µL of plasma or 100 pg purified RNA [18]. Therefore, FirePlex technology for multiplex miRNA profiling may overcome the obstacles of the current techniques by eliminating the errors induced during sample purification and insufficient methods for analytical performance. The objective of the present study was to use FirePlex miRNA assay to develop plasma miRNA biomarkers for lung cancer early detection and histological classification.

Materials and Methods

Patients and Clinical Specimens

The Institutional Review Boards of University of Maryland Baltimore and Veterans Affairs Maryland Health Care System

approved this study. We recruited lung cancer patients and cancer-free smokers by using the inclusion and/or exclusion criteria recommended by U.S. Preventive Services Task Force for lung cancer screening in heavy smokers [19]. We collected blood in BD Vacutainer spray-coated K2EDTA Tubes (BD, Franklin Lakes, NJ) and prepared plasma using the standard operating protocols developed by The NCI-Early Detection Research Network within 2 hours of the collection [20]. A total of 56 NSCLC patients and 28 cancer-free smokers were recruited. Among the cancer patients, 9 patients were female and 47 were male. Twelve had stage I NSCLCs, 6 with stage II, 12 with stage III, 18 with stage IV, and 7 with unknown stage. Histologically, 28 lung cancer patients were diagnosed with AC, while 28 with SCC. Of the cancer-free smokers, 2 patients were female and 28 were male. The cancer-free patients who were smokers and served as control subjects had granulomatous inflammation (n = 14), nonspecific inflammatory changes (n = 12) or lung infections (n = 2). The cancer-free smokers had been followed for at least 2 years, and none had any evidence of cancer. The demographic and clinical variables of the cohort are shown in Table 1.

miRNA Quantification by the FirePlex Assay

Using next-generation deep sequencing to comprehensively characterize miRNA profiles in eight lung tumor tissues, we identified miRNAs whose dysregulation was associated with NSCLC [21]. From the miRNAs, we found that 11 miRNAs might be potential plasma biomarkers for lung cancer [12,13,21–26]. The 11 miRNAs were analyzed in this study, including miRs-205-5p, -145, -422a, -34a-5p, -93-5p, -223-3p, -210-3p, and 628-3p, and lets-7d-5p, -7g-5p, and -7i-5p.

Plasma 20 µL was mixed with 36 µL Digest Buffer, 20 µL water and 4 µL Protease Mix and incubated at 60°C for 45 minutes with shaking. For each sample run, FirePlex Particles (35 µL) were added to a well of a 96-well filter plate and filtered. Twenty-five µL Hybe Buffer was added to each well. One ng total RNA was loaded. The plate was incubated at 37°C for 60 minutes with shaking. After rinsing twice with 1× Rinse A, 75 µL of 1× Labeling Buffer was added to each well. The plate was incubated at room temperature for 60 minutes with shaking. After two rinses with 1× Rinse B followed by one rinse with 1× Rinse A, a catch plate was added to the vacuum manifold and the filter plate put under constant vacuum. 65 µL of 95°C RNase-free water was added twice to each well to elute the ligated sample. 30 µL of this melt-off was added to a clean PCR plate and mixed with 20 µL PCR master mix. The mixture underwent 32 cycles of PCR amplification. 60 µL of Hybe Buffer was added back to each well of the original particles followed by 20 µL of the PCR product, and the plate was incubated at 37°C for 30 minutes with shaking. After rinsing twice with 1× Rinse B followed by one rinse with 1× Rinse A, 75 µL of 1× Reporting Buffer was added to each well and the plate incubated at room temperature for 15 minutes with shaking. After rinsing twice with 1× Rinse A, 175 µL of Run Buffer was added to each well. The samples were then scanned on an EMD Millipore Guava 6HT flow cytometer (MilliporeSigma, Germany). Flow cytometry quantification data was analyzed with the FirePlex™ Analysis Workbench software as previously described [18].

RNA Isolation and qRT-PCR

Plasma 200 µL was used for RNA extraction by Trizol LS reagent (Invitrogen Carlsbad, CA) and RNeasy Mini Kit (Qiagen, Hilden, Germany) [22,26,27]. RT was carried out to generate cDNA by using

Table 1. Characteristics of NSCLC Patients and Cancer-Free Smokers

	NSCLC Cases (n = 56)	Controls (n = 28)	P-Value
Age	65.68 (SD 10.09)	62.37 (SD 9.26)	0.262
Sex			
Female	9	2	0.2
Male	47	26	
Pack-years (median)	52.88	48.75	0.37
Stage			
Stage I	12		
Stage II	6		
Stage III	13		
Stage IV	18		
Unknown stage	7		
Histological type			
Adenocarcinoma	28		
Squamous cell carcinoma	28		

Abbreviations: NSCLC, non-small cell lung cancer; SD, standard deviation.

a RT Kit (Applied Biosystems, Foster City, CA) as described in our published works [22,26,27]. PCR was performed to measure expressions of target genes by using a PCR kit (Applied Biosystems) on a Bio-Red IQ5 Multi-color RT-PCR Detection System (Bio-Red, Hercules, CA). Expression levels of the genes were determined using comparative cycle threshold (C_T) method with miR-1228 as an internal control. The targeted genes with C_T values >35 were considered to be below the detection level of qRT-PCR [28]. All reactions were run in triplicate. C_T values were converted into absolute copy numbers using a standard curve from a synthetic miRNA as previously described [22,26,27].

Statistical Analysis

We used the area under receiver operating characteristic (ROC) curve (AUC) with the null hypothesis (H_0) of 0.5 to determine sample size. 28 subjects were required in each category of cases and controls to show a minimum difference of interest between an AUC of 0.75 versus an AUC of 0.5 with 80% power at the 5% significance level. To determine if a particular miRNA signal was detectable, assay background was subtracted from the miRNA signal measured in the samples and the result was compared to a limit of detection. The detection threshold was set based on comparison of fluorescence intensity of sample wells to negative (water) wells by calculating $[(3S_{ww})^2 + (4S_{pp})^2]^{1/2}$. S_{ww} was the well-to-well standard deviation of the mean signals, and S_{pp} was the mean of the particle-particle standard errors. The detection threshold was set at 2.00 as previously established [18]. We generated the heatmap based on a grid of raw and background subtracted miRNA signal intensities indicated by color. We used an ANOVA to determine if the groups exhibit significantly different expression levels for a particular miRNA, and the ANOVA results were reported as an unadjusted P -value. Differential expression with P -values less than 0.05 indicated that the two groups were different with at least a 95% confidence level. Pearson's correlation analysis was utilized to assess relationship between plasma miRNA expressions and demographic characteristics of the patients and cancer-free controls. Correlation between FirePlex versus qRT-PCR output analyses was tested by using linear regression models. The precision of miRNA measurements was estimated by using coefficient of variation [$CV = (SD/mean) \times 100$] of quadruplicate measures for each sample. The ROC and AUC analyses were applied to determine sensitivity and specificity of each miRNA. We used logistic regression models with constrained parameters as in least absolute shrinkage and selection operator (LASSO) to eliminate the irrelevant genes and develop prediction models with the highest sensitivity and specificity.

Results

Directly Targeting Plasma Samples for Quantification of miRNAs

The releases of miRNAs in plasma by hemolysis of blood cells often produce nonspecific and low reproducible results [16,29–32]. To determine if the plasma had hemolysis of blood cells, we included hemolysis-associated miRNA markers (miRs-451a and -486-5) in the FirePlex miRNA Assay. All the specimens had negative results of the hemolysis miRNA markers, suggesting no hemolysis in the plasma samples. All the 11 miRNAs, including miRs-205-5p, -145, -422a, -34a-5p, -93-5p, -223-3p, -210-3p, and 628-3p, and lets-7d-5p, -7g-5p, and -7i-5p, had background-subtracted signal

above the detection threshold (Figure 1). Therefore, the FirePlex miRNA Assay could quantify the 11 miRNAs in plasma.

To determine the robustness of the FirePlex assay in miRNA quantification, we performed qRT-PCR in plasma samples of 20 cancer-free controls and 20 lung cancer patients for quantitation of a miRNA, miRs-34a-5p. There was no statistical difference of the CVs of miRNA levels determined by FirePlex vs. qRT-PCR in the same samples ($P = 0.26$) (Supplementary Figure 1), implying that FirePlex had comparable precision and repeatability compared with qRT-PCR for quantification of plasma miRNA. Furthermore, there was a close correlations of expression levels of the miRNA determined by the two platforms (Pearson $r = 0.9961$, $P < 0.0001$) (Supplementary Figure 2). In addition, the total time required for the FirePlex assay is 3–5 hours, whereas 24 hours for qRT-PCR. Altogether, the FirePlex assay is a less time-consuming technique for quantification of plasma miRNAs with an equivalent analytic performance compared with qRT-PCR.

Plasma Biomarker Panels for the Early Detection and Histological Classification of Lung Cancer

Of the 11 miRNAs that robustly detectable in the plasma samples, 10 miRNAs (miRs-93-5p, 628-3p, -422a, -34a-5p, -223-3p, -210-3p, and -205-5p, and lets-7d-5p, -7g-5p, and -7i-5p) had a different expression level in lung cancer patients vs. cancer-free individuals (All $P < 0.05$) (Figure 2) (Supplementary Figure 3). We used ROC to evaluate the diagnostic performance of the miRNAs for lung cancer. The 10 miRNAs showed 0.663–0.77 AUC values with 58.93–73.21% sensitivity and 60.25–82.14% specificity (Table 2), implying that the miRNAs held promise as plasma biomarkers for lung cancer.

We used logistic regression models with constrained parameters as in LASSO and AUCs to optimize a marker panel for lung cancer. From the 10 genes, a logistic regression model with 2 miRNAs (miRs-205-5p and -210-3p) was developed for diagnosing lung cancer: $U = -1.44 + 0.05 \times \text{miR205-5p} + 0.07 \times \text{miR-210-3p}$. Incorporated analysis of the 2 biomarkers by using the logistic regression model produced a higher AUC (0.85) than did the individual biomarkers (Figure 3A) (Table 2) ($P < 0.05$). We calculated the distance to the perfect point (0, 1) with varying cut-offs for U , and the cut-off corresponding to the shortest distance in the AUC analysis was considered the optimal one. The optimal cut-off for the logistic regression model was $U = 0.62$. Any subject with $U \geq 0.62$ was classified as a lung cancer case. As a result, combined analysis of the 2 miRNAs by using the logistic regression model yielded significantly higher sensitivity (78.6%) and specificity (89.3%) compared with the individual biomarkers (all $P < 0.05$) (Table 2). Furthermore, based on the 2 miRNAs, a second logistic regression model was developed for diagnosis of AC with AUC of 0.82: $U = -1.69 + 0.05 \times \text{miR-205-5p} + 0.03 \times \text{miR-210-3p}$ (Figure 3B). When cut-off was set at $U = 0.53$, this logistic regression model produced 75.0% sensitivity and 89.3% specificity for diagnosis of AC. In addition, using the same 2 miRNAs, a third logistic regression model was established for diagnosis of SCC with AUC of 0.88: $U = -2.62 + 0.043 \times \text{miR-205-5p} + 0.1 \times \text{miR-210-3p}$ (Figure 3C). When cut-off was set at $U = 0.35$, this logistic regression model produced 82.1% sensitivity and 89.3% specificity for diagnosis of SCC.

Combined use of all the 10 miRNA genes did not produce higher sensitivity and specificity compared with the two miRNAs (miRs-205-5p and -210-3p) used in combination ($P > 0.05$).

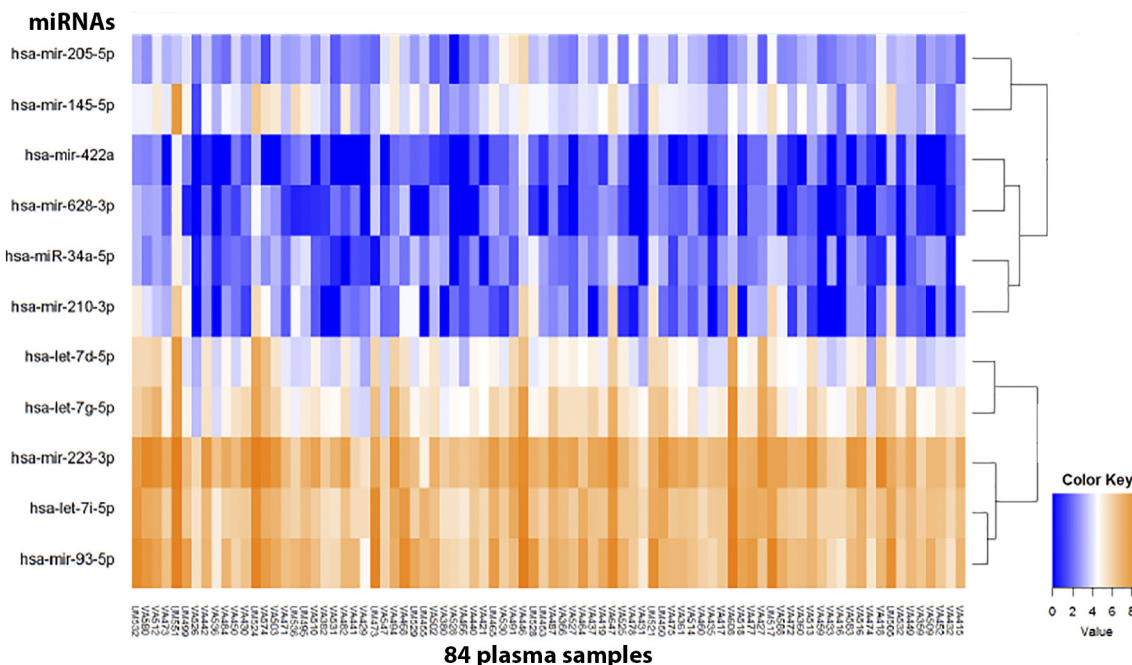


Figure 1. Directly targeting 84 plasma samples for quantifying 11 miRNAs. The heatmap is a grid of raw, background subtracted miRNA signal intensities indicated by color. The 11 genes have background-subtracted signal above the detection threshold (2.00) and thus are reliably quantified in plasma.

Furthermore, the estimated correlations among levels of the 2 genes were very low (Pearson $r = -0.0219$, $P = 0.7720$) (Supplementary Figure 4), implying that the integration of the two molecular biomarkers has complementary classification. In addition, the logistic regression models had no special association with age, gender, and smoking status of the participants (All $P > 0.05$). Moreover, the three logistic regression models with the same two miRNAs had similar sensitivity and specificity for diagnosis of different stages of NSCLC (Supplementary Figure 5).

Discussion

The current techniques for the assessment of plasma miRNAs require a large plasma volume followed by the complicated downstream handling. Small differences in the processing will lead to poor analytical and diagnostic values of plasma biomarkers. We propose to avoid the multiple procedures by direct analyzing plasma using the

FirePlex technology for quantifying multiple miRNAs. Our head-to-head comparison of the FirePlex Assay and qRT-PCR demonstrates that the FirePlex can eliminate intra-assay variables and improve reliability of miRNA quantification by directly targeting crude plasma and minimizing the complicated workflow. Furthermore, unlike qRT-PCR requiring a large sample volume (200 μL), the FirePlex assay can simultaneously quantify multiple miRNAs from input of as little as 20 μL of plasma. In addition, without the expensive procedures including RNA isolation, RNA quantification, and RT, the FirePlex technology can be used at less cost compared with qRT-PCR. The FirePlex assay only takes 3–5 hours from sample to data, and thus requires a shorter turnaround time compared with qRT-PCR demanding 24 hours. Moreover, the data handling for the FirePlex technology is straightforward for fast analysis and interpretation, as the results are obtained by using conventional flow cytometry with simple software. Therefore, the FirePlex assay might

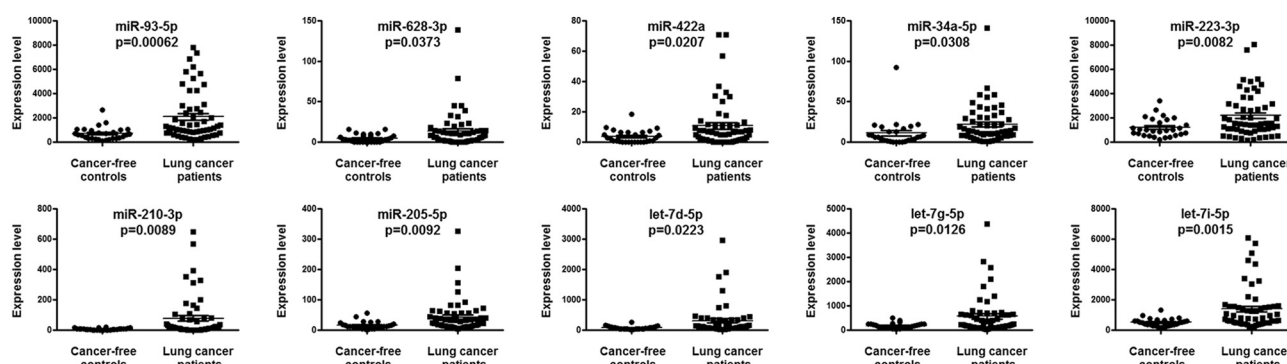


Figure 2. Of the 11 miRNAs tested in plasma samples of cancer-free smokers vs. lung cancer patients, 10 miRNAs have a significantly different level in the lung cancer patients vs. cancer-free smokers (all $P < 0.05$). The inside line denotes the median.

Table 2. miRNAs with a Different Expression Level in Lung Cancer Patients vs. Cancer-Free Smokers

miRNAs	AUC	Standard Deviation	P Value	Sensitivity (%)	Specificity (%)
mir-93-5p	0.7691	0.0518	<0.0001	71.43	71.43
miR-628-3p	0.6633	0.0600	0.0152	58.93	60.71
mir-422a	0.6773	0.0599	0.0084	62.5	60.71
miR-34a-5p	0.7165	0.0597	0.0013	67.86	67.86
mir-223-3p	0.6709	0.0592	0.0110	73.21	60.71
mir-210-3p	0.7545	0.0518	0.0002	64.29	82.14
miR-205-5p	0.7519	0.0543	0.0002	69.64	67.86
lets-7d-5p	0.6823	0.0368	0.0013	63.92	60.25
let-7i-5p	0.6678	0.0156	0.0002	59.66	61.26
let-7g-5p	0.7126	0.0387	0.0036	68.23	67.53
A panel of 2 miRNAs	0.8510	0.0162	0.0001	78.57	89.29

Abbreviations: AUC, area under the receiver-operator characteristic curve.

provide an efficient means for quantification of the plasma miRNAs whose changes are the hallmark of lung cancer.

Importantly, we develop 2 miRNA biomarkers, which used in combination could diagnose lung cancer. Since the biomarkers show similar sensitivity and specificity in the early vs. advanced stages of NSCLC, it might be a useful approach for the early detection of lung cancer, a clinically challenging. Furthermore, the use of the 2 miRNA biomarkers would potentially improve personalized treatment of lung cancer by precisely classifying the subtypes. In addition, our analysis of the plasma miRNAs by simply calculating the equation with the established cut-off could be a convenient tool in the clinics. Therefore, the developed biomarkers with the easy-to-use prediction models may represent promising noninvasive diagnostic tools not only for early detection of lung cancer, but improving the efficacy of therapeutics of NSCLC.

Increased miR-205-5p expression contributes to the development and progression of NSCLC [21,33]. We have shown that miR-205-5p is one of the miRNAs that could be used as biomarkers for the early detection of lung cancer [21,24,34]. miR-210-3p stimulates a hypoxic phenotype and upsurges radioresistance in NSCLCs [35]. Hypoxia-induced miR-210-3p can regulate tumor cell susceptibility to cytolytic T-lymphocyte-mediated lysis by a mechanism involving its downstream targets PTPN1, HOXA1, and

TP53I11 [36]. The results created from our present study support these previous findings, and further provide evidence of using the miRNAs as potential biomarkers for diagnosis of lung cancer.

There are limitations in the present study. 1), the sample size is small. Furthermore, this single and retrospective cohort of cases and controls may produce selection bias. To diminish the bias, we will perform pivotal evaluation of the diagnostic assay in a large cohort by using prospective-specimen collection, retrospective-blinded-evaluation design developed by NCI [37]. 2), in the present study, we only assess 11 miRNAs whose changes are associated with lung cancer, from which 2 miRNA biomarkers are optimized. Although showing promise, the 2 miRNAs have moderate sensitivity and specificity that are not sufficient in routine laboratory settings. We will use this platform to evaluate more lung tumor-specific miRNAs to identify additional plasma miRNA biomarkers that can be added to the current ones so that the diagnostic efficacy of the plasma assays could be improved. Furthermore, the assessments of cell-free circulating tumor DNA or DNA methylation status of gene promoters have attracted increasing attention as potential liquid biopsy tests for lung cancer [38–40]. Our ongoing efforts are to compare the diagnostic efficiency of the plasma miRNA signature with those of the cell-free DNA biomarkers in the early detection of lung cancer. 3), given that the changes of the miRNAs are not specific for lung cancer, the patents with positive results might have other type of cancers. In future clinical trials, the patients who have positive results with the molecular analysis should undertake clinical examinations and tests to have final clinical diagnosis.

Conclusions

This study demonstrates that an approach for directly targeting less blood and skipping the multipole procedures that are associated with large analytic variations would improve the efficacy of miRNA assessment. The developed plasma miRNA biomarkers may help diagnose lung cancer at the early stage and classify the subtypes. Nevertheless, the continued development of this technology and validation of the biomarkers are required.

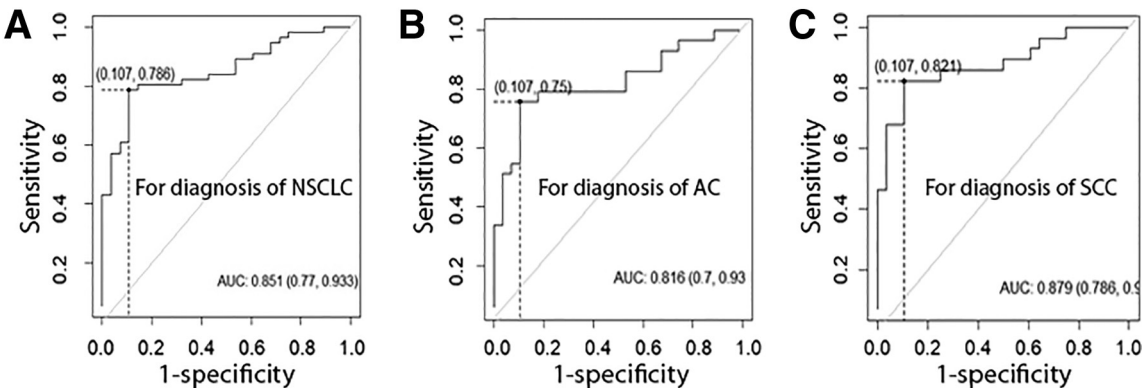


Figure 3. Receiver-operator characteristic (ROC) curves of 2 plasma miRNAs (miRs-205-5p and -210-3p). The area under the ROC curve (AUC) for the biomarkers conveys the accuracy. (A), a logistic regression model based on the 2 miRNAs creates an AUC value of 0.85 with 78.6% sensitivity and 89.3% specificity for diagnosis of non-small-cell lung cancer (NSCLC). (B), a logistic regression model based on the 2 miRNAs produces 0.82 AUC with 75.0% sensitivity and 89.3 % specificity for diagnosis of lung adenocarcinoma (AC). (C), a logistic regression model based on the 2 miRNAs generates 0.88 AUC with 82.1% sensitivity and 89.3% specificity for diagnosis of lung squamous cell carcinoma (SCC).

Declarations

Ethical Approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors.

The Institutional Review Boards of University of Maryland Baltimore and Veterans Affairs Maryland Health Care System approved this study. reference number is H40666.

Competing Interests

The authors declare no conflict of interest.

Funding

Grant support: This work was supported in part by NCI R21CA205746, VA Merit Award I01 CX000512, Award from the Geaton and JoAnn DeCesaris Family Foundation, UMD-UMB Research and Innovation Seed Grant, DoD-Idea Development Award, and Maryland Innovation Initiative (MII) Commercialization Program- Phase 1/2 Grant (F.J.).

Authors' Contributions

QL, YW, and FJ conducted the experiments and participated in study design, coordination, and data interpretation, and preparing the manuscript. All authors read and approved the final manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.tranon.2018.05.001>.

References

- [1] Siegel R, Naishadham D, and Jemal A (2013). Cancer statistics, 2013. *CA Cancer J Clin* **63**, 11–30.
- [2] Travis WD, Brambilla E, and Geisinger KR (2016). Histological grading in lung cancer: one system for all or separate systems for each histological type? *Eur Respir J* **47**, 720–723.
- [3] Patnaik S, Mallick R, Kannisto E, Sharma R, Bshara W, Yendamuri S, and Dhillon SS (2015). MiR-205 and MiR-375 microRNA assays to distinguish squamous cell carcinoma from adenocarcinoma in lung cancer biopsies. *J Thorac Oncol* **10**, 446–453.
- [4] Gressett SM and Shah SR (2009). Intricacies of bevacizumab-induced toxicities and their management. *Ann Pharmacother* **43**, 490–501.
- [5] Kazandjian D, Suzman DL, Blumenthal G, Mushti S, He K, Libeg M, Keegan P, and Pazdur R (2016). FDA Approval Summary: Nivolumab for the Treatment of Metastatic Non-Small Cell Lung Cancer With Progression On or After Platinum-Based Chemotherapy. *Oncologist* **21**, 634–642.
- [6] Ambros V (2003). MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing. *Cell* **113**, 673–676.
- [7] Galasso M, Sana ME, and Volinia S (2010). Non-coding RNAs: a key to future personalized molecular therapy? *Genome Med* **2**, 12–16.
- [8] Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, and Ferrando AA, et al (2005). MicroRNA expression profiles classify human cancers. *Nature* **435**, 834–838.
- [9] Iorio MV and Croce CM (2012). microRNA involvement in human cancer. *Carcinogenesis* **33**, 1126–1133.
- [10] Yanaihara N, Caplen N, Bowman E, Seike M, Kumamoto K, Yi M, Stephens RM, Okamoto A, Yokota J, and Tanaka T, et al (2006). Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. *Cancer Cell* **9**, 189–198.
- [11] Pritchard CC, Kroh E, Wood B, Arroyo JD, Dougherty KJ, Miyaji MM, Tait JF, and Tewari M (2012). Blood cell origin of circulating microRNAs: a cautionary note for cancer biomarker studies. *Cancer Prev Res (Phila)* **5**, 492–497.
- [12] Shen J and Jiang F (2012). Applications of MicroRNAs in the Diagnosis and Prognosis of Lung Cancer. *Expert Opin Med Diagn* **6**, 197–207.
- [13] Shen J, Stass SA, and Jiang F (2013). MicroRNAs as potential biomarkers in human solid tumors. *Cancer Lett* **329**, 125–136.
- [14] Levy B, Hu ZI, Cordova KN, Close S, Lee K, and Becker D (2016). Clinical Utility of Liquid Diagnostic Platforms in Non-Small Cell Lung Cancer. *Oncologist* **21**, 1121–1130.
- [15] Chen X, Ba Y, Ma L, Cai X, Yin Y, Wang K, Guo J, Zhang Y, Chen J, and Guo X, et al (2008). Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. *Cell Res* **18**, 997–1006.
- [16] Mitchell PS, Parkin RK, Kroh EM, Fritz BR, Wyman SK, Pogosova-Agadjanyan EL, Peterson A, Noteboom J, O'Brian KC, and Allen A, et al (2008). Circulating microRNAs as stable blood-based markers for cancer detection. *Proc Natl Acad Sci U S A* **105**, 10513–10518.
- [17] Tiberio P, Callari M, Angeloni V, Daidone MG, and Appierto V (2015). Challenges in using circulating miRNAs as cancer biomarkers. *Biomed Res Int* **2015**, 731–738.
- [18] Tackett MR and Diwan I (2017). Using FirePlex() Particle Technology for Multiplex MicroRNA Profiling Without RNA Purification Methods. *Mol Biol* **1654**, 209–219.
- [19] Humphrey LL, Deffebach M, Pappas M, Baumann C, Artis K, Mitchell JP, Zakher B, Fu R, and Slatore CG (2013). Screening for lung cancer with low-dose computed tomography: a systematic review to update the US Preventive services task force recommendation. *Ann Intern Med* **159**, 411–420.
- [20] Tuck MK, Chan DW, Chia D, Godwin AK, Grizzle WE, Krueger KE, Rom W, Sanda M, Sorbara L, and Stass S, et al (2009). Standard operating procedures for serum and plasma collection: early detection research network consensus statement standard operating procedure integration working group. *J Proteome Res* **8**, 113–117.
- [21] Ma J, Mannoor K, Gao L, Tan A, Guarnera MA, Zhan M, Shetty A, Stass SA, Xing L, and Jiang F (2014). Characterization of microRNA transcriptome in lung cancer by next-generation deep sequencing. *Mol Oncol* **8**, 1208–1219.
- [22] Shen J, Todd NW, Zhang H, Yu L, Lingxiao X, Mei Y, Guarnera M, Liao J, Chou A, and Lu CL, et al (2011). Plasma microRNAs as potential biomarkers for non-small-cell lung cancer. *Lab Invest* **91**, 579–587.
- [23] Leng Q, Lin Y, Jiang F, Lee CJ, Zhan M, Fang H, and Wang Y (2017). A plasma miRNA signature for lung cancer early detection. *Oncotarget* **8**, 111902–111911.
- [24] Li H, Jiang Z, Leng Q, Bai F, Wang J, Ding X, Li Y, Zhang X, Fang H, and Yfantis HG, et al (2017). A prediction model for distinguishing lung squamous cell carcinoma from adenocarcinoma. *Oncotarget* **8**, 50704–50714.
- [25] Ma J, Guarnera MA, Zhou W, Fang H, and Jiang F (2017). A Prediction Model Based on Biomarkers and Clinical Characteristics for Detection of Lung Cancer in Pulmonary Nodules. *Transl Oncol* **10**, 40–45.
- [26] Shen J, Liu Z, Todd NW, Zhang H, Liao J, Yu L, Guarnera MA, Li R, Cai L, and Zhan M, et al (2011). Diagnosis of lung cancer in individuals with solitary pulmonary nodules by plasma microRNA biomarkers. *BMC Cancer* **11**, 374–379.
- [27] Ma J, Jemal A, and Smith R (2013). Reply to lung cancer deaths averted by screening should be considered in the context of tobacco control policies. *Cancer* **119**, 3420–3421.
- [28] Guthrie JL, Seah C, Brown S, Tang P, Jamieson F, and Drews SJ (2008). Use of Bordetella pertussis BP3385 to establish a cutoff value for an IS481-targeted real-time PCR assay. *J Clin Microbiol* **46**, 3798–3799.
- [29] Jarry J, Schadenbord D, Greenwood C, Spatz A, and van Kempen LC (2014). The validity of circulating microRNAs in oncology: five years of challenges and contradictions. *Mol Oncol* **8**, 819–829.
- [30] Hunter MP, Ismail N, Zhang X, Aguda BD, Lee EJ, Yu L, Xiao T, Schafer J, Lee ML, and Schmittgen TD, et al (2008). Detection of microRNA expression in human peripheral blood microvesicles. *PLoS One* **3**, e3694.
- [31] Moldovan L, Batte KE, Trgovcich J, Wisler J, Marsh CB, and Piper M (2014). Methodological challenges in utilizing miRNAs as circulating biomarkers. *J Cell Mol Med* **18**, 371–390.
- [32] Arroyo JD, Chevillet JR, Kroh EM, Ruf IK, Pritchard CC, Gibson DF, Mitchell PS, Bennett CF, Pogosova-Agadjanyan EL, and Stirewalt DL, et al (2011). Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma. *Proc Natl Acad Sci U S A* **108**, 5003–5008.

- [33] Lebanony D, Benjamin H, Gilad S, Ezagouri M, Dov A, Ashkenazi K, Gefen N, Izraeli S, Rechavi G, and Pass H, et al (2009). Diagnostic assay based on hsa-miR-205 expression distinguishes squamous from nonsquamous non-small-cell lung carcinoma. *J Clin Oncol* **27**, 2030–2037.
- [34] Xing L, Todd NW, Yu L, Fang H, and Jiang F (2010). Early detection of squamous cell lung cancer in sputum by a panel of microRNA markers. *Mod Pathol* **23**, 1157–1164.
- [35] Grosso S, Doyen J, Parks SK, Bertero T, Paye A, Cardinaud B, Gounon P, Lacas-Gervais S, Noel A, and Pouyssegur J, et al (2013). MiR-210 promotes a hypoxic phenotype and increases radioresistance in human lung cancer cell lines. *Cell Death Dis* **4**e544.
- [36] Noman MZ, Buat S, Romero P, Ketari S, Janji B, Mari B, Mami-Chouaib F, and Chouaib S (2012). Hypoxia-inducible miR-210 regulates the susceptibility of tumor cells to lysis by cytotoxic T cells. *Cancer Res* **72**, 4629–4641.
- [37] Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, Winget M, and Yasui Y (2001). Phases of biomarker development for early detection of cancer. *J Natl Cancer Inst* **93**, 1054–1061.
- [38] Nie K, Jia Y, and Zhang X (2015). Cell-free circulating tumor DNA in plasma/serum of non-small cell lung cancer. *Tumour Biol* **36**, 7–19.
- [39] Balgouranidou I, Chimonidou M, Milaki G, Tsaroucha E, Kakolyris S, Georgoulas V, and Lianidou E (2016). SOX17 promoter methylation in plasma circulating tumor DNA of patients with non-small cell lung cancer. *Clin Chem Lab Med* **54**, 1385–1393.
- [40] Normanno N, Denis MG, Thress KS, Ratcliffe M, and Reck M (2017). Guide to detecting epidermal growth factor receptor (EGFR) mutations in ctDNA of patients with advanced non-small-cell lung cancer. *Oncotarget* **8**(7), 12501–12516.

Integrating Circulating Immunological and Sputum Biomarkers for the Early Detection of Lung Cancer

Jian Su, Qixin Leng, Yanli Lin, Jie Ma, Fangran Jiang, Cheng-Ju Lee, HongBin Fang and Feng Jiang

Department of Pathology, University of Maryland School of Medicine, Baltimore, MD, USA.

Biomarkers in Cancer
Volume 10: 1–6
© The Author(s) 2018
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1179299X18759297



ABSTRACT: We have demonstrated that assessments of microRNA (miRNA) expressions in circulating peripheral blood mononucleated cell (PBMC) and sputum specimens, respectively, may help diagnose lung cancer. To assess the individual and combined analysis of the miRNAs across the different body fluids for lung cancer early detection, we analyse a panel of 3 sputum miRNAs (miRs-21, 31, and 210) and a panel of 2 PBMC miRNAs (miRs-19b-3p and 29b-3p) in a discovery cohort of 68 patients with lung cancer and 66 cancer-free smokers. We find that integrating 2 sputum miRNAs (miRs-31 and 210) and 1 PBMC miRNA (miR-19b-3p) has higher sensitivity (86.8%) and specificity (92.4%) compared with the individual panels. The synergistic value of the integrated panel of 3 biomarkers is confirmed in a validation cohort, independent of stage and histological type of lung cancer, and patients' age, sex, and ethnicity. Integrating circulating immunological and sputum biomarkers could improve the early detection of lung cancer.

KEYWORDS: Lung cancer, early detection, biomarkers

RECEIVED: November 19, 2017. **ACCEPTED:** January 23, 2018.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in part by NC1 R21CA205746, VA Merit Award I01 CX000512, Award from the Geaton and JoAnn DeCesaris Family Foundation, UMD-UMB Research and Innovation Seed Grant, DoD-Idea Development Award, and Maryland Innovation Initiative (MII) Commercialization Program-Phase 1/2 grant (F.J.)

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Feng Jiang, Department of Pathology, University of Maryland School of Medicine, 10 S. Pine St. Baltimore, MD 21201, USA. Email: fjiang@som.umaryland.edu

Introduction

Lung cancer is the biggest cancer killer in both men and women. More than 85% lung cancers are non-small-cell lung cancers (NSCLCs), which mainly consist of adenocarcinoma (AC) and squamous cell carcinoma (SCC). Because the early detection of the disease can reduce the mortality, non-invasive biomarkers that can accurately and cost-effectively diagnose early-stage lung cancer are urgently needed.¹

Cigarette smoking is the leading cause of NSCLC. Smoking can cause damages across the entire respiratory epithelial cells. The process is called field cancerization. The field cancerization could be observed in histologically normal-appearing lung epithelium, which may anchor the same molecular aberrations as the ones in primary lung tumours.^{2,3} Examination of the exfoliated bronchial epithelia of airway in sputum can identify the lung tumour-related molecular alterations and hence provide a non-invasive and specific means for diagnosis of NSCLC.^{4–7} MicroRNAs (miRNAs) are small molecules and the aberrations contribute to carcinogenesis.⁸ We have found that aberrant miRNA expression detected in sputum biologically reflects those in primary lung cancer.^{9–14} We have also developed a 3-miRNA biomarker panel in sputum for diagnosis of NSCLC with 83% sensitivity and 88% specificity.^{15–17}

Immune evasion occurs as a primary and important event in tumorigenesis.^{18–20} Peripheral blood mononuclear cells (PBMCs) are the front-line defence against tumorigenesis within immune system.²¹ Furthermore, miRNAs have important effects on the immune cells' function, and the abnormalities contribute to the evasion proceedings of cancer cells in

immunological surveillance.^{22–24} Therefore, miRNA changes of PBMCs would provide novel biomarkers for diagnosis of cancer.^{18,19,21,25} In addition, miRNA biomarkers of PBMCs may overwhelm the difficulties of cell-free (eg, plasma or serum)-based biomarkers because of the following reasons: (1) circulating PBMCs are a rich and readily accessible cell source in blood and thus produce large RNA quantity that has high quality for consistent analysis,¹⁹ and (2) miRNAs are purified from the PBMCs and thus do not have other sources of extracellular miRNAs (RBC-, protein-, exosome-, and vesicle-related miRNAs). We recently identified 2 PBMC miRNA (miRs-19b-3p and 29b-3p) as circulating immunologic biomarkers for NSCLC with 73% sensitivity and 83% specificity.²⁶

Given that the miRNAs in sputum and peripheral PBMC have diverse functions in lung carcinogenesis through various cellular paths; here, we investigated whether integrating sputum and circulating immunological miRNA biomarkers would have a synergistic influence on the early detection of lung cancer.

Materials and Methods

Patients

The Institutional Review Boards at University of Maryland Baltimore and Veterans Affairs Maryland Health Care System reviewed and approved this study. Using the inclusion and/or exclusion criteria recommended by US Preventive Services Task Force for lung cancer screening in heavy smokers,²⁷ we enrolled subjects between the ages of 55 to 80 years who had not less than a 30 pack-year smoking



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<http://www.creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

Table 1. Characteristics of patients with NSCLC and cancer-free smokers in a discovery cohort.

	NSCLC CASES (N=68)	CONTROLS (N=66)	P VALUE
Age	64.46 (SD: 11.37)	64.68 (SD: 11.32)	.31
Sex			.32
Female	28	28	
Male	40	38	
Race			.40
White	51	50	
African American	17	16	
Pack-years (median)	34.56	33.27	.47
Stage			
Stage I	22		
Stage II	20		
Stage III-IV	26		
Histological type			
Adenocarcinoma	37		
Squamous cell carcinoma	31		

Abbreviation: NSCLC, non-small-cell lung cancer.

history and were former smokers (quit within 15 years). Exclusion criteria included pregnancy, current pulmonary infection, surgery within 6 months, radiotherapy within 1 year, and life expectancy of <1 year. A total of 117 patients with NSCLC and 116 cancer-free smokers were recruited. No difference of age, race, sex, and smoking status was observed in the lung cancer cases vs cancer-free smokers. The cases and controls were randomly grouped into 2 cohorts: a discovery cohort and a validation cohort. The discovery cohort comprised 68 patients with lung cancer and 66 cancer-free smokers. The discovery cohort was used to simultaneously evaluate sputum miRNAs and PBMC miRNAs in the same set of clinical specimens and develop an integrated panel of miRNA biomarkers for NSCLC. The validation cohort included 49 patients with lung cancer and 50 cancer-free smokers and was used to confirm the synergistic effect of the integrated panel of miRNAs on the early detection of NSCLC. Detailed demographic and clinical characteristics of the 2 cohorts are shown in Tables 1 and 2.

Sample collection and preparation

From the participants before they had any treatment, sputum and blood samples were collected as previously described.^{9-12,28-36} The preparation of airway bronchial epithelia from the sputum specimens was performed using a protocol developed in our previous studies.^{7,9-12,28-36} Peripheral blood was collected in BD Vacutainer spray-coated K₂EDTA Tubes

(BD, Franklin Lakes, NJ, USA), and PBMCs were isolated using Ficoll gradient centrifugation as previously described.²⁶

Assessing expressions of the miRNAs using quantitative reverse transcription polymerase chain reaction

RNA was extracted from the samples using our established protocols.^{9-12,26,30,31} The expressions of 3 miRNAs (miRs-21, 31, and 210) in sputum, and 2 miRNAs (miRs-19b-3p and 29b-3p) in PBMC were determined by using TaqMan miRNA assays (Applied Biosystems, Foster City, CA, USA).^{9-12,26,30,31} The levels of miRNAs were determined using a threshold cycle method.^{9-12,15,16} We included controls in each experiment: RNA extracted from a H358 NSCLC cell line (a positive control for RT [reverse transcription], preamplification, and ddPCR [Droplet Digital polymerase chain reaction]), genomic DNA (a positive control for genomic DNA detection), and nuclease-free water (control for contamination). All experiments were repeated 3 times.

Statistical analysis

We used univariate analysis to identify the miRNAs whose expression levels being related to NSCLC. The significantly associated factors were then analysed using multivariate logistic regression models with stepwise regression based on receiver operator characteristic (ROC) curve to select an optimal prediction model for NSCLC. We also generated a 95% confidence

Table 2. Characteristics of patients with NSCLC and cancer-free smokers in a validation cohort.

	NSCLC CASES (N = 49)	CONTROLS (N = 50)	P VALUE
Age	65.23 (SD: 10.48)	64.70 (SD: 11.15)	.30
Sex			.35
Female	19	20	
Male	30	30	
Race			.36
White	37	38	
African American	12	12	
Pack-years (median)	35.52	35.16	.43
Stage			
Stage I	15		
Stage II	15		
Stage III-IV	19		
Histological type			
Adenocarcinoma	26		
Squamous cell carcinoma	23		

Abbreviation: NSCLC, non-small-cell lung cancer.

interval for the difference in the area under the ROCs (AUCs) using the bootstrap.³⁷ The optimal cut-off value was generated using the Youden index.^{38,39} To compare different miRNA biomarker panels, we computed their AUCs to determine the sensitivity and specificity as previously described.⁴⁰

Results

The individual sputum miRNAs and PBMC miRNAs could distinguish patients with lung cancer from cancer-free controls

In the discovery cohort, the 3 sputum miRNAs (miRs-21, -31, and -210) and 2 PBMC miRNAs (miRs-19b-3p and 29b-3p) exhibited a different level in patients with NSCLC vs cancer-free smokers (all P 's < .05). The analysis of the 3 sputum miRNAs created 0.923 AUC (Figure 1). Successively, the examination of the 3 sputum miRNAs produced 82.3% sensitivity and 87.9% specificity for diagnosis of lung cancer (Table 3). The level of miR-21 in sputum was related to AC (P < .05); however, miR-210 was correlated to SCC (P < .05). Overall, the 3 sputum miRNA panel did not have specific relationship with stage and histological type of lung cancer, and patients' age, race, and sex (all P 's > .05), but smoking history (P < .05).

The analysis of the 2 PBMC miRNAs had an AUC of 0.837 (Figure 1), producing 72.1% sensitivity and 81.8% specificity for diagnosis of NSCLC (Table 3). Furthermore, the 2-PBMC miRNA biomarker panel created 80.7% sensitivity and 89.4% specificity for SCC, and 75.7% sensitivity

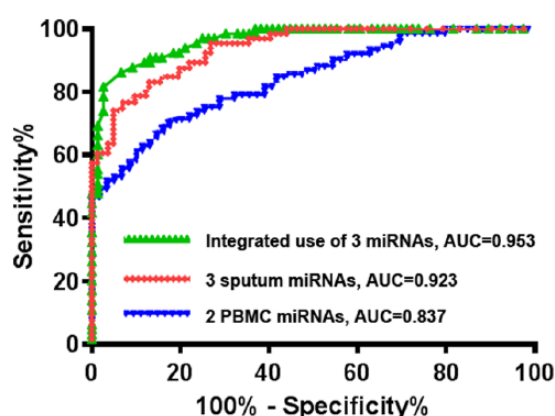


Figure 1. The comparison of integrated miRNA biomarkers with panels of sputum and PBMC miRNAs in a discovery cohort. A prediction model based on integrated use of 3 miRNA biomarkers (miRs-31, 210, and 19b-3p) across sputum and PBMC specimens was developed for distinguishing lung cancer patients from cancer-free smokers. The ROC curve of the integrated miRNA biomarkers produced a higher AUC (0.953), as compared with the panel of sputum miRNAs (0.923) and the panel of PBMC miRNAs (0.837) (all P 's < .05). AUC indicates area under the ROC; miRNAs, microRNAs; PBMC, peripheral blood mononucleated cell.

and 68.2% specificity for AC. As a result, the 2-PBMC miRNA biomarker panel had a higher diagnostic value for SCC compared with AC. The 2 PBMC miRNA biomarkers did not have relationship with stage of NSCLC, and patients' age, sex, and ethnicity of the participants (all P 's ≥ .05), except the smoking history (P = .03). Altogether, the results from

Table 3. The comparison of integrated panel of 2 sputum and 1 PBMC miRNA biomarkers with individual panels of sputum and PBMC miRNA biomarkers in a discovery cohort and a validation cohort.

	A DISCOVERY COHORT		A VALIDATION COHORT	
	SENSITIVITY (95% CI)	SPECIFICITY (95% CI)	SENSITIVITY (95% CI)	SPECIFICITY (95% CI)
3 sputum miRNAs	82.35 (71.20-90.53)	87.88 (77.51-94.62)	81.63 (67.98-91.24)	86.00 (73.26-94.18)
2 PBMC miRNAs	72.06 (59.85-82.27)	81.82 (70.39-90.24)	71.43 (56.74-83.42)	80.00 (66.28-89.97)
Integrated 2 sputum and 1 PBMC miRNAs	86.76 (76.36-93.77)	92.42 (83.20-97.49)	85.71 (72.76-94.06)	92.00 (80.77-97.78)

Abbreviations: CI, confidence interval; miRNAs, microRNAs; PBMC, peripheral blood mononucleated cell.

this study confirmed our prior discoveries^{15,26} that both the sputum and PBMC miRNAs could have the potential as biomarkers for NSCLC.

Incorporated examination of 2 miRNA and 1 PBMC miRNA biomarkers has a synergistic influence on the early detection of NSCLC

In the discovery cohort, logistic regression model was applied to define performance of different patterns of the 3 sputum and 2 PBMC miRNA biomarkers. From the 5 genes, 2 sputum miRNAs (miR-31 and 210) and 1 PBMC miRNA (miR-19b-3p) were selected as the best markers (all P 's < .001) and integrated into a logistic model. Integrated use of the 3 biomarkers created a greater AUC (0.953) than did the 3-sputum biomarker panel (0.923) or the 2-PBMC biomarker panel (0.837) (P < .05) (Figure 1). Using Youden index, we set up optimal cut-off at 0.64. As a result, integrated use of the 2 sputum miRNAs and 1 PBMC miRNA produced higher sensitivity (86.8%) and specificity (92.4%) than did the 3-sputum biomarkers and the 2-PBMC biomarkers (all P 's < .05) (Table 3). Furthermore, combined use of all the 5 biomarkers (3 miRNA and 2 PBMC biomarkers) did not produce higher sensitivity and specificity compared with the integrated panel of the only 3 biomarkers (P > .05). Moreover, integrated analysis of the 3 biomarkers across the different body fluids did not show special association with stage and histology of lung cancer, and patients' age, race, and sex (all P 's > .05). Importantly, the estimated Pearson correlation among levels of the 3 miRNAs was very low (all P 's > .05), suggesting that the combination had complementary classification for diagnosis of lung cancer.

Validating the synergistic effect of the integrated panel of 3 miRNAs across the body fluids for detection of NSCLC

In the validation cohort, the integrated panel of 3 biomarkers was blindly confirmed in the specimens of 49 patients with lung cancer and 50 cancer-free smokers. The integration of the 3 biomarkers across sputum and blood specimens showed higher sensitivity (85.7% vs 81.6% and 71.4%) and specificity (92.0% vs 86.0% and 80.0%) than did the 3-sputum miRNA

panel and 2-PBMC miRNA panel (all P 's < .05) (Table 3). Inconsistent with the findings in the discovery cohort, integrated analysis of the 3 biomarkers had no special relationship with stage and histological type of lung cancer, and patients' age, race, and sex (all P 's > .05). Furthermore, the Pearson correlation between the 3 miRNAs' expression levels was very low (all P 's > .05), supporting the synergistic value of this integrated panel of the 3 biomarkers for the early detection of NSCLC.

Discussion

The development of sputum biomarkers is based on the concept that lung tumorigenesis developed in a field defect characterized by molecular aberrations in the airways. According to the field defect concept, the alterations of miRNAs in bronchial airways reflect those in primary lung tumours in the distal lungs.⁶ The analysis of exfoliated airway bronchial epithelia in sputum for the altered miRNA expressions might indicate the lung tumour-related molecular alterations and hence identify NSCLC. On the contrary, the development of PBMC miRNAs as circulating tumour markers is based on the concept that PBMCs act as the primary defence line against cancer, and the abnormal miRNA expressions of PBMC contributes to the escaping events in immunological surveillance.^{18,19,21,22,25} Based on the different concepts, miRNA expression profile in PBMCs, unlike miRNAs detected in sputum cells, may not signify the abnormal miRNA changes in the lung tumours or mimic that in the primary cancer tissues.⁴¹⁻⁴³ Therefore, there was no overlap between the PBMC miRNA vs sputum miRNA profiles of patients with lung cancer.^{10,11,15,26,44} Given that altered miRNAs in sputum and PBMCs contribute to lung tumorigenesis via the different mechanisms,^{5,9,10,22,26} we hypothesize that integrating the miRNAs across the different body fluids may have a synergistic effect and thus improve the diagnosis of lung cancer. Indeed, the combined analysis of the 2-sputum miRNA and 1-PBMC miRNA generated a better performance than did a single category of the markers. Also, there was no association among the miRNA aberrations in the 2 surrogate clinical materials, further supporting that the performances of the 2 categories of miRNA biomarkers might have a synergistic value. In addition, unlike PBMC miRNA markers that are

more specific to SCC, the integration of sputum and PBMC biomarkers was not related to histology of lung tumour and hence substantiates the utility for predicting lung cancer. Moreover, integrated use of the 3 miRNAs had a comparable diagnostic performance for lung tumour at the early vs late stages. Altogether, the findings would be an important feature if the assay is used for identifying early-stage lung cancer.

Dysregulations of miRs-31, 210, and 19b-3p have been found to associate with a variety of tumours. For example, miR-31 downregulation in breast cancer is related with tumour metastases, whereas elevated miR-31 expression in colorectal cancer may associate with late stage of the tumours.^{45,46} MiR-31 dysregulation endorses KRAS mutation-driven NSCLC and hence plays a key role in lung tumorigenesis.⁴⁷ MiR-210 stimulates a hypoxic phenotype and upsurges radioresistance in NSCLCs.⁴⁸ The analysis of miR-210 in serum can diagnose several types of malignancies, including NSCLC.^{49–51} Abnormal miR-19b-3p expression has not been described in the immune system. Our ongoing project is to investigate the mechanisms, by which miR-19b-3p may involve in lung cancer cells' evasion of immunological surveillance.

This study may have some limitations. (1) Clinically useful biomarkers should have very high sensitivity and specificity. Integrated use of 3 miRNAs has a sensitivity of 88.2% and a specificity of 92.4%, which may not be sufficient in the laboratory settings. We will find new biomarkers that can be used together with the current ones to improve clinical performance of the assay. Furthermore, an ongoing project in our laboratory is to simultaneously evaluate cell-free miRNAs (eg, plasma- and serum-derived miRNAs) and PBMC miRNAs using the same clinical specimens to compare the diagnostic values of the 2 different miRNA biomarkers and investigate whether there is synergistic diagnosis between them. (2) The clinical samples were obtained from the patients who already have clinical diagnosis. The subjects might not represent heavy smokers in a lung cancer screening programme. We will conduct a large lung cancer screening trial to confirm the performance of the integrated panel of miRNA biomarkers. (3) Using low-dose computed tomography (LDCT) for the early detection of lung cancer can reduce the mortality.⁵² However, LDCT always produces over-diagnosis or a false-positive rate, presenting a major clinical obstacle for lung cancer early detection.⁵² It would be interesting to evaluate whether the 3 miRNA biomarkers could complement LDCT for the early detection of lung cancer by specifically reducing its over-diagnosis.

Conclusions

We have for the first time demonstrated that integrating circulating and sputum markers has a synergistic value and presents a potential approach for the early detection of lung cancer. Nevertheless, a large multi-centre trial is required to confirm the integrated method before it could be translated in laboratory settings.

Author Contributions

JS, QL, YL, JM, FJ, and C-JL conducted the experiments and participated in data interpretation. HF and FJ participated in study design, coordination, and data analysis and interpretation, and prepared the manuscript. All authors read and approved the final manuscript.

Disclosures and Ethics

The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material.

REFERENCES

1. American Cancer Society. *Cancer Facts & Figures for African Americans*. Atlanta, GA: American Cancer Society.
2. Tang X, Shigematsu H, Bekele BN, et al. EGFR tyrosine kinase domain mutations are detected in histologically normal respiratory epithelium in lung cancer patients. *Cancer Res*. 2005;65:7568–7572.
3. Solis LM, Behrens C, Raso MG, et al. Histologic patterns and molecular characteristics of lung adenocarcinoma associated with clinical outcome. *Cancer*. 2012;118:2889–2899.
4. Kadara H, Shen L, Fujimoto J, et al. Characterizing the molecular spatial and temporal field of injury in early-stage smoker non-small cell lung cancer patients after definitive surgery by expression profiling. *Cancer Prev Res (Phila)*. 2013;6:8–17.
5. Kadara H, Wistuba II. Field cancerization in non-small cell lung cancer: implications in disease pathogenesis. *Proc Am Thorac Soc*. 2012;9:38–42.
6. Spira A, Beane JE, Shah V, et al. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nat Med*. 2007;13:361–366.
7. Belinsky SA, Liechty KC, Gentry FD, et al. Promoter hypermethylation of multiple genes in sputum precedes lung cancer incidence in a high-risk cohort. *Cancer Res*. 2006;66:3338–3344.
8. Croce CM, Calin GA. miRNAs, cancer, and stem cell division. *Cell*. 2005;122:6–7.
9. Shen J, Liao J, Guarnera MA, et al. Analysis of microRNAs in sputum to improve computed tomography for lung cancer diagnosis. *J Thorac Oncol*. 2014;9:33–40.
10. Xie Y, Todd NW, Liu Z, et al. Altered miRNA expression in sputum for diagnosis of non-small cell lung cancer. *Lung Cancer*. 2010;67:170–176.
11. Xing L, Todd NW, Yu L, Fang H, Jiang F. Early detection of squamous cell lung cancer in sputum by a panel of microRNA markers. *Mod Pathol*. 2010;23:1157–1164.
12. Yu L, Todd NW, Xing L, et al. Early detection of lung adenocarcinoma in sputum by a panel of microRNA markers. *Int J Cancer*. 2010;127:2870–2878.
13. Pang W, Tian X, Bai F, et al. Pim-1 kinase is a target of miR-486-5p and eukaryotic translation initiation factor 4E, and plays a critical role in lung cancer. *Mol Cancer*. 2014;13:240.
14. Wang J, Tian X, Han R, et al. Downregulation of miR-486-5p contributes to tumor progression and metastasis by targeting protumorigenic ARHGAP5 in lung cancer. *Oncogene*. 2014;33:1181–1189.
15. Xing L, Su J, Guarnera MA, et al. Sputum microRNA biomarkers for identifying lung cancer in indeterminate solitary pulmonary nodules. *Clin Cancer Res*. 2015;21:484–489.
16. Su J, Liao J, Gao L, et al. Analysis of small nucleolar RNAs in sputum for lung cancer diagnosis. *Oncotarget*. 2016;7:5131–5142.
17. Su Y, Guarnera MA, Fang H, Jiang F. Small non-coding RNA biomarkers in sputum for lung cancer diagnosis. *Mol Cancer*. 2016;15:36.
18. Bluth M, Lin YY, Zhang H, Viterbo D, Zenilman M. Use of gene expression profiles in cells of peripheral blood to identify new molecular markers of acute pancreatitis. *Arch Surg*. 2008;143:227–233 (discussion 233–234).
19. Baine MJ, Chakraborty S, Smith LM, et al. Transcriptional profiling of peripheral blood mononuclear cells in pancreatic cancer patients identifies novel genes with potential diagnostic utility. *PLoS ONE*. 2011;6:e17014.
20. Melief CJ, Finn OJ. Cancer immunology. *Curr Opin Immunol*. 2011;23:234–236.

21. Twine NC, Stover JA, Marshall B, et al. Disease-associated expression profiles in peripheral blood mononuclear cells from patients with advanced renal cell carcinoma. *Cancer Res.* 2003;63:6069–6075.
22. Sasaki K, Kohanbash G, Hoji A, et al. miR-17-92 expression in differentiated T cells: implications for cancer immunotherapy. *J Transl Med.* 2010;8:17.
23. Bhatia A, Kumar Y. Cellular and molecular mechanisms in cancer immune escape: a comprehensive review. *Expert Rev Clin Immunol.* 2014;10:41–62.
24. Kohlhaup FJ, Mitra AK, Lengyel E, Peter ME. MicroRNAs as mediators and communicators between cancer cells and the tumor microenvironment. *Oncogene.* 2015;34:5857–5868.
25. Whitney AR, Diehn M, Popper SJ, et al. Individuality and variation in gene expression patterns in human blood. *Proc Natl Acad Sci U S A.* 2003;100:1896–1901.
26. Ma J, Lin Y, Zhan M, Mann DL, Stass SA, Jiang F. Differential miRNA expressions in peripheral blood mononuclear cells for diagnosis of lung cancer. *Lab Invest.* 2015;95:1197–1206.
27. Humphrey LL, Deffebach M, Pappas M, et al. Screening for lung cancer with low-dose computed tomography: a systematic review to update the US preventive services task force recommendation. *Ann Intern Med.* 2013;159:411–420.
28. Romeo MS, Sokolova IA, Morrison LE, et al. Chromosomal abnormalities in non-small cell lung carcinomas and in bronchial epithelia of high-risk smokers detected by multi-target interphase fluorescence in situ hybridization. *J Mol Diagn.* 2003;5:103–112.
29. Yu L, Shen J, Mannoor K, Guarnera M, Jiang F. Identification of ENO1 as a potential sputum biomarker for early-stage lung cancer by shotgun proteomics. *Clin Lung Cancer.* 2014;15:372–378.
30. Li N, Ma J, Guarnera MA, Fang H, Cai L, Jiang F. Digital PCR quantification of miRNAs in sputum for diagnosis of lung cancer. *J Cancer Res Clin Oncol.* 2014;140:145–150.
31. Anjuman N, Li N, Guarnera M, Stass SA, Jiang F. Evaluation of lung flut in sputum samples for molecular analysis of lung cancer. *Clin Transl Med.* 2013;2:15.
32. Jiang F, Todd NW, Li R, Zhang H, Fang H, Stass SA. A panel of sputum-based genomic marker for early detection of lung cancer. *Cancer Prev Res (Phila).* 2010;3:1571–1578.
33. Jiang F, Todd NW, Qiu Q, Liu Z, Katz RL, Stass SA. Combined genetic analysis of sputum and computed tomography for noninvasive diagnosis of non-small-cell lung cancer. *Lung Cancer.* 2009;66:58–63.
34. Katz RL, Zaidi TM, Fernandez RL, et al. Automated detection of genetic abnormalities combined with cytology in sputum is a sensitive predictor of lung cancer. *Mod Pathol.* 2008;21:950–960.
35. Qiu Q, Todd NW, Li R, et al. Magnetic enrichment of bronchial epithelial cells from sputum for lung cancer diagnosis. *Cancer.* 2008;114:275–283.
36. Li R, Todd NW, Qiu Q, et al. Genetic deletions in sputum as diagnostic markers for early detection of stage I non-small cell lung cancer. *Clin Cancer Res.* 2007;13:482–487.
37. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology.* 1983;148:839–843.
38. Luo J, Xiong C. Youden index and associated cut-points for three ordinal diagnostic groups. *Commun Stat Simul Comput.* 2013;42:1213–1234.
39. Schisterman EF, Faraggi D, Reiser B, Hu J. Youden index and the optimal threshold for markers with mass at zero. *Stat Med.* 2008;27:297–315.
40. Samawi HM, Yin J, Rochani H, Panchal V. Notes on the overlap measure as an alternative to the Youden index: how are they related? *Stat Med.* 2017;36:4230–4240.
41. Shen J, Liu Z, Todd NW, et al. Diagnosis of lung cancer in individuals with solitary pulmonary nodules by plasma microRNA biomarkers. *BMC Cancer.* 2011;11:374.
42. Sozzi G, Boeri M, Rossi M, et al. Clinical utility of a plasma-based miRNA signature classifier within computed tomography lung cancer screening: a correlative MILD trial study. *J Clin Oncol.* 2014;32:768–773.
43. Jarry J, Schadendorf D, Greenwood C, Spatz A, van Kempen LC. The validity of circulating microRNAs in oncology: five years of challenges and contradictions. *Mol Oncol.* 2014;8:819–829.
44. Ma J, Guarnera MA, Zhou W, Fang H, Jiang F. A prediction model based on biomarkers and clinical characteristics for detection of lung cancer in pulmonary nodules. *Transl Oncol.* 2017;10:40–45.
45. Wang S, Hu J, Zhang D, Li J, Fei Q, Sun Y. Prognostic role of microRNA-31 in various cancers: a meta-analysis. *Tumour Biol.* 2014;35:11639–11645.
46. Laurila EM, Kallioniemi A. The diverse role of miR-31 in regulating cancer associated phenotypes. *Gene Chromos Cancer.* 2013;52:1103–1113.
47. Edmonds MD, Boyd KL, Moyo T, et al. MicroRNA-31 initiates lung tumorigenesis and promotes mutant KRAS-driven lung cancer. *J Clin Invest.* 2016;126:349–364.
48. Grosso S, Doyen J, Parks SK, et al. MiR-210 promotes a hypoxic phenotype and increases radioresistance in human lung cancer cell lines. *Cell Death Dis.* 2013;4:e544.
49. Chang W, Lee CY, Park JH, et al. Survival of hypoxic human mesenchymal stem cells is enhanced by a positive feedback loop involving miR-210 and hypoxia-inducible factor 1. *J Vet Sci.* 2013;14:69–76.
50. Lawrie CH, Gal S, Dunlop HM, et al. Detection of elevated levels of tumour-associated microRNAs in serum of patients with diffuse large B-cell lymphoma. *Br J Haematol.* 2008;141:672–675.
51. Papaconstantinou IG, Manta A, Gazouli M, et al. Expression of microRNAs in patients with pancreatic cancer and its prognostic significance. *Pancreas.* 2013;42:67–71.
52. Patz EF Jr, Pinsky P, Gatsonis C, et al; NLST Overdiagnosis Manuscript Writing Team. Overdiagnosis in low-dose computed tomography screening for lung cancer. *JAMA Intern Med.* 2014;174:269–274.

A CRISPR Test for Detection of Circulating Nuclei Acids



Jen-Hui Tsou, Qixin Leng and Feng Jiang*

Department of Pathology, University of Maryland School of Medicine, 10 South Pine Street, Baltimore, MD, USA

Abstract

Emerging CRISPR-based nucleic acid detection shows great promise in molecular diagnosis of diseases. CRISPR-Cas12a can sensitively and specifically detect human papillomavirus (HPV) DNA in anal swabs. However, the current CRISPR-Cas12a system needs auxiliary and expensive equipment, which limit its application as a point-of-care (POC) diagnostic tool. This study aimed to develop CRISPR-Cas12a as a POC test to directly target plasma for circulating HPV DNA detection by immediately reading results with naked eyes. Cell-cultured supernatants of either HPV16- or 18-positive cancer cells were treated with lysis buffer followed by isothermal amplification without DNA isolation. Cas12a, crRNA, and fluorescent-biotin reporters were incubated with the lysates. Our data showed that integrating CRISPR-Cas12a with lateral-flow strips could directly and specifically detect HPV16 and 18 in the liquid samples with the same limit of detection (0.24 fM) as did polymerase chain reaction but requiring less time. Furthermore, the CRISPR-Cas12a system could rapidly detect presence of HPV16 and HPV18 in plasma samples of 13 of 14 and 3 of 10 the patients with histopathological diagnosis of cervical cancer, respectively. Therefore, a CRISPR-Cas12a-based POC system was developed for conveniently detecting circulating nuclei acid targets in body fluids without requiring technical expertise and ancillary machineries.

Translational Oncology (2019) 12, 1566–1573

Introduction

The development of accurate and rapid nucleic acid detection for cancer and pathogen diagnoses, genotyping, and disease monitoring is clinically imperative. The current techniques, including polymerase chain reaction (PCR), have been used for nuclei acid detection [1]. However, they are costly, labor intensive, and time consuming. Furthermore, the techniques require extensive sample manipulation and expensive machineries. Rapid, reliable, and easy-to-use tests of circulating nucleic acids allowing for point-of-care (POC) without requiring special technical expertise and ancillary equipment are urgently needed.

Clustered regularly interspaced short palindromic repeats (CRISPR) are a family of DNA sequences found within the genomes of prokaryotic organisms [2]. CRISPR-associated (Cas) immune system has been applied in molecular biology to target and cleave specific nucleic acid sequences, which is commonly used in gene editing. Furthermore, upon binding to target double-stranded DNA (dsDNA) or RNA, several Cas proteins can be activated and unleash the nonspecific endoribonuclease activity to degrade the single-stranded DNA (ssDNA) and RNA and thus provide a novel diagnostic approach for nuclei acid detection [3–7]. For instance, based on Cas12a ssDNase activation, Chen et al. recently developed a

method termed DNA endonuclease-targeted CRISPR trans reporter (DETECTR), which could detect human papillomavirus (HPV) in cell-based clinical specimens (anal swabs) [5]. However, DETECTR has some limitations in its application as a field-deployable POC tool: 1) requiring complicated process of sample and DNA preparations; 2) reliance on fluorescent detection equipment for readout; and 3) applied only to anal swabs, which are cellular specimens.

To overcome the limitations, in this study, we developed CRISPR-Cas12a as a potential POC testing for HPV by directly targeting plasma without DNA isolation and by using a visual readout with naked eyes on a paper strip.

Address all correspondence to: Feng Jiang, Department of Pathology, University of Maryland School of Medicine, 10 South Pine Street, Baltimore, MD, USA. E-mail: fjiang@som.umaryland.edu

Received 5 August 2019; Revised 26 August 2019; Accepted 29 August 2019

© 2019 The Authors. Published by Elsevier Inc. on behalf of Neoplasia Press, Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1936-5233/19

<https://doi.org/10.1016/j.tranon.2019.08.011>

Materials and Methods

Cell Culture

All cancer cell lines [SiHa (ATCC HTB-35), Ca Ski (ATCC CRL-1550), UPCI: SCC152 (ATCC CRL-3240), HeLa (ATCC® CCL-2), C-4 I (ATCC® CRL-1594)] were obtained from the ATCC (Manassas, VA). The cell lines were cultured in complete growth medium according to ATCC's instruction and were grown at 37°C under 5% CO₂.

Clinical Specimens

From July 2006 to July 2012, we have recruited cancer-free female patients without any cancer under our protocol approved by the Institutional Review Boards of University of Maryland Baltimore. EDTA-anticoagulant blood samples were collected at the time of the interview by venipuncture from the consented subjects as previously described [8]. The samples were centrifuged at 750×g for 10 minutes, and the plasma fractions were stored at –80°C [8]. We randomly obtained plasma from 14 female subjects without diagnosis of any cancer as a control group. Fifteen plasma samples of cervical cancer patients were obtained from Tissue Solutions (Glasgow, UK) and used as a case group. The 15 cervical cancer patients underwent complete physical and gynecologic examinations, and their disease was staged according to the guidelines of the International Federation of Gynecology and Obstetrics. The demographic and clinical variables of the cases and controls are shown in Table 1.

Crude DNA Preparation and Recombinase Polymerase Amplification (RPA)

For cultured cell supernatants, 25 µl of each supernatant sample was mixed with 25 µl of phosphate buffered saline (PBS) (Corning) containing 0.2% Triton X-100 (Promega) followed by boiled mixture at 95°C for 5 minutes. Five microliters of the preparation was used as direct input into RPA. RPA reaction was run as instructed with TwistAmp Basic (TwistDx). The 50-µl reaction mixture contained 0.48 µM forward and reverse primers (Table 2), 29.5 µl primer free rehydration buffer, 5 µl DNA from boiled liquid fluids, and 14 mM magnesium acetate (MgOAc). The PRA mixture was incubated at 37°C for 20 minutes. Five microliters of RPA reaction was used as direct input. The samples were then preceded with heating unextracted diagnostic samples to obliterate nucleases (HUDSON) [3]. For human plasma, we made a minor modification to

Table 2. Sequences of RPA and PCR Primers, Guide RNA, and Report Substrates

Primers for RPA and PCR	
HPV16-L1_F	TTGTTGGGGTAACCACTATTGTGTACTGTT
HPV16-L1_R	CCTCCCCATGTCTGAGGTACTCCTTAAAG
HPV18-L1_F	GCATAATCAATTATTTGTACTGTGGTAGATACCACT
HPV18-L1_R	GCTATACTGCTTAAATTTGGTAGCATCATATTGC
Guide RNA	Sequence (5'→3')
Cas12a crRNA-HPV16-L1	UAAUUUCUACUAGUGUAGAUUGAAGUAGAUUGGCAGCAC
Cas12a crRNA-HPV18-L1	UAAUUUCUACUAGUGUAGAUACAUAUGUGCUUCUACACA
Substrates	Sequence (5'→3')
ssDNA-FQ reporter	/56-FAM/TTATT/3IABKFQ/
ssDNA-FB reporter	/56-FAM/TTATT/3Bio/

Abbreviations: RPA, recombinase polymerase amplification; PCR, polymerase chain reaction; and FQ, FAM-quencher.

HUDSON. Briefly, plasma was first diluted 1:3 to avoid solidification. Twenty-five microliters of plasma sample was mixed with 50 µl of PBS (Corning) containing 0.53% Triton X-100 (Promega) followed by boiled mixture at 95°C for 5 minutes.

Plasmid DNA Preparation and 10-Fold Serial Dilution

Plasmid of HPV type 16 (ATCC 45113) and HPV type 18 (ATCC 45152) in *Escherichia coli* DH5α were obtained from ATCC and grown in LB medium with 50 µg/ml ampicillin at 37°C for 16 hours in a shaking incubator. Plasmid DNA was isolated using QIAprep spin miniprep kit (QIAGEN) according to manufacturer's instruction. Briefly, the bacterial culture was harvested and lysed. The plasmid DNA was adsorbed on a QIAprep membrane, washed, and then eluted with nuclease free water. Concentration was measured by Qubit 4 (ThermoFisher) and prepared in 10 ng/µl followed by 10-fold serial dilution into water. One microliter of the sample was used as direct input into RPA or PCR reaction.

Polymerase Chain Reaction

Each PCR was prepared in a total volume of 50 µl containing 1× TaqMan Universal PCR Master Mix, no AmpErase UNG (ThermoFisher), 1 µl plasmid with serial diluted concentrations, and HPV specific primer pairs (Table 2). Each reaction was performed using Bio-Rad CFX96 Real-time system (Bio-Rad). Thermal cycle conditions are as follow: an initial incubation at 95°C for 10 minutes followed by 50 cycles of alternating 95°C for 10 seconds, 50°C for 10 seconds and 72°C for 30 seconds. A further incubation at 72°C for 10 minutes was done to complete the extension step. Five microliters of amplified DNA was used for performing electrophoresis in 2% agarose gel (Promega).

Fluorescent Readout of CRISPR-Cas12a Activity Using FAM-Quencher (FQ) Reporters

LbCas12a was preassembled with an HPV16- or HPV18-targeting CRISPR RNA (crRNA) at 37°C for 30 minutes and diluted in 1× binding buffer (20 mM Tris-HCl, pH 7.5, 100 mM KCl, 5 mM MgCl₂, 1 mM DTT, 5% glycerol, 50 µg/ml heparin) and custom ssDNA-FQ reporter (IDT) (Table 2) [5]. Differently diluted RPA reactions were transferred to a 384-well microplate (Corning). Different concentrations of LbCas12a, crRNA, and custom

Table 1. Characteristics of Cervical Cancer Patients and Cancer-Free Subjects

Cervical Cancer Cases (n = 15)	Controls (n = 14)	
Age	38.28 (SD 9.35)	43.34 (SD 9.98)
Sex	Female	Female
Race	All are Caucasian	All are Caucasian
Stage		
Stage 0 (carcinoma in situ)	5	
Stage IA1	1	
Stage IB1	4	
Stage IB2	3	
Stage IIIA	1	
Stage IIIB	1	
Histological type	All are squamous cell carcinoma	

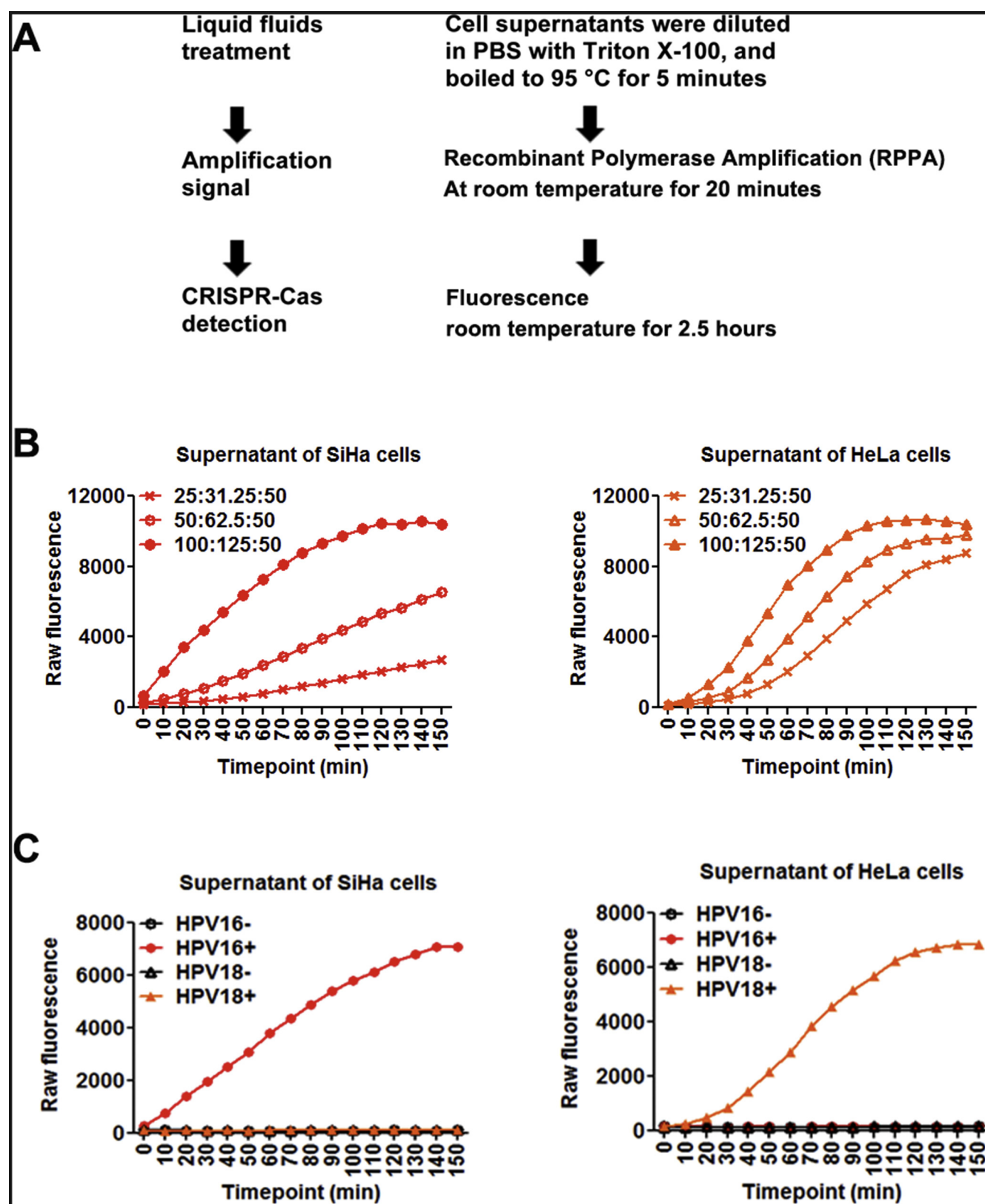


Figure 1. Specific distinguishing HPV16 and HPV18 DNA in liquid fluids with the CRISPR-Cas12a system.(A) Schematic of HPV detection by using fluorescence plate reader from liquid fluids without extraction of DNA. (B) Time courses of fluorescence detection with different concentrations (nM) of LbCas12a:crRNA complexes and fixed 50 nM ssDNA-FQ reporter to target HPV16 in supernatant of SiHa cells (left) and target HPV18 in supernatant of HeLa cells (right). (C) Detection of HPV16 (red circle) or HPV18 (orange triangle) in supernatant of SiHa cells containing HPV16 genomic fragment (left) and supernatant of HeLa cells containing HPV18 genomic fragment. + or -: incubating with or without LbCas12a-crRNA that targets either HPV16 or HPV18.

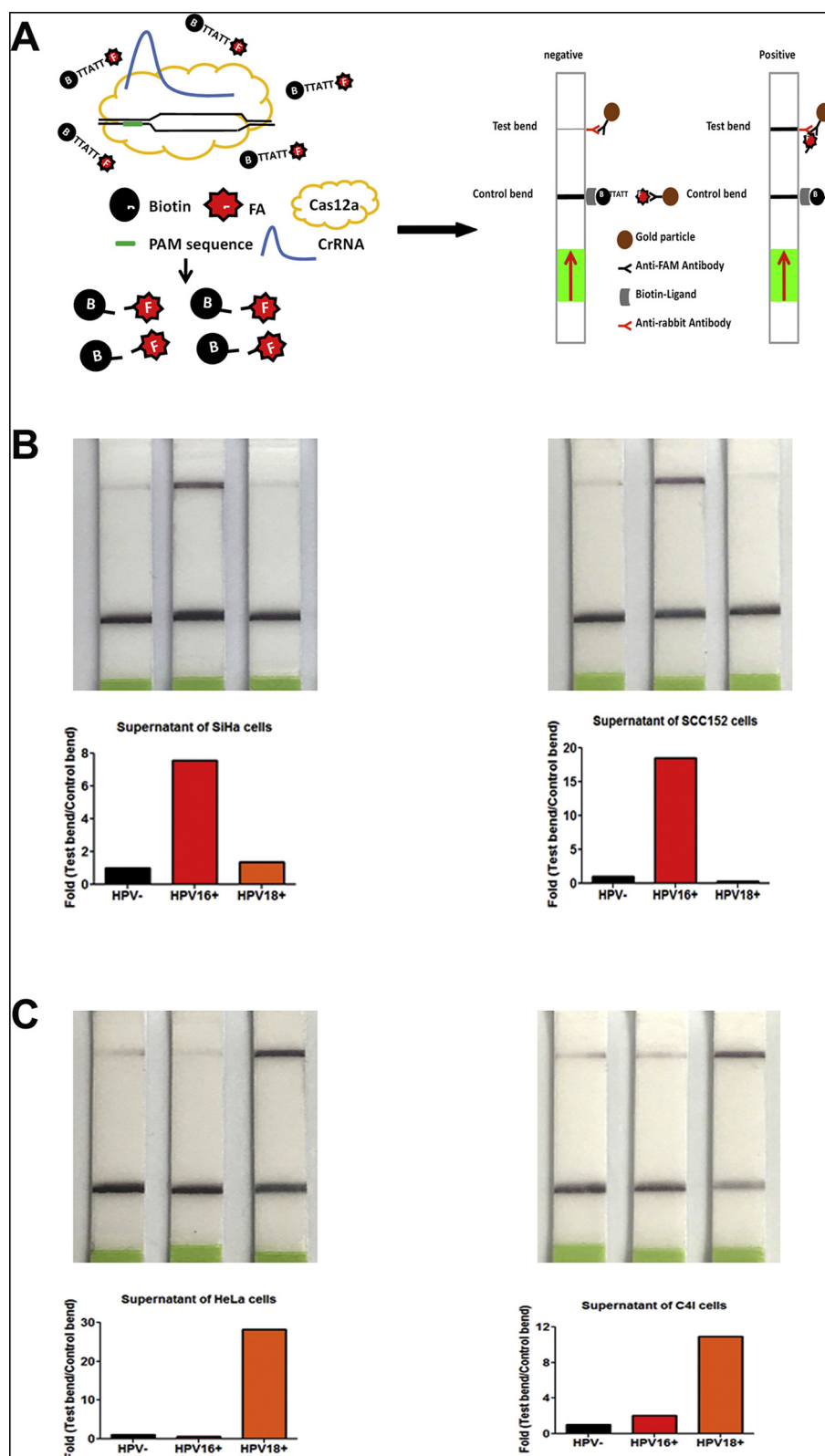


Figure 2. Applying the CRISPR-Cas12a system with paper strips to rapidly and directly visualize the result with naked eyes.(A) Schematic of CRISPR-Cas12a system with lateral flow technology by paper strips to read out the negative or positive results. (B) Lateral flow detection of HPV16 in supernatants of SiHa cells (left) and SCC152 cells (right). (C) Lateral flow detection of HPV18 in supernatants of HeLa cells (left) and C41 cells (right). Quantification of band intensity from detection is shown below. Black bar: incubating without LbCas12a-crRNA as control. Red/orange bar: incubating with LbCas12a-crRNA that targeted HPV16 or HPV18, respectively.

ssDNA-FQ reporter were added directly to the reactions in total 20 μ l volume. Reactions were proceeded at 37°C for 2.5 hours on a fluorescence plate reader (Biotek Synergy H1) with fluorescence kinetics measured every 10 minutes.

Lateral Flow Readout of CRISPR-Cas12a Activity Using FAM-Biotin (FB) Reporters

Lateral flow readout was based on cleavage of dual-labeled ssDNA [4]. Preparation was performed with the modification using FAM-biotin labeled ssDNA reporter (IDT) as a substrate but not quenched fluorescent ssDNA reporter. After incubation at 37°C for 3 hours, 80- μ l HybriDetect assay buffer was added to the 20- μ l reaction mixture. Milena HybriDetect1 lateral flow dipstick (TwistDx) was dipped into the reaction mixture and incubated for 5 minutes at room temperature. Paper strip was removed, and results were interpreted immediately by photograph using a smartphone camera. A sample with signal significantly increasing on the second line (test bend) and decreasing on the first line (control line) was considered to have a positive result. A sample without incubation of CRISPR-Cas12a was used as a control and should have intense line at the control line.

Statistical Analysis

We performed statistical analysis using the Statistical Package for the Social Sciences Software version 11.5 (SPSS, Chicago, IL). Graphical presentation was plotted and analyzed by using Prism 4 software (Graphpad Inc., San Diego CA). *P* values less than .05 were considered as statistically significant in all the analyses. Differences in values were evaluated using Student *t* test. We used Bland-Altman plots to analyze the correlation of two different techniques for quantification of HPV DNA.

Results

The CRISPR-Cas12a System Can Directly and Specifically Detect HPV DNA in Liquid Fluids

To evaluate whether CRISPR-Cas12a could directly detect HPV in liquid fluids without DNA isolation, we collected cell-cultured supernatants of three HPV16-positive cell lines (SiHa, Ca Ski, and SCC152) and two HPV18-positive cell lines (HeLa and C-41). The cell supernatants were diluted 1:2 in PBS containing 0.2% Triton X-100 and boiled at 95°C for 5 minutes. Five microliters of boiled liquid fluids was directly applied in RPA. RPA is a regular temperature isothermal amplification method to exponentially amplify DNA amount without requiring an additional temperature step [4]. LbCas12a with its crRNAs and ssDNA FQ reporter was directly added in the RPA reactions without any purification step. The results were determined by using a fluorescence plate reader with fluorescence kinetics (Figure 1A). Several reaction parameters, including the concentrations of LbCas12a, crRNA, ssDNA-FQ reporter, and RPA reactions, were investigated in the cell supernatants. When 1:10 diluted RPA reactions were used, the optimized concentrations of LbCas12a, crRNA, and ssDNA FAM-biotin reporter were 100 nM, 125 nM, and 50 nM, respectively. At the optimum reaction condition, LbCas12a with either HPV16 or 18 crRNA unambiguously and rapidly identified HPV16 or HPV18 in the cell-supernatants of HPV16- or 18-positive cells, respectively (Figure 1, B-C). Therefore, the CRISPR-Cas12a system could be directly applied

to liquid fluids for sensitivity detecting and specifically classifying HPV DNA.

The CRISPR-Cas12a System Can Rapidly and Directly Read the Results on Paper Strips

To determine whether the results of CRISPR-Cas12a could be directly visualized with the naked eye, we used lateral flow dipsticks to read out the results (Figure 2A). The lateral-flow readout is based on the destruction of a FAM-biotin reporter, by which plentiful reporter gathers anti-FAM antibody-gold nanoparticle conjugates at the first line on the strip, hindering binding of the antibody-gold conjugates to protein A on the second line. In our experiments, when CRISPR-Cas12a cleaved ssDNA FAM-biotin reporter, the signal significantly increased on the second line as the test bend and indicated a positive result on the paper strip (Figure 2A). Using the optimized protocol established above, we incubated RPA reactions of the cell supernatants with 100 nM LbCas12a: 125 nM crRNA: 50 nM FAM-biotin reporter at room temperature. CRISPR-Cas12a intensely increased the test bend and therefore clearly identified HPV16 and 18 in the cell supernatants of HPV16-positive (Figure 2B) or 18-positive cells (Figure 2C), respectively. Furthermore, the total turnaround time for combining CRISPR-Cas12a with lateral-flow readout was about 3 hours, implying that the CRISPR-Cas12a system could be used as a rapid and portable test for directly reading HPV DNA.

The CRISPR-Cas12a System Has a Similar Sensitivity for Detecting HPV DNA in Liquid Fluids Compared with PCR-Based Assay

To define analytic performance of the CRISPR-Cas12a system, the plasmids containing the complete genomic DNA of HPV genotype 16 and 18 were serially (10-fold) diluted into water. Each diluted sample was processed with the CRISPR-Cas12a-paper test and PCR analysis. CRISPR-Cas12a and PCR had the same limit of detection of 170.6 copies/ μ l of input or 0.28 fM and 151.8 copies/ μ L or 0.24 fM in quantification of HPV-16 and 18, respectively (Figure 3). Furthermore, the CRISPR-Cas12a-paper test and PCR analyses exhibited excellent linearity between the target input and measured values in a dynamic range of six orders of magnitude of input. The two techniques displayed an estimated slope coefficient that was in the range of 0.89-0.94 and R^2 of 0.90-0.92 for HPV detection, suggesting an excellent correlation between the CRISPR-Cas12a techniques and PCR assays.

The CRISPR-Cas12a System Can Be Used as a POC Diagnostic Test for Detecting HPV DNA in Clinical Plasma Specimens

We obtained plasma of 15 cervical cancer patients and 14 cancer-free subjects (Table 1). The final diagnosis of the cancer cases and control subjects was made by clinical standard procedures, including HPV testing, Pap testing, and HPV/Pap contesting. Clinically, of the 15 cervical cancer patients, 5 are positive for HPV16, 1 for HPV18, and 9 for both HPV16 and 18.

Among the five HPV 16-positive cervical cancer patients, all had the evidence of HPV16 in plasma determined by the CRISPR-Cas12a system. The one HPV18-positive cancer patient also had the evidence of HPV-18 in plasma. Of cervical cancer cases that were both HPV16 and 18 positive, eight had positive HPV16, while 2 had

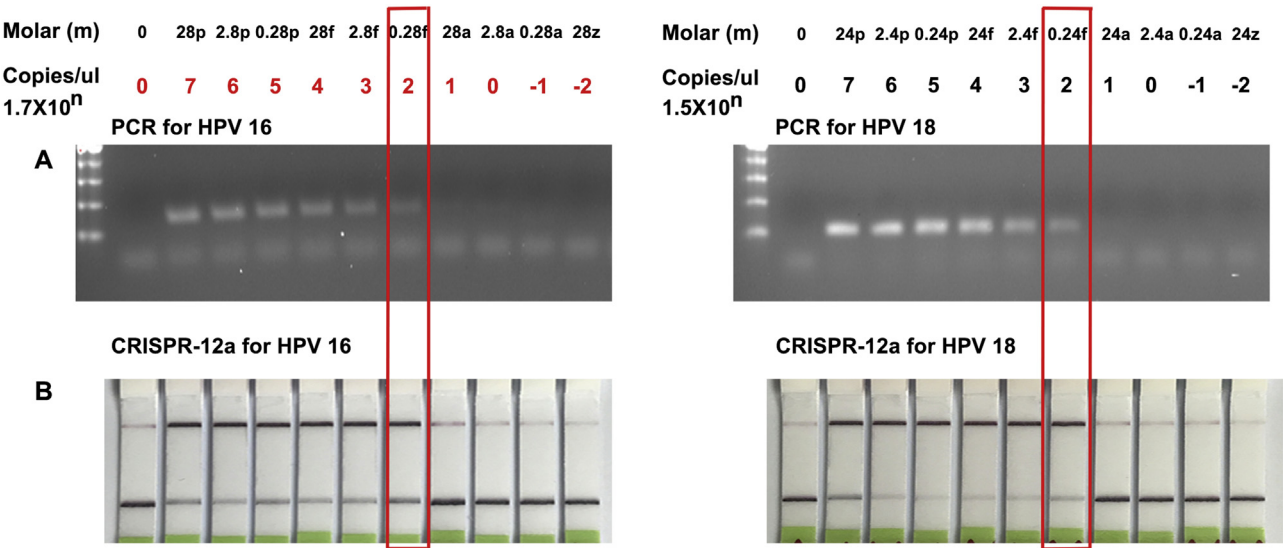


Figure 3. The limit of detection of PCR assay and CRISPR-Cas12a-paper test. Ten-fold serial dilutions of HPV16 plasmid (left) and HPV18 plasmid (right) in nuclease-free water were tested by PCR (A) and CRISPR-Cas12a system with lateral flow detection (B), respectively.

positive HPV 18 results in the plasma samples (Figure 4A). Furthermore, both of the HPV DNA 16 and 18 statuses in plasma had no special relationship with histology and stage of cervical cancer and age of the patients (all $P > .05$). Of the 14 control individuals, 5 had the evidence of HPV16 in plasma (Figure 4B) (Supplementary Figure 1A). Therefore, the CRISPR-Cas12a system could rapidly detect circulating nuclei acid targets in plasma without requiring technical expertise and ancillary machineries.

Discussion

CRISPR-Cas biology has revolutionized the field of molecular diagnostics for infectious and noninfectious diseases, including cancer [7]. Based on RNA-guided Cas13a-mediated collateral cleavage of RNA reporter, Gootenberg et al. recently developed specific high-sensitivity enzymatic reporter UnLOCKing (SHERLOCK) [9]. SHERLOCK could detect multiple synthetic ssRNA targets (RNA1, ZIKV, and DENV) and one synthetic dsDNA target in the same reaction [3,4,9]. However, since Cas13a targets RNA, SHERLOCK requires reverse transcription (RT)-RPA coupled with T7 RNA polymerase to transcribe amplified DNA to RNA for subsequent detection [3,4,9]. The multiple procedures may lead to the associated analytical variations and cost. Furthermore, since RNA is less stable than DNA in human specimens, SHERLOCK might

generate unreliable results, presenting challenges to be used in the clinics.

In contrast to SHERLOCK, the Cas12a-based DETECTR could detect HPV DNA without RT reaction and thus is more practically useful [5]. However, some obstacles exist in the use of DETECTR as a field-deployable POC tool, including requiring complex procedure of sample and DNA preparations and needing expensive fluorescent detection equipment for readout. Furthermore, many pathogens, such as viruses, are shed in body fluids (e.g., plasma, urine, or saliva), which provide noninvasively and easily obtained surrogate materials. A field-deployable, rapid diagnostic test that can directly detect the circulating nucleic acids in the body fluids would be clinically useful. However, the DETECTR system is only applicable to cell-based specimens rather than circulating body fluids.

In this study, we demonstrate that CRISPR-Cas12a can directly target body fluids for sensitivity detecting and specifically classifying circulating HPV DNA without extraction step. Furthermore, we use lateral flow dipstick to rapidly visualize the results of CRISPR-Cas12a without requiring special technical expertise and ancillary equipment. In addition, the CRISPR-Cas12a system has a similar sensitivity for detecting HPV DNA in liquid fluids as does PCR. Moreover, combining CRISPR-Cas12a with a lateral-flow paper strip could be a rapid and portable test for directly reading HPV DNA. Therefore, our further development of CRISPR-Cas12a for detecting HPV DNA in plasma might overcome the obstacles of DETECTR and hence provide a potential field-deployable POC tool.

HPV16 and HPV18 are considered as the highest-risk viruses in the development of cervical cancer, and almost all cervical cancer patients bear HPV DNA in their affected tissues [10]. PCR-based assays have been used for the detection of HPV DNA in tissue samples. Furthermore, the presence of circulating HPV DNA in body fluids can act as a determining factor for identifying the patients at risk for the development and progression of cervical cancer [11,12]. However, PCR is costly, labor intensive, and time consuming, leading to difficulties in the detection of HPV infection in plasma [13,14]. Intriguingly, the CRISPR-Cas12a system could rapidly and

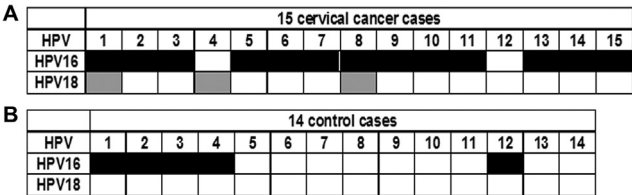


Figure 4. The CRISPR-Cas12a system for detecting HPV DNA in clinical plasma specimens. Using CRISPR-Cas12a-paper to detect HPV DNA in plasma of 15 cervical cancer patients (A) and 14 cancer-free female individuals without any cancer (B). Black cells indicate positive HPV16 results, while gray cells indicate positive HPV18 results in plasma.

cost-effectively detect presence of HPV16 and HPV18 in the plasma of 13 of 14 and 3 of 10 the patients with histopathological diagnosis of cervical cancer, respectively.

HPV16 was also detected in the plasma samples of 5 of 14 women with no sign of malignancy. There are two possible explanations for the observation. First, although these women had no symptoms of cervical cancer, some had signs of inflammation or bacterial infection in their Pap smear test. It would be possible that such conditions facilitate the spread of the existing virus into the blood circulation even in the absence of malignancies. Nevertheless, long-term follow-up of the subjects who have no cervical cancer with positive HPV in plasma is required. Second, CRISPR-Cas12a needs a crRNA as a guide RNA for the activities to genomic targets [15]. The specificity of the CRISPR-Cas12a system is mainly determined by how specific the crRNA sequence is for the genomic target. However, a crRNA targeting sequence might occasionally have additional sites throughout the genome where partial homology exists, causing off-target effect [16]. When used in the detection of a gene of interest, the off-target effect of the crRNA may cause a false-positive result. To reduce the possible off-target-related false-positive results, we are using new programs [17–19] that particularly take consideration of on-target cutting efficiency to design crRNAs for specifically detecting HPV16 and 18, respectively.

A major limitation of this study is that the size of the cohort (15 cases of cervical cancer and 14 healthy control individuals) is small. With the small sample size, the diagnostic performance of the CRISPR-Cas12a system in the clinical specimens could not be reliably evaluated. We are performing a large and prospective study to determine its diagnostic value for detection of circulating HPV DNA.

Conclusions

The CRISPR-Cas12a system might be a promising POC test for noninvasively monitoring and tracking other nucleic acid-related infectious and noninfectious diseases in body fluids. However, the CRISPR-Cas12a system needs further optimization for improved detecting efficiency.

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.tranon.2019.08.011>.

Declarations

Ethical Approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors.

The Institutional Review Boards of University of Maryland Baltimore.

Competing Interests

The authors declare no conflict of interest.

Authors' Contributions

J. T., Q. L., and F. J. conducted the experiments and participated in study design, coordination, and data interpretation,

and preparing the manuscript. All authors read and approved the final manuscript.

Availability of Data and Materials

Not applicable.

Disclosure of Potential Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgements

This work was supported in part by: National Institutes of Health/ National Cancer Institute (USA) grant CA240556, Department of Defense (USA) grant LC160422, the Geaton and JoAnn DeCesaris Family Foundation (USA), and University of Maryland Marlene & Stewart Greenebaum Comprehensive Cancer Center Pilot Grant (USA) (All to F.J.).

References

- [1] Persing DH (1996). Nucleic acid-based pathogen discovery techniques: potential application to zoonoses. *Mol Diagn* **1**, 243–254.
- [2] Barrangou R (2015). The roles of CRISPR-Cas systems in adaptive immunity and beyond. *Curr Opin Immunol* **32**, 36–41.
- [3] Myhrvold C, Freije CA, Gootenberg JS, Abudayyeh OO, Metsky HC, Durbin AF, Kellner MJ, Tan AL, Paul LM, and Parham LA, et al (2018). Field-deployable viral diagnostics using CRISPR-Cas13. *Science* **360**, 444–448.
- [4] Gootenberg JS, Abudayyeh OO, Kellner MJ, Joung J, Collins JJ, and Zhang F (2018). Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a, and Csm6. *Science* **360**, 439–444.
- [5] Chen JS, Ma E, Harrington LB, Da Costa M, Tian X, Palefsky JM, and Doudna JA (2018). CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science* **360**, 436–439.
- [6] Abudayyeh OO, Gootenberg JS, Essletzbichler P, Han S, Joung J, Belanto JJ, Verdine V, Cox DBT, Kellner MJ, and Regev A, et al (2017). RNA targeting with CRISPR-Cas13. *Nature* **550**, 280–284.
- [7] Chertov DS (2018). Next-generation diagnostics with CRISPR. *Science* **360**(6387), 381–382.
- [8] Ma J, Li N, Guarnera M, and Jiang F (2013). Quantification of plasma miRNAs by digital PCR for cancer diagnosis. *Biomark Insights* **8**, 127–136.
- [9] Gootenberg JS, Abudayyeh OO, Lee JW, Essletzbichler P, Dy AJ, Joung J, Verdine V, Donghia N, Daringer NM, and Freije CA, et al (2017). Nucleic acid detection with CRISPR-Cas13a/C2c2. *Science* **356**, 438–442.
- [10] Ferlay J, Colombet M, Soerjomataram I, Mathers C, Parkin DM, Pineros M, Znaor A, and Bray F (2019). Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int J Cancer* **144**, 1941–1953.
- [11] Dong SM, Pai SI, Rha SH, Hildesheim A, Kurman RJ, Schwartz PE, Mortel R, McGowan L, Greenberg MD, and Barnes WA, et al (2002). Detection and quantitation of human papillomavirus DNA in the plasma of patients with cervical carcinoma. *Cancer Epidemiol Biomarkers Prev* **11**, 3–6.
- [12] Ho CM, Yang SS, Chien TY, Huang SH, Jeng CJ, and Chang SF (2005). Detection and quantitation of human papillomavirus type 16, 18 and 52 DNA in the peripheral blood of cervical cancer patients. *Gynecol Oncol* **99**, 615–621.
- [13] Gnanamony M, Peedicayil A, Subhashini J, Ram TS, Rajasekar A, Gravitt P, and Abraham P (2010). Detection and quantitation of HPV 16 and 18 in plasma of Indian women with cervical cancer. *Gynecol Oncol* **116**, 447–451.
- [14] Jaberipour M, Samsami A, Sahraian F, Kazerooni T, Hashemi M, Ghaderi A, and Habibagahi M (2011). Elevation of HPV-18 and HPV-16 DNA in the plasma of patients with advanced cervical cancer. *Asian Pac J Cancer Prev* **12**, 163–167.
- [15] Lee K, Zhang Y, Kleinstiver BP, Guo JA, Aryee MJ, Miller J, Malzahn A, Zarecor S, Lawrence-Dill CJ, and Joung JK (2019). Activities and specificities of CRISPR/Cas9 and Cas12a nucleases for targeted mutagenesis in maize. *Plant Biotechnol J* **17**, 362–372.

-
- [16] Chen X, Chen Y, Huang HM, Li HD, Bu FT, Pan XY, Yang Y, Li WX, Li XF, and Huang C (2019). SUN2: a potential therapeutic target in cancer. *Oncol Lett* **17**, 1401–1408.
- [17] Wang M, Tang C, and Chen J (2018). Drug-target interaction prediction via dual Laplacian graph regularized matrix completion. *Biomed Res Int* **18**, 560–568.
- [18] Chen Y, Zhao B, Liu M, Wang J, Qiu X, Zhu C, and Wu X (2018). MicroRNAs profiling identifies miR-125a and its target gene Wnt2 in skins of different haired rabbits. *Front Genet* **9**, 628–632.
- [19] Wang L, Zhang H, Xie Y, Chen H, Ren C, and Chen X (2019). Target-mediated hyperbranched amplification for sensitive detection of human alkyladenine DNA glycosylase from HeLa cells. *Talanta* **194**, 846–851.



Published in final edited form as:

Med Phys. 2018 April ; 45(4): 1537–1549. doi:10.1002/mp.12820.

Radiomics analysis of pulmonary nodules in low-dose CT for early detection of lung cancer

Wookjin Choi,

Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY 10065

Jung Hun Oh,

Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY 10065

Sadegh Riyahi,

Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY 10065

Chia-Ju Liu,

Department of Radiology, Memorial Sloan Kettering Cancer Center, New York, NY 10065

Feng Jiang,

Department of Pathology, University of Maryland School of Medicine, Baltimore, MD 21201

Wengen Chen,

Department of Diagnostic Radiology and Nuclear Medicine, University of Maryland School of Medicine, Baltimore, MD 21201

Charles White,

Department of Diagnostic Radiology and Nuclear Medicine, University of Maryland School of Medicine, Baltimore, MD 21201

Andreas Rimner,

Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York, NY 10065

James G. Mechalakos,

Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY 10065

Joseph O. Deasy, and

Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY 10065

Wei Lu

Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY 10065

Abstract

Purpose—To develop a radiomics prediction model to improve pulmonary nodule (PN) classification in low-dose CT. To compare the model with the American College of Radiology

^aAuthor to whom correspondence should be addressed. luw@mskcc.org; Telephone: (212)-639-3285.

The authors have no relevant conflicts of interest to disclose.

(ACR) Lung CT Screening Reporting and Data System (Lung-RADS) for early detection of lung cancer.

Methods—We examined a set of 72 PNs (31 benign and 41 malignant) from the Lung Image Database Consortium image collection (LIDC-IDRI). 103 CT radiomic features were extracted from each PN. Before the model building process, distinctive features were identified using a hierarchical clustering method. We then constructed a prediction model by using a support vector machine (SVM) classifier coupled with a least absolute shrinkage and selection operator (LASSO). A 10-fold cross-validation (CV) was repeated ten times (10×10-fold CV) to evaluate the accuracy of the SVM-LASSO model. Finally, the best model from the 10×10-fold CV was further evaluated using 20×5- and 50×2-fold CVs.

Results—The best SVM-LASSO model consisted of only two features: the bounding box anterior-posterior dimension (BB_AP) and the standard deviation of inverse difference moment (SD_IDM). The BB_AP measured the extension of a PN in the anterior-posterior direction and was highly correlated ($r=0.94$) with the PN size. The SD_IDM was a texture feature that measured the directional variation of the local homogeneity feature IDM. Univariate analysis showed that both features were statistically significant and discriminative ($P=0.00013$ and 0.000038 respectively). PNs with larger BB_AP or smaller SD_IDM were more likely malignant. The 10×10-fold CV of the best SVM model using the two features achieved an accuracy of 84.6% and 0.89 AUC. By comparison, Lung-RADS achieved an accuracy of 72.2% and 0.77 AUC using four features (size, type, calcification, and spiculation). The prediction improvement of SVM-LASSO comparing to Lung-RADS was statistically significant (McNemar's test $P=0.026$). Lung-RADS misclassified 19 cases because it was mainly based on PN size, whereas the SVM-LASSO model correctly classified 10 of these cases by combining a size (BB_AP) feature and a texture (SD_IDM) feature. The performance of the SVM-LASSO model was stable when leaving more patients out with 5- and 2-fold CVs (accuracy 84.1% and 81.6% respectively).

Conclusion—We developed an SVM-LASSO model to predict malignancy of PNs with two CT radiomic features. We demonstrated that the model achieved an accuracy of 84.6%, which was 12.4% higher than Lung-RADS.

Keywords

Lung Cancer; Radiomics; Pulmonary Nodule; CT; SVM

I. Introduction

Lung cancer is the leading cause of cancer death in the world. The National Lung Cancer Screening Trial (NLST) showed a clear survival benefit for screening with a low-dose computed tomography (LDCT) in current and former smokers.¹ The early detection of lung cancer by LDCT can reduce mortality. Recently, the Lung Imaging Reporting and Data System (Lung-RADS) was developed by the American College of Radiology (ACR) to standardize the screening of lung cancer on CT images.^{2,3} However, LDCT dramatically increases the number of indeterminate pulmonary nodules (PNs) and produces a high false-positive diagnostic rate, which leads to overdiagnosis.⁴ Therefore, it is important to develop new approaches to improve accuracy.

Computer-aided detection/diagnosis (CAD; specifically, CAdE for detection and CAdx for diagnosis) systems have been investigated to detect PNs and classify malignant and benign PNs.⁵⁻¹⁰ In the mid-90s, Gurney and Swensen conducted a PN characterization study with an artificial neural network (ANN) and features that were subjectively assessed by radiologists.¹¹ Kawata et al. proposed quantitative surface characterization to classify malignant and benign PNs.¹² McNitt-Gray et al. proposed a pattern classification approach to characterize PNs, which used quantitative features including attenuation (intensity), size, shape, and texture.¹³ Aoyama et al. proposed an automatic scheme which segment PNs using dynamic programming technique and determine malignant PNs using linear discriminant analysis (LDA) classification.¹⁴ Armato et al. developed a serial approach PN classification following automatic PN detection.¹⁵ A rule-based scheme was applied to reduce PN candidates in the detection, and two LDA classifiers were applied for the detection and the classification of PNs. Shah et al. investigated the utility of a CAD system using volumetric and contrast enhancement features.¹⁶ Suzuki et al. developed the massive training ANN to filter out benign appearances in CT image.¹⁷ Way et al. developed an automatic 3D active contour segmentation method and extracted surface features from the segmented PNs.^{18,19} Lee et al. developed ensemble classifications using random subspace method or genetic algorithm feature selection with LDA classifier.²⁰ Han et al. investigated texture feature analysis to differentiate malignant and benign PNs.²¹ The early successes of the CAD systems illustrated that quantitative medical image analysis has the potential to improve the performance of detecting cancer on chest CT.

Recently radiomics studies, which extract a large number of quantitative features from medical images and subsequently perform data mining, have been proposed for various clinic applications.²²⁻²⁶ For instance, radiomics has been studied for the prediction of tumor responses and patient outcomes, resulting in more accurate prediction of local control and overall survival.^{22-24,26-29} Lung cancer screening using radiomics has also been studied.³⁰⁻³³ Hawkins et al. proposed a random forest classifier³⁰ using 23 stable radiomic features.^{34,35} Ma et al. proposed a random forest classifier using 583 radiomic features.³¹ Buty et al. developed a random forest classifier using 4096 appearance features extracted with a pre-trained deep neural network and 400 shape features extracted with spherical harmonics.³² Kumar et al. developed a deep neural network model using 5000 features.³³ Liu et al. proposed a linear classifier based on 24 image traits visually scored by physicians.³⁶

Despite the improved prediction accuracy reported in these radiomics studies, there are limitations including the possibility of overfitting the model to the data and lack of clinical/biological interpretations of the intimidatingly large number of radiomic features. To overcome these limitations, we first identified distinctive radiomic features using hierarchical clustering and then constructed a support vector machine (SVM) model with only two important features chosen by a least absolute shrinkage and selection operator (LASSO). We compared the performance of this model and Lung-RADS on a public database.

II. Methods and Materials

II. A. Dataset and Lung-RADS

The Lung Image Database Consortium image collection (LIDC-IDRI) in The Cancer Imaging Archive (TCIA) contains 1018 cases with low-dose screening thoracic CT scans and marked-up annotated lesions.³⁷ Four experienced thoracic radiologists performed contouring and image annotation. A subset of cases (n=157) has associated diagnostic data regarding the screening CT scans. Of these 157 patients, 36 had benign lesions, 43 had malignant primary lung cancers, and the remainders were unknown or had metastatic tumors. Forty two malignant lung cancer cases and 6 benign cases were diagnosed by biopsy or surgical resection, and 18 benign diseases were determined by stability at 2 year follow-up. Lastly, 4 cases (1 malignant and 3 benign) were determined by lesional progression or response. However, 5 benign cases and 2 malignant cases had missing PN contours. Thus, we evaluated 72 cases (31 benign and 41 malignant) who had both diagnostic data and PN contours. Each evaluated PN had at least one to four contours delineated by the four radiologists. Figure 1 shows distributions of size, type, calcification, and spiculation of PNs in the data set.

The tube peak potential energy used for most scan acquisition was 120 kV (n=71) and only one scan was 140 kV. Tube current ranged from 80 to 570 mA (mean: 322.5 mA). Slice thicknesses were 1.0 mm (n=5), 1.25 mm (n=13), 1.5 mm (n=1), 2.0 mm (n=11), and 2.5 mm (n=42). Reconstruction interval ranged from 0.625 to 2.5 mm (mean: 1.80 mm). The in-plane pixel size ranged from 0.547 to 0.898 mm (mean: 0.721 mm). Each axial slide of CT scan was 512×512 pixels. While the convolution kernels used for image reconstruction differ among manufacturers, these convolution kernels may be classified broadly as “standard/non enhancing” (n=43), “slightly enhancing” (n=17), and “over enhancing” (n=12) (in order of increasing spatial frequencies accentuated by each class). 44 scans were contrast-enhanced CT.

We performed Lung-RADS categorization based on the PN contour and annotations made by the four radiologists in the LIDC-IDRI dataset.² A study radiologist (CL) reviewed the categorization results. As shown in Table 1, the Lung-RADS categorization is mainly based on PN size (the average of the longest and shortest diameters on axial slice) with some consideration to calcification, PN type (solid, part-solid, and non-solid or ground glass nodule/GGN), and additional suspicious features. To match the original LIDC-IDRI diagnosis, categories 3 and lower PNs are deemed as benign and category 4 (4A, 4B, and 4X) PNs as malignant.

II. B. Radiomic Features

Figure 2 shows the flowchart for the extraction of radiomic features and the construction of a prediction model. To extract radiomic features from a CT, we built the following image analysis pipeline using the Insight Segmentation and Registration Toolkit (ITK, National Library of Medicine; Bethesda, MD).³⁸ Firstly, we re-sampled the CT images to make voxels isotropic (1mm). Secondly, we generated a consensus contour for each PN with 2 or more contours by using the simultaneous truth and performance level estimation (STAPLE).

^{39,40} Thirdly, we extracted 103 radiomic features from each PN to quantify its intensity, shape, and texture (spatial variations).^{23,27,41} Finally, we performed a univariate analysis using Wilcoxon rank-sum test and the area under the receiver operating characteristic curve (AUC) to evaluate the significance of each feature. P-values were adjusted using Bonferroni correction because we tested multiple features ($n=103$) for a single outcome.⁴²

Three types of radiomic features were extracted for each PN. Intensity features are first-order statistical measures that quantify the level and distribution of CT attenuations in a PN. Shape features describe geometric characteristics (e.g., volume, diameter, elongation, and flatness) of a PN. CT texture features quantify the spatial patterns of tissue density, such as homogeneity, coarseness, and correlation of CT intensity in a PN by using Gray-level co-occurrence matrix (GLCM)⁴³ and gray-level run-length matrix (GLRM).^{44,45} For texture features, the CT intensity was first normalized to the range of for contrast stretching to simplify the spatial complexity due to a wide dynamic range of CT attenuation. The texture features were then computed on the GLCM and GLRM of the normalized volumes. The average value of each texture feature was computed over all 13 directions to obtain rotationally invariant features. Furthermore, the length of runs was normalized by the diagonal length of the PN's bounding box to make the GLRM scale-invariant.

II. C. Prediction Model

Before constructing a prediction model, we identified distinctive radiomic features using Ward's hierarchical clustering method⁴⁶ which maximized the total within-cluster (Pearson) correlation (r). Each feature started in its cluster. Pairs of clusters were merged if the total within-cluster correlation was larger after merging than before merging, as one moved up the hierarchy. This resulted in a hierarchical feature cluster tree. The tree was then divided into several prominent clusters (feature groups) by cutting using a threshold $r \geq 0.85$. If a feature was the representative feature of a feature group, which had the smallest within-cluster correlation, or if a feature was independent to all other features, it was identified as a distinctive feature. Other features were redundant and were removed from subsequent analysis.

An SVM classifier was then constructed for the prediction of PN malignancy coupled with a LASSO feature selection. All distinctive features were fed to the SVM classifier in a manner of a 10-fold cross-validation (CV). Within each fold CV of the model building process, LASSO was applied to select the ten most important distinctive features by using another (inner loop) 10-fold CV. An SVM classifier was then constructed to predict PN malignancy. A radial basis kernel function was employed in the SVM classifier, with its parameters experimentally chosen: $\gamma=0.001$ and $C=64$. We repeated the outer-loop 10-fold CV ten times to obtain the model accuracy (10 $\times$ 10-fold CV). In each repetition, all patients were randomly partitioned into a training set (90% patients) and a testing set (10% patients). Finally, the stability of the best model from the 10 $\times$ 10-fold CV was evaluated using 20 $\times$ 5- and 50 $\times$ 2-fold CVs, in which 20% and 50% patients were partitioned into the testing set, respectively.²⁸ McNemar's test was used to compare prediction performance between the proposed radiomics model and Lung-RADS.

III. Results

The hierarchical clustering identified 44 distinctive features from the total of 103 radiomic features. Among the 44 distinctive features, 14 were significant in univariate analysis ($P < 0.05$ after Bonferroni correction). These included 4 shape, 8 texture, and 2 shape + intensity (intensity weighted shape features using image moments) features (Supplemental Table 1).

Figure 3 shows that the model achieved the highest accuracy when two features were selected into the SVM classifier. The two most frequently selected features were: (1) the PN bounding box (BB) anterior-posterior (AP) dimension (BB_AP), and (2) the standard deviation (SD) of inverse difference moment (SD_IDM, a texture feature that measures the directional variation of the local homogeneity feature IDM⁴³, see IV.A). BB_AP and SD_IDM were selected as the first feature 57 times and 43 times, respectively. They were selected as the second feature 43 times and 57 times, respectively. Therefore, the two features were always selected into the best SVM classifier. The best single-feature model achieved $75.1 \pm 6.0\%$ accuracy (0.75 ± 0.04 AUC). When adding the second feature in the model, the prediction accuracy was improved to $84.6 \pm 1.5\%$ (0.89 ± 0.01 AUC). However, the performance was worse when adding more than two features as shown in Figure 3.

Figure 4 shows the difference between benign and malignant PNs for the two features and the PN size. As expected, the BB_AP was highly correlated with the PN size ($r = 0.94$), and the larger the BB_AP, the more likely a PN was malignant. The AP or left-right (LR) dimensions might be a better predictor than the superior-inferior (SI) dimension because of the typically higher axial resolutions (≤ 1 mm) than the longitudinal resolution (2-3 mm) in CT. The SD_IDM is a texture feature that measures the directional variation of the local homogeneity feature IDM. The smaller the SD_IDM, the more likely a PN was malignant.

Table 2 compares the prediction performance of Lung-RADS and the SVM-LASSO model in 10×10-fold CV. Figure 5 shows their ROC curves. The ROC curve for Lung-RADS was generated by using each category as a cutoff for malignancy classification. The ROC curve for the SVM-LASSO model was generated by computing the probability of nodule malignancy using the improved Platt's method.⁴⁷ The Lung-RADS achieved an accuracy of 72.2% with four features (size, type, calcification, and suspicious features or image findings). The SVM-LASSO model achieved an accuracy of 84.6% with two features (BB_AP and SD_IDM), which represented a 12.4% improvement over the Lung-RADS. The performance difference was statistically significant ($P = 0.026$).

Figure 6 shows a scatter plot of the two features and the classification curve by the SVM-LASSO model. All five malignant PNs in the elliptical region **a** (red) were misclassified as benign by Lung-RADS since they were small (size < 8 mm). Two example cases (c) and (d) were shown in Figure 7. Note that case (c) was part-solid with a solid component of 4 mm, and was thus classified as Category 3. On the other hand, all seven benign PNs in the elliptical region **b** (green) were misclassified as malignant by Lung-RADS since they were large (size ≥ 8 mm). Two example cases (a) and (b) were shown in Figure 8. The SVM-LASSO model correctly classified these 12 PNs in both regions **a** and **b** by combining a size

feature (BB_AP) and a texture feature (SD_IDM). There was one malignant case that was correctly classified by Lung-RADS but misclassified by the SVM-LASSO model. This case was indicated by arrow **c** in Figure 6 and shown as the case (e) in Figure 7. This PN was correctly classified by Lung-RADS as malignant (category 4B) based on its size. Finally, nine cases were misclassified by both Lung-RADS and the SVM-LASSO model. Overall, the SVM-LASSO model showed a clear advantage over Lung-RADS. Figure 8 shows Lung-RADS categorization on the scatter plots for solid and part-solid PNs, respectively.

The performance of the SVM-LASSO model was stable when more patients were partitioned into the testing set with 5- and 2-fold CVs compared to 10-fold CV (Table 3). Only a small reduction in each accuracy measurement was observed, and even the accuracy of 2-fold CV (50% patients in the training set and 50% in the testing set) was 6.3% higher than that of the Lung-RADS.

IV. Discussion

We demonstrated that the SVM-LASSO model achieved 84.6% accuracy, which was 11.0% higher than that of the Lung-RADS. Accurate diagnosis of malignant PNs on LDCT screening is critical because LDCT dramatically increases the number of indeterminate PNs, leading to overdiagnosis.⁴ This model has the potential to spare individuals with benign growths from the biopsies and 2-year multiple follow-up examinations while allowing effective treatments to be immediately initiated for lung cancer.

IV. A. SD_IDM feature

IDM is a texture feature that measures the local homogeneity.⁴³ The SD_IDM measures the directional variation of the IDM in 13 directions. As shown in Figure 9 and Figure 10, benign PNs appeared to be more homogeneous with generally higher IDMs in each direction as well as in the Mean_IDM (average of IDMs in all 13 directions) than malignant PNs of similar size. More importantly, benign PNs tended to be more homogeneous with much higher IDMs in AP and/or SI directions than in the other 12 or 11 directions. This led to a larger directional variation of the IDM or higher SD_IDM for benign PNs. The SD_IDM was even higher (≥ 0.041) for all four cases with diffuse solid calcification (not shown), which is a typical pattern in benign PNs (for example, in PN with granulomatous inflammation). For the studied cohort, SD_IDM showed a highly significant difference between malignant and benign PNs ($P=0.000038$, Figure 4) and provided complementary information to size (Figure 6 and 7). These observations supported using SD_IDM in addition to size in the classification of PN malignancy.

IV. B. Comparison with CADx systems

Table 4 shows the comparisons with CADx systems for lung cancer screening. The proposed method showed comparable or better accuracy than others. For both comparisons with CADx systems (Table 4) and with radiomics models (Table 5), when not specified ground-truth was obtained by biopsy, resection, or 2 year follow-up. Also it should be noted that each method was evaluated on different datasets or with different validation methods. Most methods reported only AUC.

Aoyama et al.¹⁴ proposed segmentation-based scheme which achieved 0.828 AUC (single slice) and 0.846 AUC (multiple slice). The performance (AUC 0.882) of Suzuki's scheme¹⁷ based on the multiple massive training ANNs was greater than that of Aoyama's scheme¹⁴ for the same database. Shah et al.¹⁶ applied contrast enhancement features extracted from pairs of CTs without contrast and with intravenous contrast injection. They achieved AUCs from 0.69 to 0.92 in a leave-one-out CV. Way et al.¹⁹ developed novel surface features, and the AUC was improved from 0.821 to 0.857 when the surface features were added to morphological and texture features. Han et al.²¹ achieved the highest AUC of 0.894, but it was the only study that ground-truth was rated by radiologist's assessment. LUNGx Challenge reported the performance of 11 CAD systems (0.50 - 0.68 AUC) and six radiologists (0.70 - 0.85 AUC) for diagnosis of PNs on CT scans.¹⁰ Only three CAD methods performed statistically better than random guessing. Three radiologists performed statistically better than the best CAD system (0.68 AUC), which was based on SVM model.

IV. C. Comparison with recently reported radiomics models

Table 5 shows the comparisons with recently reported radiomics models for lung cancer screening. The proposed method showed comparable or better accuracy than others. Some accuracy measurements were not reported in all studies.

Hawkins et al. proposed a random forest classifier using 23 stable (high reproducibility – concordance correlation coefficients ≥ 0.95 in test-retest) radiomic features.³⁰ Ma et al. proposed a random forest classifier using 583 radiomic features.³¹ Its performance on the same LIDC dataset as used in the present study was comparable to the proposed SVM-LASSO model. However, the number of features used was more than eight times of the number of patients, which may cause a model overfitting problem.

Both Buty et al. and Kumar et al. applied deep learning techniques to predict malignancy of PNs.^{32,33} Buty et al. extracted 4096 appearance features using the pre-trained deep neural network (AlexNet⁴⁸) and 400 shape features from spherical harmonics. They fed these features into a random forest classifier, which achieved an accuracy of 82.4%.³² Since the neural network was pre-trained using general color images, it is questionable that it can capture the salient features of a PN in the LDCT images. Kumar et al. used a deep neural network for both feature extraction and malignancy classification. They extracted a total of 5000 radiomic features and achieved an accuracy of 77.5%.³³ Deep learning is a rapidly emerging technology, but it needs large training dataset to avoid model overfitting because an intimidatingly large number of nodes and features are used. Liu et al. applied 24 image traits scored by radiologists to predict malignancy of PNs.³⁶ They achieved an accuracy of 80.0% and 0.81 AUC using four image traits that characterized PN size, contour/margin, concavity, and PNs in nontumor lobes. However, the image traits are semi-quantitative with inter- and intra-observer variation depending on radiologists' training and preferences.

Compared to the above methods, the main advantage of our method was that we used only two important features in the SVM-LASSO model while achieving comparable or better accuracy.

IV. D. Repeatability Analysis

Using the group consensus contour from four manual segmentations is a good way to reduce inter-observer variability, however, it is not representative of clinical practice where an automatic or semi-automatic segmentation method with manual correction would be most likely used. In this study, we used the manual segmentations readily available and conducted a repeatability analysis across the four individual contours and the group consensus contours. First, we evaluated the inter-observer agreement by comparing the individual contours and the consensus contour (Supplemental Table 2). All the STAPLE-estimated sensitivity, specificity, and Jaccard index were greater than 80% except the Jaccard index (77.7%) of R4. For feature agreement evaluation, we examined Bland-Altman plots⁴⁹ for BB_AP and SD_IDM (Supplemental Figure 1) and calculated the intra-class correlation coefficient (ICC)⁵⁰ between the consensus and individual contours. Supplemental Figure 1 shows that the mean differences were close to zero, indicating that there was no systematic error, there was no trend in the plots, and the 95% limits of agreement were small. The ICCs were 0.95 for BB_AP and 0.78 for SD_IDM (Supplemental Table 3), showing good agreement.⁵⁰ Lastly, Supplemental Table 4 shows agreements in malignancy predictions between the consensus contours and individual contours. The predictions based on individual contours (80.6~83.6%) were slightly less accurate than the prediction based on the consensus contours (84.6%) but were not significantly different (all $P > 0.05$). Overall, the proposed method showed consistent results across different contours.

IV. E. Limitations

Limitations of the present study include that the model was developed from a moderate-size cohort of 72 patients, and there were no GGN cases in this cohort. Although 10-, 5- and 2-fold CV showed that the model was not notably affected by overfitting, the performance of the model should be validated in a larger, independent patient cohort. Also, there were no radiomic features that specifically characterized lobulated or spiculated margins.

IV. F. Future work

In this study, we applied a feature discovery approach, which extracted a large number of image features (>100) first and then selected the most valuable ones that are independent, robust, and prominent in the data.²⁴ We plan to add a candidate feature approach, in which only a few important features are selected based on prior knowledge of their physiological, biochemical or functional associations with the disease and therapy.²⁴ For example, quantification of lobulated or spiculated margins is a good candidate feature because it is known that a PN with smooth and well-defined margins is more likely benign, while a PN with lobulated or spiculated margins is more likely malignant.⁵¹ A more accurate PN segmentation method is required to delineate lobulated or spiculated margins, and advanced feature extraction methods are needed to characterize such margins. Other potential candidate features include calcification, attachment, solidity, and cavitation of a PN.

It was difficult to diagnose small PNs (diameter=6-15mm and the probability of malignancy=1-15%) based on radiomics features only. Our model achieved an accuracy of only 50% for PNs smaller than 15 mm. We showed that when combining plasma biomarkers with clinical variables and image features, the prediction was more accurate (AUC = 0.91

in⁵² and 0.95 in⁵³). These studies suggested that the biomarkers, clinical variables, and image features have complementary information. Therefore, we plan to integrate all these parameters in the SVM-LASSO model and expect further improvement in the prediction accuracy, particularly for small PNs.

V. Conclusion

We developed an SVM-LASSO model to predict malignancy of PNs with two CT radiomic features (the bounding box anterior-posterior dimension and the directional variation of local homogeneity). We demonstrated that the model achieved an accuracy of 84.6%, which was 12.4% higher than that for Lung-RADS.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by the NIH/NCI Grant No. R01 CA172638 and the NIH/NCI Cancer Center Support Grant P30 CA008748.

References

1. Aberle DR, DeMello S, Berg CD, et al. Results of the two incidence screenings in the National Lung Screening Trial. *The New England journal of medicine*. 2013; 369(10):920–931. [PubMed: 24004119]
2. McKee BJ, Regis SM, McKee AB, Flacke S, Wald C. Performance of ACR Lung-RADS in a Clinical CT Lung Screening Program. *J Am Coll Radiol*. 2015; 12(3):273–276. [PubMed: 25176499]
3. Pinsky PF, Gierada DS, Black W, et al. Performance of Lung-RADS in the National Lung Screening Trial: a retrospective assessment. *Annals of internal medicine*. 2015; 162(7):485–491. [PubMed: 25664444]
4. Patz EF Jr, Pinsky P, Gatsonis C, et al. Overdiagnosis in low-dose computed tomography screening for lung cancer. *JAMA Intern Med*. 2014; 174(2):269–274. [PubMed: 24322569]
5. Armato SG, Giger ML, Moran CJ, Blackburn JT, Doi K, MacMahon H. Computerized detection of pulmonary nodules on CT scans. *Radiographics*. 1999; 19(5):1303–1311. [PubMed: 10489181]
6. Doi K. Current status and future potential of computer-aided diagnosis in medical imaging. *The British Journal of Radiology*. 2005; 78(Suppl_1):s3–s19. [PubMed: 15917443]
7. Suzuki K. A review of computer-aided diagnosis in thoracic and colonic imaging. *Quantitative Imaging in Medicine and Surgery*. 2012; 2(3):163–176. [PubMed: 23256078]
8. El-Baz A, Beache GM, Gimel'farb G, et al. Computer-Aided Diagnosis Systems for Lung Cancer: Challenges and Methodologies. *Int J Biomed Imaging*. 2013 Article ID 942353: 46 pages.
9. Choi W, Choi T-S. Automated pulmonary nodule detection based on three-dimensional shape-based feature descriptor. *Comput Methods Programs Biomed*. 2014; 113(1):37–54. [PubMed: 24148147]
10. Armato SG III, Drukker K, Li F, et al. LUNGx Challenge for computerized lung nodule classification. *Journal of Medical Imaging*. 2016; 3(4):044506. [PubMed: 28018939]
11. Gurney JW, Swensen SJ. Solitary pulmonary nodules: determining the likelihood of malignancy with neural network analysis. *Radiology*. 1995; 196(3):823–829. [PubMed: 7644650]
12. Kawata Y, Niki N, Ohmatsu H, et al. Quantitative surface characterization of pulmonary nodules based on thin-section CT images. *IEEE Transactions on nuclear science*. 1998; 45(4):2132–2138.

13. McNitt-Gray MF, Hart EM, Wyckoff N, Sayre JW, Goldin JG, Aberle DR. A pattern classification approach to characterizing solitary pulmonary nodules imaged on high resolution CT: preliminary results. *Medical physics*. 1999; 26(6):880–888. [PubMed: 10436888]
14. Aoyama M, Li Q, Katsuragawa S, Li F, Sone S, Doi K. Computerized scheme for determination of the likelihood measure of malignancy for pulmonary nodules on low-dose CT images. *Medical Physics*. 2003; 30(3):387–394. [PubMed: 12674239]
15. Armato SG, Altman MB, Wilkie J, et al. Automated lung nodule classification following automated nodule detection on CT: A serial approach. *Medical Physics*. 2003; 30(6):1188–1197. [PubMed: 12852543]
16. Shah SK, McNitt-Gray MF, Rogers SR, et al. Computer aided characterization of the solitary pulmonary nodule using volumetric and contrast enhancement features. *Academic radiology*. 2005; 12(10):1310–1319. [PubMed: 16179208]
17. Suzuki K, Li F, Sone S, Doi K. Computer-aided diagnostic scheme for distinction between benign and malignant nodules in thoracic low-dose CT by use of massive training artificial neural network. *IEEE Trans Med Imaging*. 2005; 24(9):1138–1150. [PubMed: 16156352]
18. Way TW, Hadjiiski LM, Sahiner B, et al. Computer-aided diagnosis of pulmonary nodules on CT scans: segmentation and classification using 3D active contours. *Medical physics*. 2006; 33(7):2323–2337. [PubMed: 16898434]
19. Way TW, Sahiner B, Chan HP, et al. Computer-aided diagnosis of pulmonary nodules on CT scans: improvement of classification performance with nodule surface features. *Medical physics*. 2009; 36(7):3086–3098. [PubMed: 19673208]
20. Lee MC, Boroczky L, Sungur-Stasik K, et al. Computer-aided diagnosis of pulmonary nodules using a two-step approach for feature selection and classifier ensemble construction. *Artificial Intelligence in Medicine*. 2010; 50(1):43–53. [PubMed: 20570118]
21. Han F, Wang H, Zhang G, et al. Texture feature analysis for computer-aided diagnosis on pulmonary nodules. *J Digit Imaging*. 2015; 28(1):99–115. [PubMed: 25117512]
22. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. 2012; 48(4):441–446. [PubMed: 22257792]
23. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014; 5:4006. [PubMed: 24892406]
24. Lu W, Chen W. Positron emission tomography/computerized tomography for tumor response assessment—a review of clinical practices and radiomics studies. *Translational Cancer Research*. 2016; 5(4):364–370. [PubMed: 27904837]
25. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology*. 2016; 278(2):563–577. [PubMed: 26579733]
26. Scrivener M, de Jong EEC, van Timmeren JE, Pieters T, Ghaye B, Geets X. Radiomics applied to lung cancer: a review. *Translational Cancer Research*. 2016; 5(4):398–409.
27. Tan S, Kligerman S, Chen W, et al. Spatial-Temporal [(18)F]FDG-PET Features for Predicting Pathologic Response of Esophageal Cancer to Neoadjuvant Chemoradiation Therapy. *International journal of radiation oncology, biology, physics*. 2013; 85(5):1375–1382.
28. Zhang H, Tan S, Chen W, et al. Modeling Pathologic Response of Esophageal Cancer to Chemoradiation Therapy Using Spatial-Temporal (18)F-FDG PET Features, Clinical Parameters, and Demographics. *Int J Radiat Oncol Biol Phys*. 2014; 88(1):195–203. [PubMed: 24189128]
29. Lu W, Tan S, Chen W, et al. Pre-Chemoradiotherapy FDG PET/CT cannot Identify Residual Metabolically-Active Volumes within Individual Esophageal Tumors. *J Nucl Med Radiat Ther*. 2015; 6(3):226. [PubMed: 26594591]
30. Hawkins S, Wang H, Liu Y, et al. Predicting malignant nodules from screening CTs. *J Thorac Oncol*. 2016
31. Ma, J., Wang, Q., Ren, Y., Hu, H., Zhao, J. Automatic lung nodule classification with radiomics approach. Paper presented at: SPIE Medical Imaging 2016: PACS and Imaging Informatics: Next Generation and Innovations; San Diego, California, United States. 2016.

32. Buty, M., Xu, Z., Gao, M., Bagci, U., Wu, A., Mollura, DJ. Characterization of Lung Nodule Malignancy Using Hybrid Shape and Appearance Features. In: Ourselin, S.Joskowicz, L.Sabuncu, MR.Unal, G., Wells, W., editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016: 19th International Conference; Athens, Greece. October 17-21, 2016; Proceedings, Part I. Vol 9900. Cham: Springer International Publishing; 2016:662-670
33. Kumar D, Shafiee MJ, Chung AG, Khalvati F, Haider MA, Wong A. Discovery Radiomics for Computed Tomography Cancer Detection. arXiv preprint. 2015:arXiv. 1509.00117.
34. Balagurunathan Y, Kumar V, Gu Y, et al. Test-retest reproducibility analysis of lung CT image features. Journal of digital imaging. 2014; 27(6):805–823. [PubMed: 24990346]
35. Balagurunathan Y, Gu Y, Wang H, et al. Reproducibility and Prognosis of Quantitative Features Extracted from CT Images. Translational Oncology. 2014; 7(1):72–87. [PubMed: 24772210]
36. Liu Y, Balagurunathan Y, Atwater T, et al. Radiological Image traits Predictive of Cancer Status in Pulmonary Nodules. Clinical cancer research : an official journal of the American Association for Cancer Research. 2016
37. Armato SG III, McLennan G, Bidaut L, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys. 2011; 38(2):915–931. [PubMed: 21452728]
38. Ibanez, L., Schroeder, W., Ng, L., Cates, J. The ITK Software Guide. Kitware; 2005.
39. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE transactions on medical imaging. 2004; 23(7):903–921. [PubMed: 15250643]
40. Choi W, Xue M, Lane BF, et al. Individually optimized contrast-enhanced 4D-CT for radiotherapy simulation in pancreatic ductal adenocarcinoma. Medical Physics. 2016; 43(10):5659–5666. [PubMed: 27782710]
41. Choi W, Choi T-S. Genetic programming-based feature transform and classification for the automatic detection of pulmonary nodules on computed tomography images. Information Sciences. 2012; 212:57–78.
42. Pagano, M., Gauvreau, K. Principles of Biostatistics. Second. Pacific Grove, CA: Duxbury; 2000.
43. Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics. 1973; 3(6):610–621.
44. Galloway MM. Texture analysis using gray level run lengths. Computer graphics and image processing. 1975; 4(2):172–179.
45. Tang X. Texture information in run-length matrices. IEEE transactions on image processing : a publication of the IEEE Signal Processing Society. 1998; 7(11):1602–1609. [PubMed: 18276225]
46. Ward JH Jr. Hierarchical Grouping to Optimize an Objective Function. 2012
47. Lin H-T, Lin C-J, Weng RC. A note on Platt's probabilistic outputs for support vector machines. Machine Learning. 2007; 68(3):267–276.
48. Krizhevsky, A., Sutskever, I., Hinton, GE. ImageNet Classification with Deep Convolutional Neural Networks. Paper presented at: Advances in Neural Information Processing Systems; 2017.
49. Bland JM, Altman DG. Agreement between methods of measurement with multiple observations per individual. J Biopharm Stat. 2007; 17(4):571–582. [PubMed: 17613642]
50. Parmar C, Rios Velazquez E, Leijenaar R, et al. Robust Radiomics feature quantification using semiautomatic volumetric segmentation. PloS one. 2014; 9(7):e102107. [PubMed: 25025374]
51. Erasmus JJ, Connolly JE, McAdams HP, Roggli VL. Solitary pulmonary nodules: Part I. Morphologic evaluation for differentiation of benign and malignant lesions. Radiographics. 2000; 20(1):43–58. [PubMed: 10682770]
52. Ma J, Guarnera MA, Zhou W, Fang H, Jiang F. A Prediction Model Based on Biomarkers and Clinical Characteristics for Detection of Lung Cancer in Pulmonary Nodules. Transl Oncol. 2016; 10(1):40–45. [PubMed: 27889655]
53. Lin Y, Leng Q, Jiang Z, et al. A classifier integrating plasma biomarkers and radiological characteristics for distinguishing malignant from benign pulmonary nodules. Int J Cancer. 2017

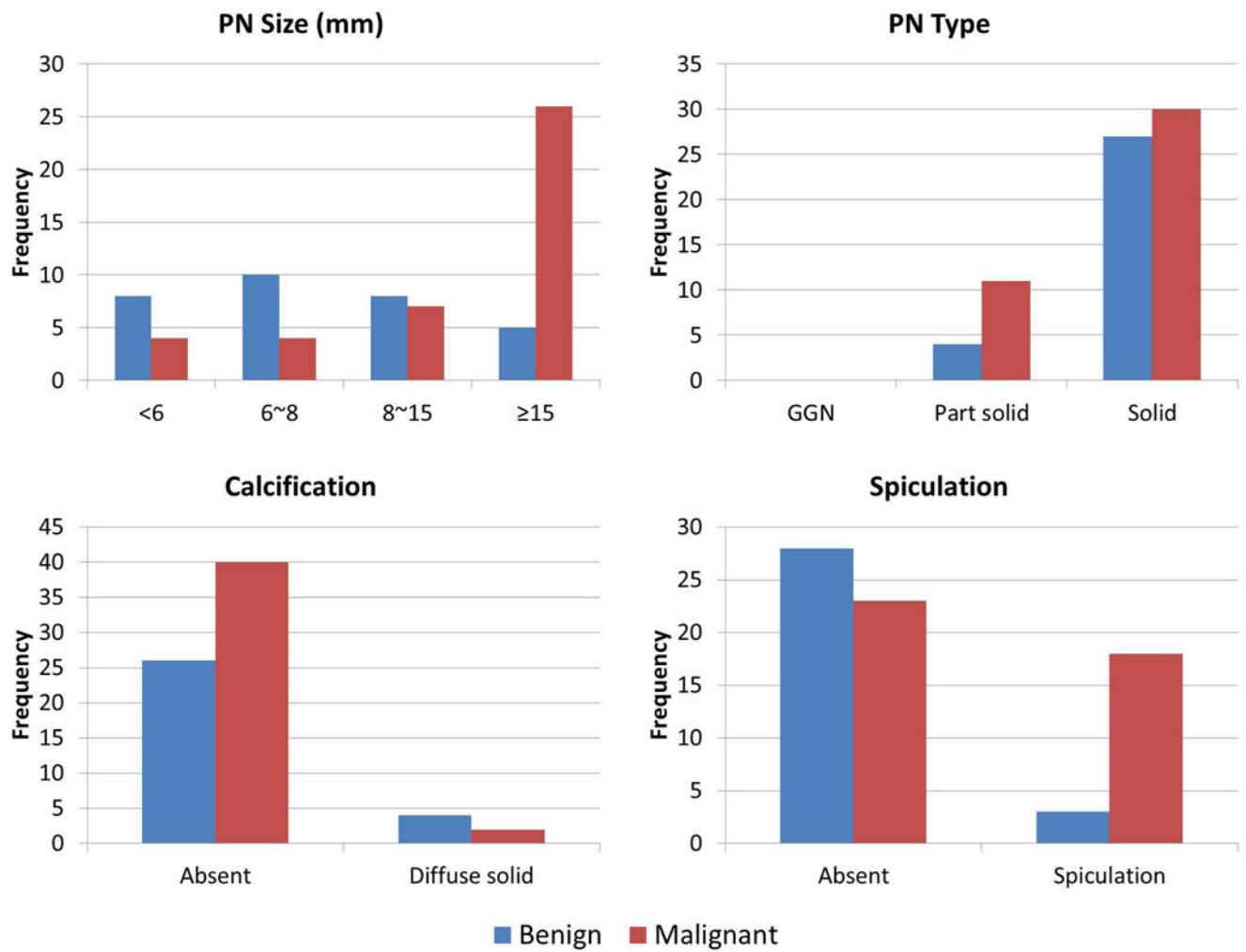


Figure 1.
Distributions of pulmonary nodule size, type, calcification, and spiculation in the dataset

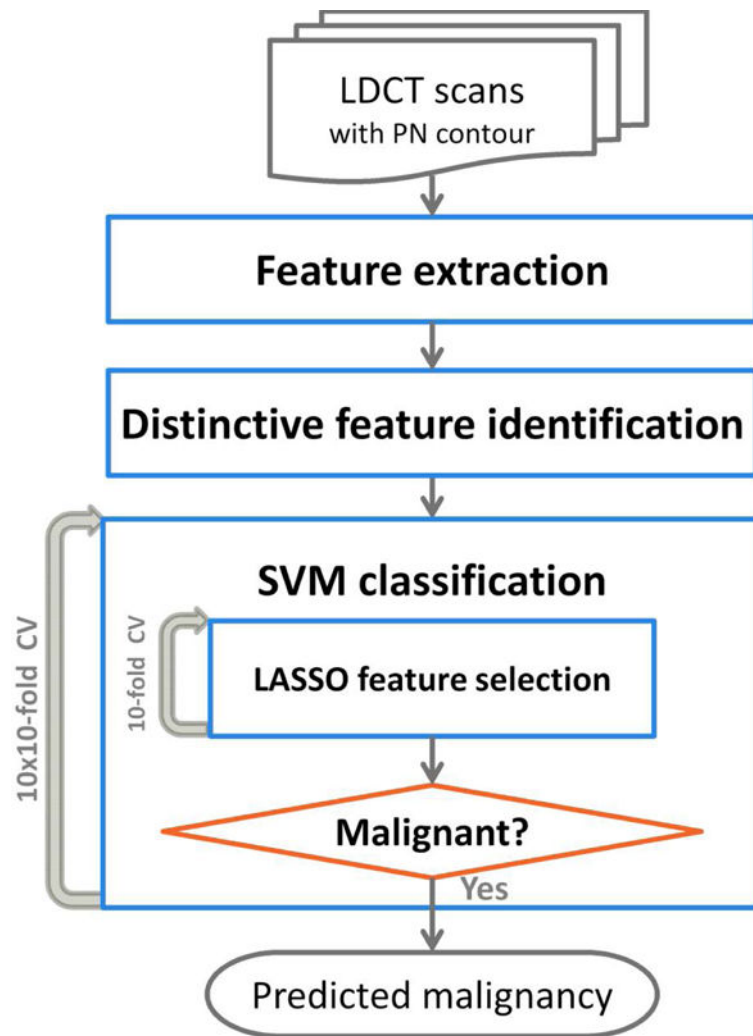


Figure 2.

A flowchart of the extraction of radiomic features and the construction of a prediction model

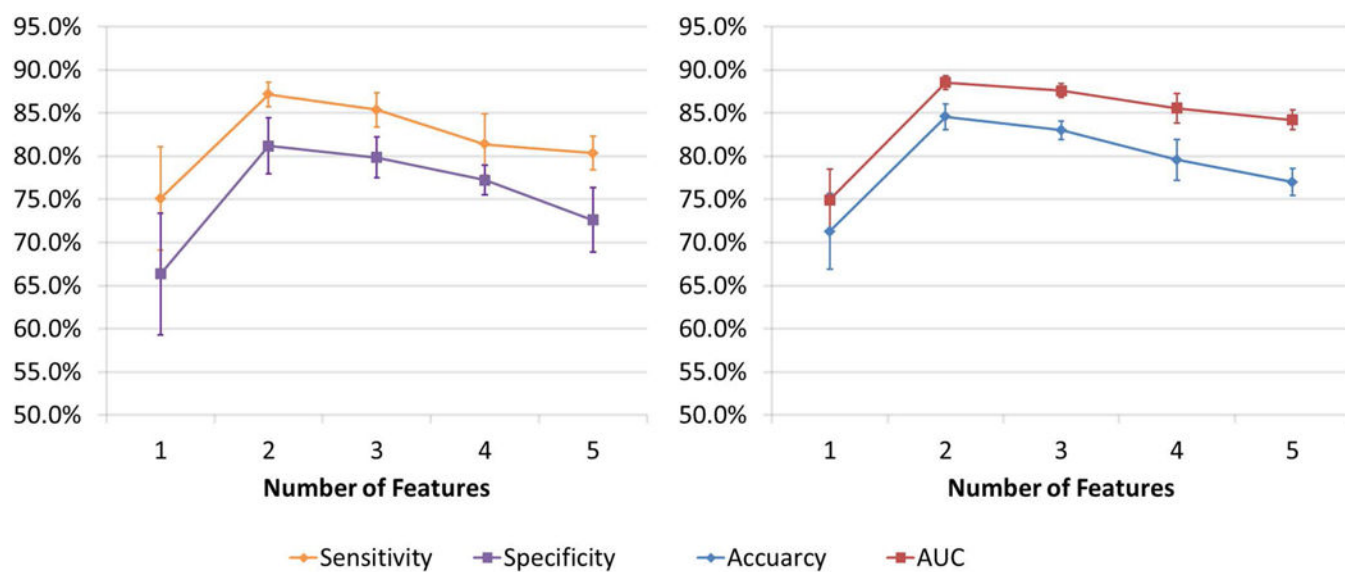


Figure 3.
Performance of the prediction model with increasing number of features in the CV

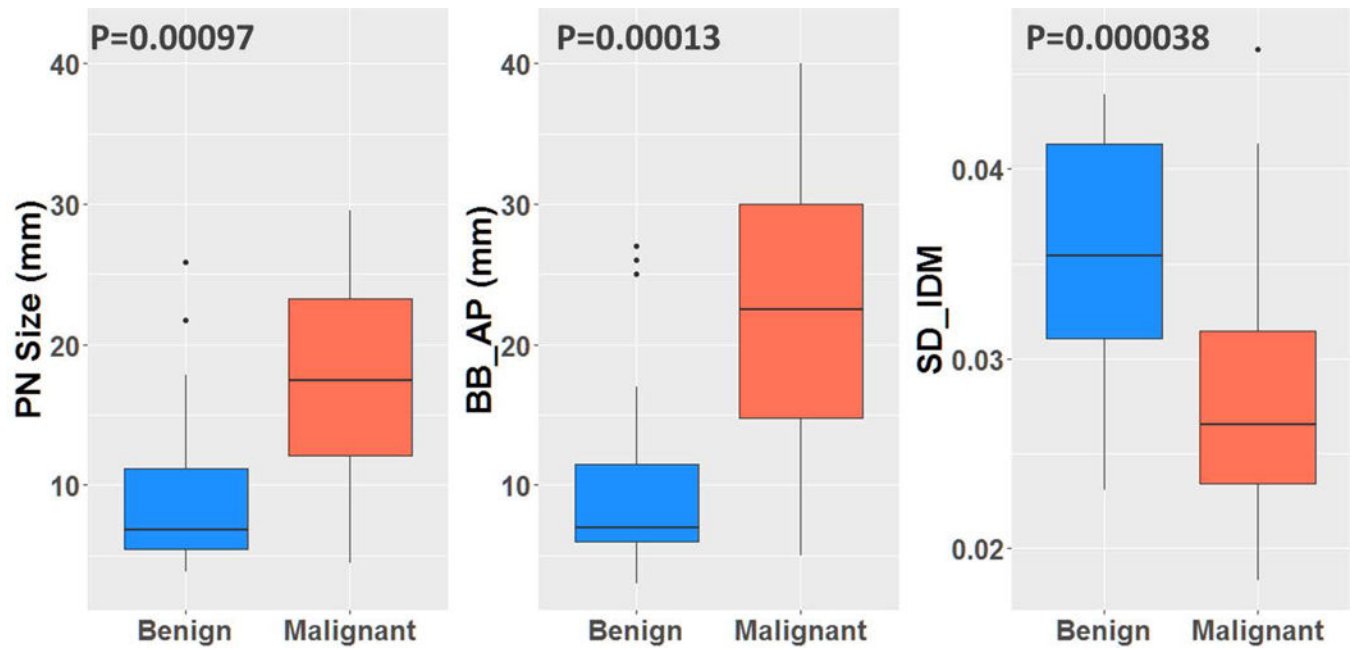


Figure 4.
The box plots show the difference between benign and malignant for PN size and the selected features (BB_AP and SD_IDM). The Wilcoxon rank-sum test obtained P-values.

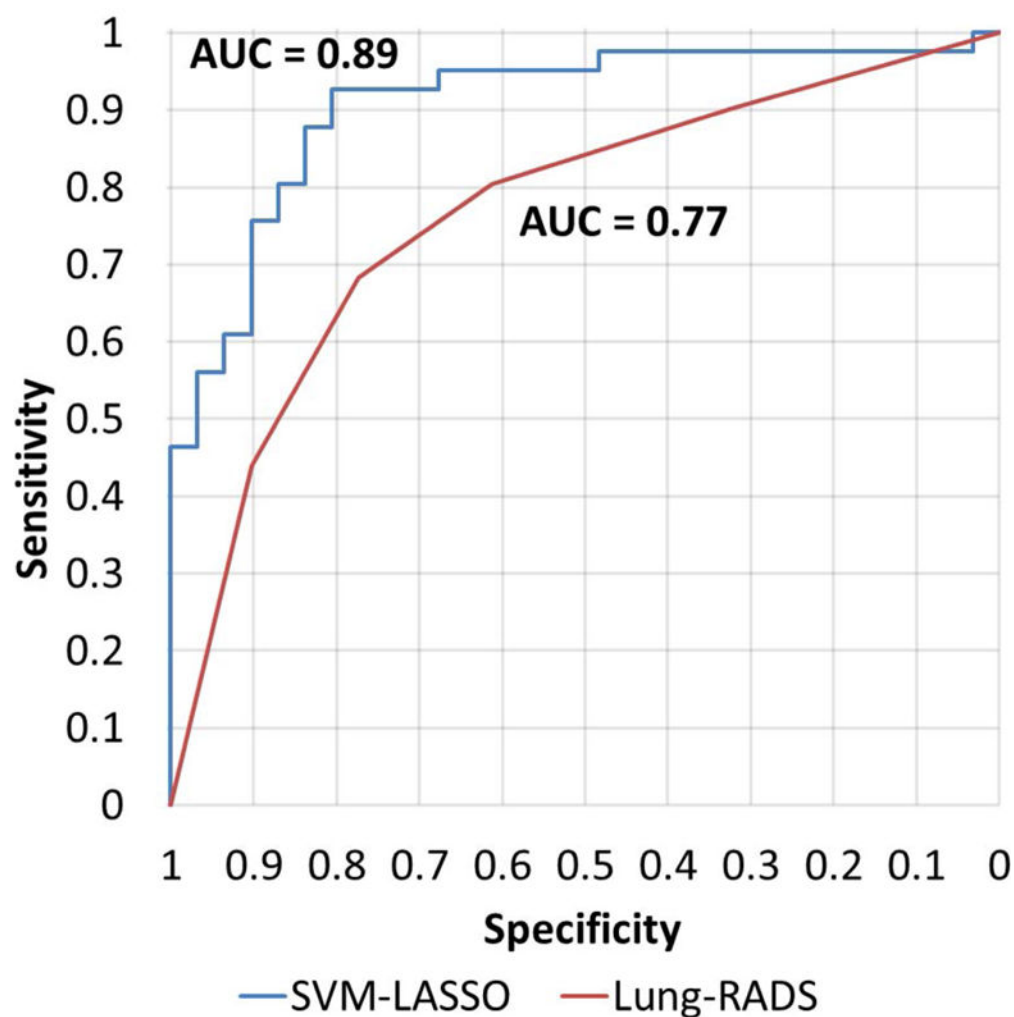


Figure 5.
ROC curve analysis on the best model of SVM-LASSO and Lung-RADS for predicting malignant PNs

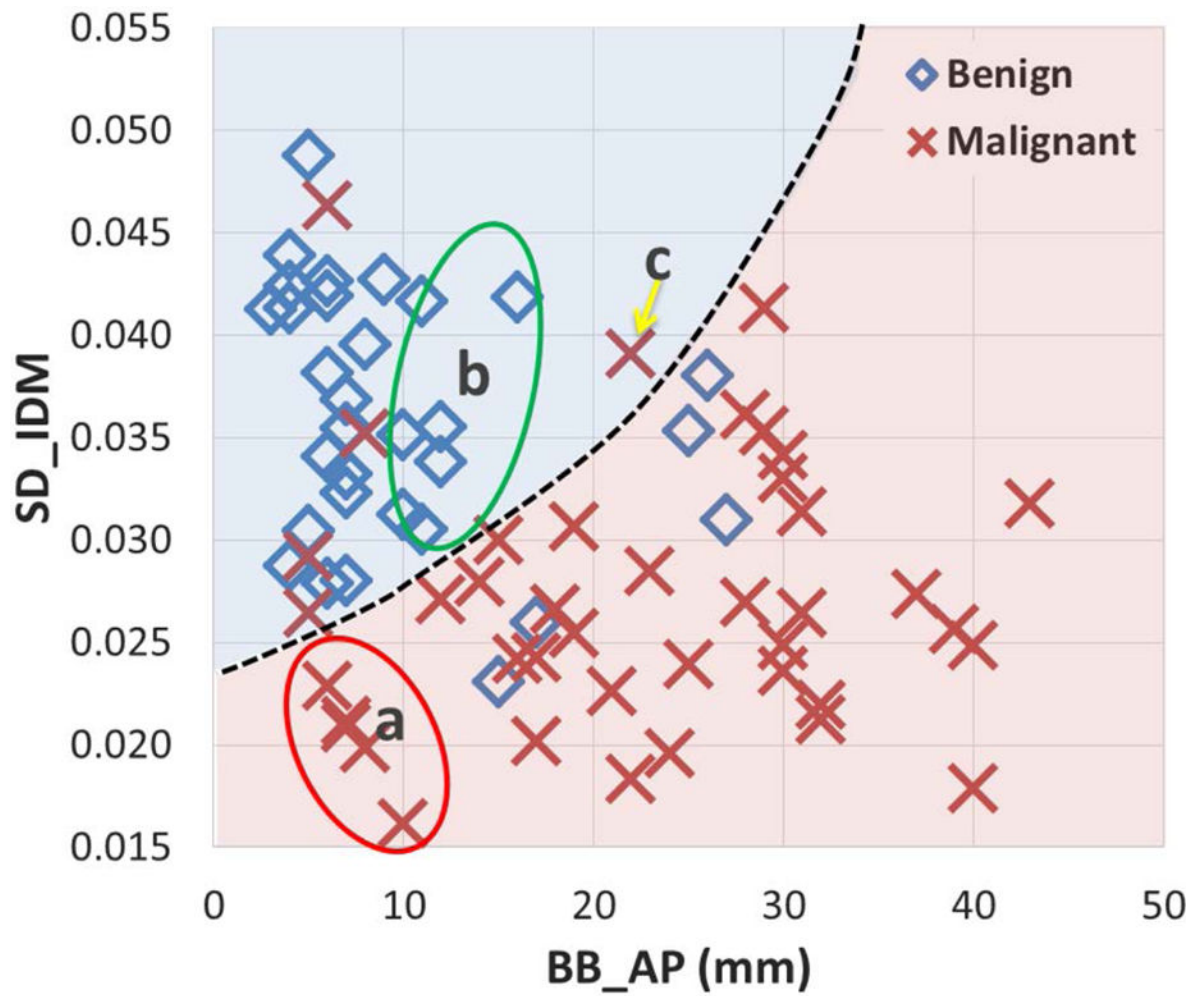


Figure 6.
Scatter plot of the two important features and the classification curve (dashed line) by the SVM-LASSO model for all PNs.

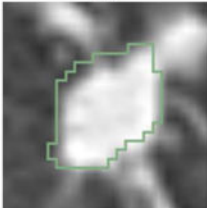
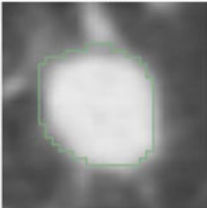
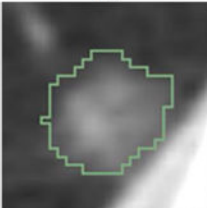
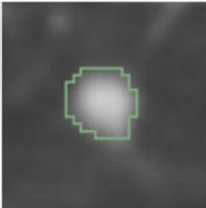
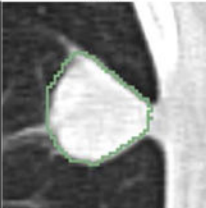
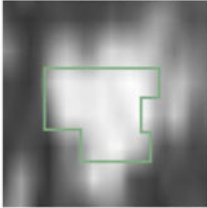
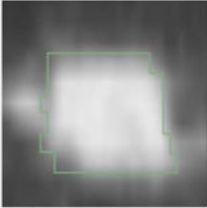
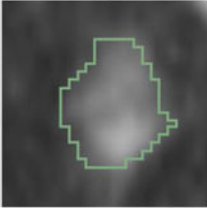
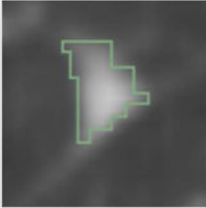
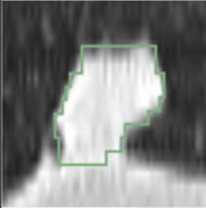
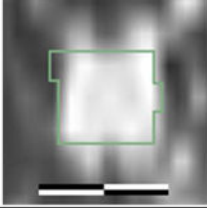
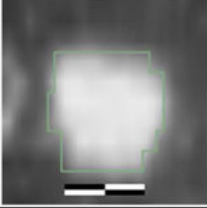
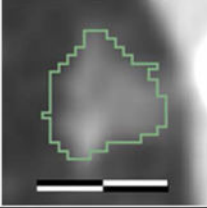
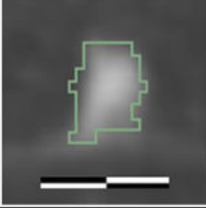
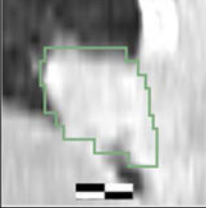
	(a)	(b)	(c)	(d)	(e)
Axial					
Sagittal					
Coronal					
True Status	Benign	Benign	Malignant	Malignant	Malignant
SVM-LASSO	Benign	Benign	Malignant	Malignant	Benign
BB_AP (mm)	10	16	10	7	22
SD_IDM	0.035	0.042	0.014	0.021	0.039
Lung-RADS	4X: Malignant	4A: Malignant	3: Benign	2: Benign	4B: Malignant
Size (mm)	10	16	9 (4)	7	20
Type	Solid	Solid	Part Solid	Solid	Solid
Calcification	Absent	Absent	Absent	Absent	Absent
Spiculation	3	1	1	1	1

Figure 7.

Cases misclassified by Lung-RADS but correctly classified by the SVM-LASSO model (a-d). A case correctly classified by Lung-RADS but misclassified by the SVM-LASSO model (e). The scale bar indicates 10 mm, window/level: 1400/-500 HU. The value in the parenthesis for size is the diameter of the solid component of a part-solid PN. Spiculation is on a 1(no) to 5(marked) scale.

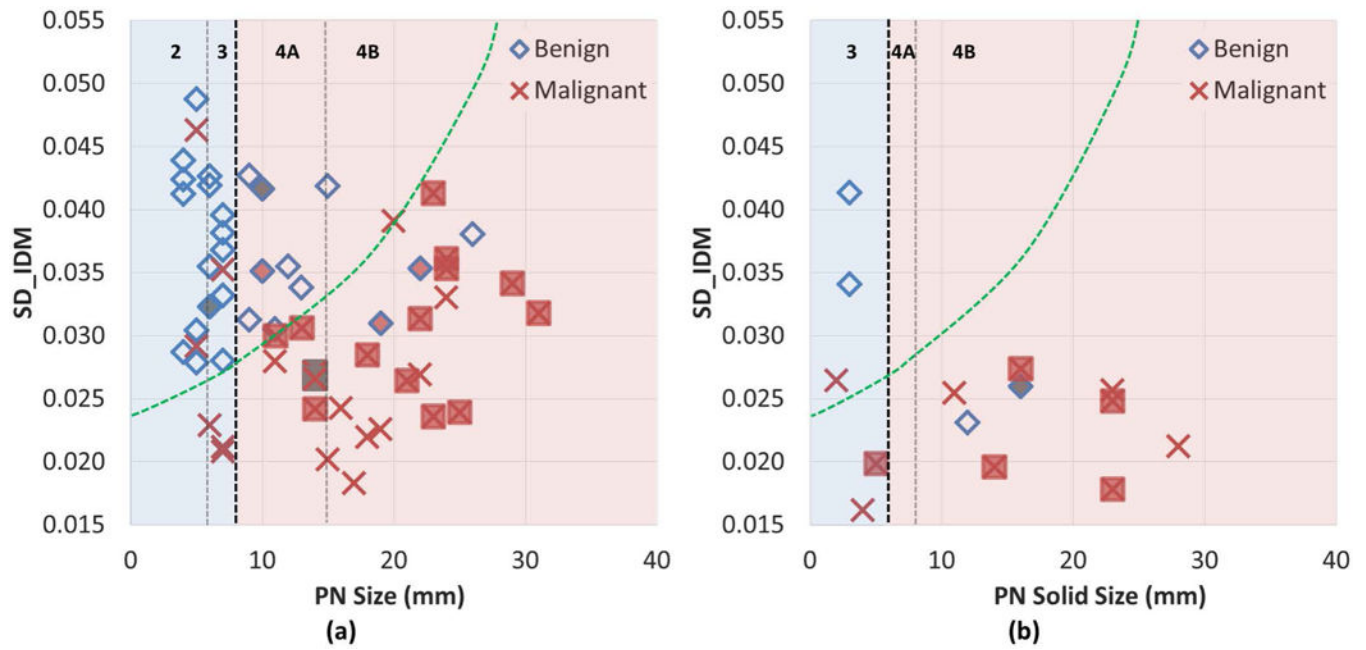


Figure 8.

Lung-RADS categorization on scatter plots for solid PNs (a), and part-solid PNs (b). The SVM-LASSO classification curve is approximately mapped on the plots (green dashed line); Lung-RADS categorization is shown on top with black vertical dashed lines (the bold line indicates classification between benign and suspicious); PNs with calcification (category 1) are filled with gray color, and PNs with spiculation (category 4X) are filled with red color.

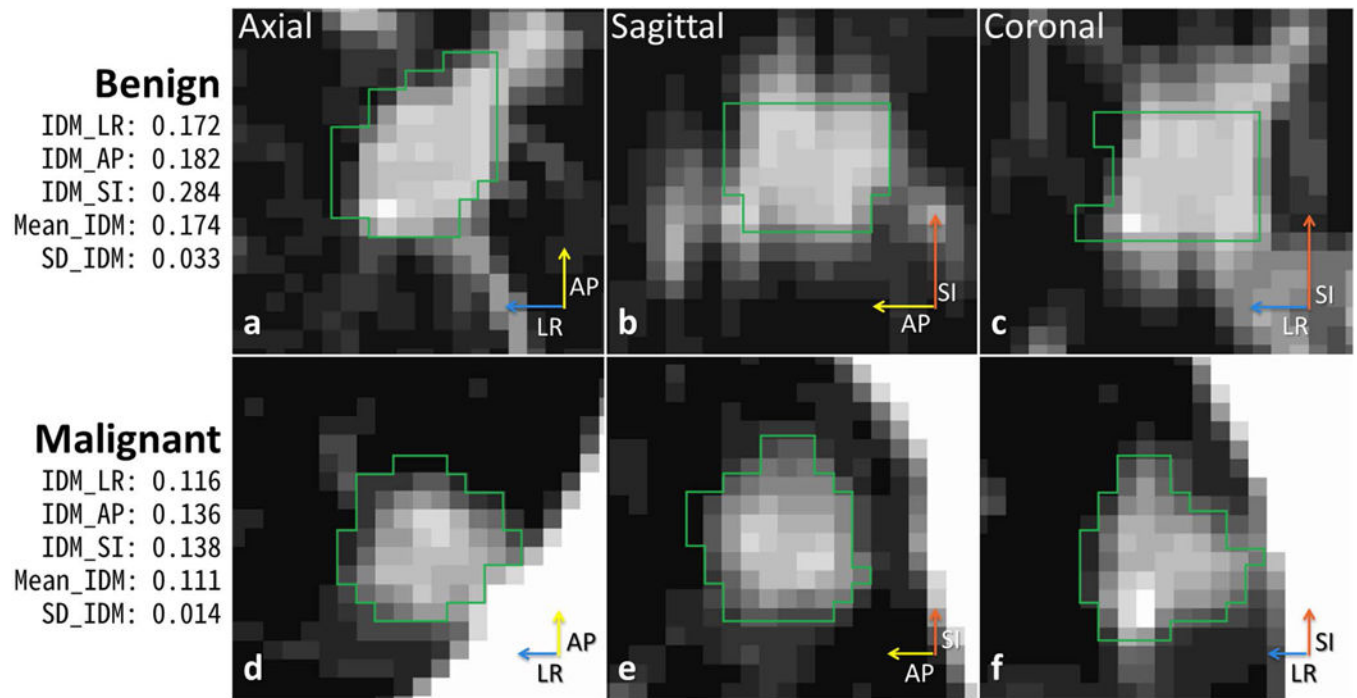


Figure 9. IDM and SD_IDM for small benign and malignant PNs (BB_AP=10mm for both). The length of each arrow indicates IDM value for left-right (LR), anterior-posterior (AP), and superior-inferior (SI) directions.

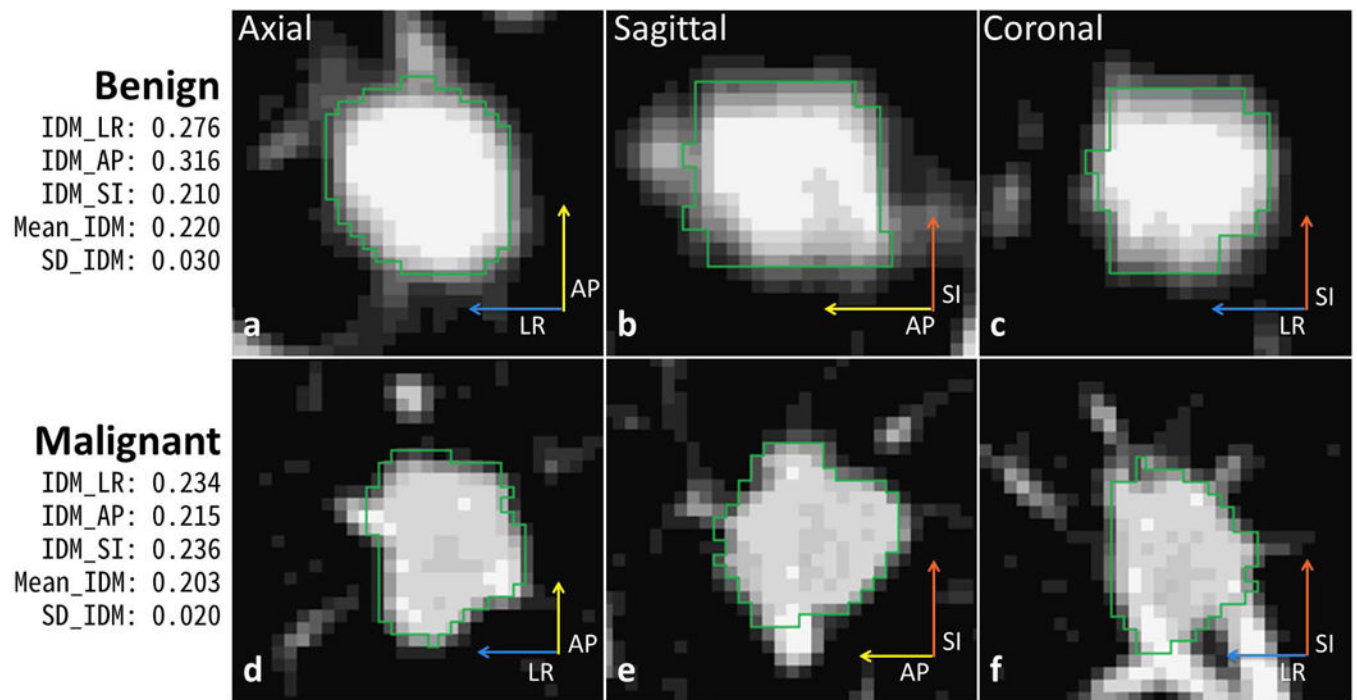


Table 1

Summary of Lung-RADS categorization for baseline screening

Category	Baseline Screening	Malignancy
1	No PNs; PNs with calcification	Negative <1% chance of malignancy
2	Solid/part-solid: <6 mm GGN: <20 mm	Benign appearance <1% chance of malignancy
3	Solid: ≥6 to <8 mm Part-solid: ≥6 mm with solid component <6 mm GGN: ≥20 mm	Probably benign 1-2% chance of malignancy
4A	Solid: ≥8 to <15 mm Part-solid: ≥8 mm with solid component ≥6 and <8 mm	Suspicious 5-15% chance of malignancy
4B	Solid: ≥15 mm Part-solid: Solid component ≥8 mm	>15% chance of malignancy
4X	Category 3 or 4 PNs with suspicious features (e.g., enlarged lymph nodes) or suspicious imaging findings (e.g., spiculation)	>15% chance of malignancy

Table 2

Prediction performance of Lung-RADS and the SVM-LASSO model

Prediction Model	Sensitivity	Specificity	Accuracy	AUC	# of Features
Lung-RADS	80.5%	61.3%	72.2%	0.77	4
SVM-LASSO	87.2±1.4%	81.2±3.2%	84.6±1.5%	0.89±0.01	2

Table 3

Prediction performance of the SVM-LASSO using two features BB_AP and SD_IDM on 10×10-, 20×5-, and 50×2-fold CVs

	Sensitivity	Specificity	Accuracy	AUC
10×10-fold	87.2±1.4%	81.2±3.2%	84.6±1.5%	0.89±0.01
20×5-fold	86.5±2.5%	80.9±2.9%	84.1±2.0%	0.88±0.02
50×2-fold	85.7±4.5%	76.1±10.6%	81.6±5.7%	0.87±0.04

Table 4

Comparison with CADx systems

	Dataset	Model description	Performance
Aoyama et al.¹⁴	76 primary lung cancer (73 pts) 413 benign (342 pts)	- Automatic segmentation using dynamic programming - LDA classification - Leave-one-out CV	- AUC 0.828 (single slice) - AUC 0.846 (multiple slice)
Shah et al.¹⁶	35 solitary PNs 19 malignant and 16 benign PNs - Pairs of CTs without contrast and with IV contrast	- Attenuation, shape, and enhancement features (n=31) - LDA, quadratic discriminant analysis, logistic regression - Leave-one-out CV	AUC 0.69 - 0.92
Suzuki et al.¹⁷	76 primary lung cancer (73 pts) 413 benign (342 pts)	- Multiple massive training ANNs - Integration ANN - Leave-one-out CV	- Sensitivity 100% - Specificity 48% - AUC 0.882
Way et al.¹⁹	256 PNs 124 malignant and 132 benign PNs (152 pts)	3D active contour segmentation - Surface, demographic, and image features - LDA and SVM two-loop leave-one-out CV	- AUC 0.857 (Primary cancers) - AUC 0.822 (Metastatic cancers)
Han et al.²¹	- LIDC 1356 PNs - Ground-truth by radiologist's assessment	- Texture feature analysis - SVM with radial basis function kernel 100 times hold-out validation	- AUC 0.839-0.927 - Average AUC 0.894
LUNGx Challenge (2015)¹⁰	LUNGx 10pts calibration set and 60 pts testing set	- SVM classification (n not reported) - Trained by in-house dataset; Independent cohort validation	- CAD AUC 0.50 - 0.68 - Observer AUC 0.70 - 0.85
Proposed SVM-LASSO	LIDC 72pts	2 important radiomic features LASSO features selection and SVM classification 10×10-fold CV	- Sensitivity 87.2% - Specificity 81.2% - Accuracy 84.6% - AUC 0.89

Table 5

Comparison with recently reported radiomics models

	Dataset	Model description	Sensitivity	Specificity	Accuracy	AUC
Hawkins et al. (2016) ³⁰	■ Baseline CT scans of 261pts in NLST	■ 23 RIDER stable radiomic features ■ Random forest classifier ■ 10×10-fold CV	51.7%	92.9%	80.0%	0.83
Ma et al.(2016) ³¹	■ LIDC 72pts	■ 583 radiomic features ■ Random forest classifier ■ 10-fold CV	80.0%	85.5%	82.7%	
Buty et al.(2016) ³²	■ LIDC 2054 PNs ■ Ground-truth by radiologist's assessment	■ Spherical Harmonics (100, 150, and 400 shape features) and AlexNet⁴⁸ (4096 appearance features) ■ Random forest classifier ■ 10-fold CV			82.4%	
Kumar et al.(2015) ³³	■ LIDC 97pts, including metastatic tumors	■ Deep convolutional neural network model (5000 features) ■ 10-fold CV	79.1%	76.1%	77.5%	
Liu et al. (2016) ³⁶	■ 172 pts (two independent cohorts 102 and 70 pts)	■ 4 image traits selected from 24 image traits scored by radiologists ■ Linear discriminant analysis ■ Trained by 102 and tested by 70 pts	71.4%	83.7%	80.0%	0.81