Jia *et al. Clinical Proteomics* (2025) 22:11
https://doi.org/10.1186/s12014-025-09532-w

Clinical Proteomics

RESEARCH Open Access

An integrated proteomic classifier to distinguish benign from malignant pulmonary nodules

Bin Jia[1], Tingting Wang[2], Liangxuan Pan[2], Xiaoyao Du[2], Jing Yang[3], Fei Gao[2], Lujian Liao[2], Bianqin Guo[3*] and Junqiang Dong[4*]

Abstract

Background Pulmonary nodule with diameters ranging 8–30 mm has a high occurrence rate, and distinguishing benign from malignant nodules can greatly improve the patient outcome of lung cancer. However, sensitive and specific liquid-biopsy methods have yet to achieve satisfactory clinical goals.

Methods We enrolled three cohorts and a total of 185 patients diagnosed with benign (BE) and malignant (MA) pulmonary nodules. Utilizing data-independent acquisition (DIA) mass spectrometry, we quantified plasma proteome from these patients. We then performed logistic regression analysis to classify benign from malignant nodules, using cohort 1 as discovery data set and cohort 2 and 3 as independent validation data sets. We also developed a targeted multi-reaction monitoring (MRM) method to measure the concentration of the selected six peptide markers in plasma samples.

Results We quantified a total of 451 plasma proteins, with 15 up-regulated and 5 down-regulated proteins from patients diagnosed as having malignant nodules. Logistic regression identified a six-protein panel comprised of APOA4, CD14, PFN1, APOB, PLA2G7, and IGFBP2 that classifies benign and malignant nodules with improved accuracy. In cohort 1, the area under curve (AUC) of the training and testing reached 0.87 and 0.91, respectively. We achieved a sensitivity of 100%, specificity of 40%, positive predictive value (PPV) of 62.5%, and negative predictive value (NPV) of 100%. In two independent cohorts, the 6-biomarker panel showed a sensitivity, specificity, PPV, and NPV of 96.2%, 35%, 65.8%, and 87.5% respectively in cohort 2, and 91.4%, 54.2%, 74.4%, and 81.3% respectively in cohort 3. We performed a targeted LC-MS/MS method to quantify plasma concentration of the six peptides and applied logistic regression to classify benign and malignant nodules with AUC of the training and testing reached 0.758 and 0.751, respectively.

Conclusions Our study identified a panel of plasma protein biomarkers for distinguishing benign from malignant pulmonary nodules that worth further development into a clinically valuable assay.

Keywords Plasma, Classification, Biomarker, Pulmonary nodule, Lung cancer

*Correspondence:
Bianqin Guo
guo_bianqin@163.com
Junqiang Dong
dongjq588@163.com

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

Background

Each year around 1.6 million people in the United States are diagnosed with pulmonary nodule [1], whereas several population-based studies estimated nodule occurrence rate between 3 and 13% (Oudkerk, Liu, Heuvelmans, Walter, & Field [2]), imposing serious health care issues to the society. The initial diagnosis of pulmonary nodule is usually made from low-dose computed tomography (LDCT), in which the probability of malignancy reaches 10–80% in nodules with a size greater than 10 mm, whereas nearly 99% of nodules are benign with a size less than 6 mm [1]. Nodules with the size between 5 ~ 30 mm presents a dilemma to clinicians, because the estimated probability of cancer between 5 ~ 65% [3] is too wide a range to be actionable: surgical procedures may result in over treatment leading to various complications, whereas nodules left untreated may bear the risk of malignancy and tumor progression. Thus, developing sensitive diagnostic methods to differentiate benign from malignant nodules is critically important.

Currently, non-invasive diagnostic methods such as liquid biopsy detect tumor-derived nucleic acids or differentially expressed proteins in the blood or other body fluids. Circulating tumor cells provide a plethora of tumor-derived nucleic acid species for detection, including circulating tumor DNA (ctDNA) [4], cell-free DNA (cfDNA) [5], as well as circulating messenger RNA (mRNA) and long non-coding RNA (lncRNA) [6]. Detection of nucleic acids derived from circulating tumor cells possess the intrinsic property of specificity for detecting cancer, therefore greatly reduces the false positive rates. Aided by newer generation sequencing methods, it also maintains a high sensitivity, especially applicable in bronchoalveolar fluids [7]. In addition, methylation of panels of tumor suppressor genes and/or oncogenes also produced satisfactory sensitivity and specificity in distinguishing between lung cancer and benign pulmonary nodules, and has the potential of clinical utility [8, 9]. Other blood-derived inflammation parameters including neutrophil-to-lymphocyte ratio, platelet-to-lymphocyte ratio, and systemic immune-inflammation index have been assessed to associate with malignant pulmonary nodules with mild odds ratios [10].

In contrast to the wide range of biomarker studies utilizing nucleic acid as targets, studies on protein biomarkers are relatively scarce. Notably, Pulmonary Nodule Plasma Proteomic Classifier (PANOPTIC) presented an integrated protein biomarker incorporating several key clinical features to distinguish benign from malignant nodules [11–13], it is by far the only mass spectrometry-based assay that has been used in clinic to manage nodule patients. A two-year follow-up study shows that with a sensitivity of 97%, specificity of 44%, and negative predictive value of 98%, this biomarker performs superior than physician estimates of probability of cancer [14]. Nevertheless, this biomarker panel still suffers from low specificity and thus relatively poor positive predictive value, therefore further development on protein biomarker panel remains warranted.

In this study, we applied quantitative mass spectrometry technology to retrospectively compare the plasma proteome between patients diagnosed with benign and malignant pulmonary nodules. The nature of the nodule was pathologically confirmed through fine-needle aspiration. We captured unique molecular features of the plasma proteome from patients with malignant nodule. Using these features, we applied logistic regression to discover a panel of protein biomarkers that perform well in distinguish benign from malignant pulmonary nodules.

Methods

Human samples

The study started in June 30th, 2020 and ended in August 31st, 2022. Peripheral venous blood samples were collected before surgery. The patient recruitment criteria included: the overall groups of benign and malignant samples were matched in gender, nodule size and age. Using a population based non-small-cell lung cancer prevalence estimate, patients whose nodule size lies between 3 mm and 30 mm were included. Patients with a history of other cancers were excluded. The study was approved by the hospital's institutional review committee. The participants provided the written informed consent. T test and chi-squared test were applied to the clinical information and to ensure that there was no statistical significance in clinical factors between the two patient groups. The study conformed to the Declaration of Helsinki and was approved by the Medical Ethics Committee of The First Affiliated Hospital of Zhengzhou University (Study license number 2020-KY-308) and Chongqing University Cancer Hospital (Study license number CZLS2023268-A).

In the discovery cohort (cohort 1), plasma samples from 40 patients diagnosed with benign and 40 patients diagnosed with malignant pulmonary nodules were collected from First Affiliated Hospital of Zhengzhou University. In two independent validation cohorts - cohort 2 (20 benign, 26 malignant) and cohort 3 (24 benign, 35 malignant), the plasma samples were collected from The First Affiliated Hospital of Zhengzhou University and Chongqing University Cancer Hospital, respectively. We calculated the effect size of the six protein markers using the normalized intensity results, taken into consideration the fold changes and pooled standard deviation of both benign and malignant groups. The Cohen's d ranges from small effect size of 0.213 to large effect size of 0.921 among the six peptides. By calculating the theoretical effect size given existing sample size of 40 in each

group, α of 0.05 and a power of 0.8, resulting the effect size d = 0.634, at least two protein markers reached this desired effect size in the discovery cohort.

Processing of plasma samples

For DIA proteome discovery assay, the plasma samples were depleted of top 14 high-abundant proteins (Cat. #A36370, Thermo Science, USA), and BCA kit was used to determine the protein concentration in plasma samples. From each sample, 25 μg proteins were suspended in NH_4HCO_3 solution with a final concentration of 50 mM. The proteins were reacted with 10 mM DTT at 95 °C for 10 min and alkylated with 15 mM iodoacetamide (Cat. # I1149, Sigma Aldrich, USA) in darkness for 30 min and digested with sequencing grade trypsin (1:50; Cat. # V5113, Promega, USA) overnight at 37 °C. The resulting peptides were desalted with 96-well SOLA solid-phase extraction apparatus, and vacuum dried for mass spectrometry analysis.

For the MRM experiments, 8 μL plasma samples were also depleted of top 14 high-abundant proteins. For each sample, 20 μg proteins were subsequently processed the same way as described in DIA experiments. The six internal standards using synthetic stable isotope-labeled peptides were spiked in the proteotypic peptides or the standard curve samples at a final concentration of 25 ng/mL for each peptide for absolute quantification (Synpeptide, China). The resulting peptides were desalted with 96-well SOLA solid-phase extraction apparatus, and vacuum dried. The resulting peptides were resuspended in 80 μL 0.1% FA, and 15 μL were injected for mass spectrometry analysis.

Data-independent acquisition mass spectrometry (DIA-MS) of plasma samples

Protein digests were analyzed using an EASY-nLC 1200 LC coupled with Q-Exactive Explories 240 mass spectrometer (ThermoScientific, USA). To assess the stability of the analytical system, a QC sample composed of a small aliquot of each sample was interspersed in every five samples. In addition, the sequence of the sample run was completely randomized. For the LC separation, the mobile phases consisted of buffer A (100% ddH2O, 0.1% formic acid) and buffer B (80% ACN, 0.1% formic acid). Peptides were resuspended in buffer A and spiked with iRT peptides (Omicsolution, China). An equivalent to 1.5 ug of protein digest from each sample was loaded onto a C18 column (Cat. #164941, Thermo Science, USA) connected with a pre-column (Cat. #164535, Thermo Scientific, USA) and separated at a flow rate of 300 nL/ min. A 90 min gradient from 1 to 8% buffer B in 1 min, 8–28% in 71 min, 28–40% in 9 min, 40–100% in 2 min, and 100% for 7 min was used.

The MS instrument was operated in the positive mode and centroid mode with a nano-electrospray through a heated ion transfer tube with a temperature setting of 320 °C. For data dependent acquisition (DDA), one full scan MS from 350 to 1500 m/z followed by 20 MS2 scans were continuously acquired. MS spectra were acquired with resolution of 60,000 for an auto maximum injection time (IT) with an automatic gain control (AGC) target value of 3e6. MS2 spectra were obtained in the higher-energy collisional dissociation (HCD) mode using a normalized collision energy of 30%, resolution at 15,000 with an auto maximum injection time, AGC target of 1e5 and isolation window at 1.6 m/z. For data independent acquisition (DIA), isolation window for MS2 was set to 20 Da window over a precursor mass range of 350–400 m/z, 9 Da window for 400–800 m/z, 12 Da window for 800–100 m/z and 25 Da window for 1000–1200 m/z. AGC target of 5e5 with an auto maximum injection time, and other parameters were set to be the same as DDA method.

Mass spectrometry data analysis

DDA MS/MS spectra were searched using Protein Discoverer 2.4 (Thermo Scientific, USA) against a UniportKB human database (UP000005640). The following settings were used: trypsin was set to the protease and one missed cleavage allowed; The precursor mass tolerance was set to 10 ppm, and the fragmentation ion mass tolerance was set to 0.02 Da. The maximum number of variable modifications was set to 2. The false discovery rate (FDR) was set at 1% at both the peptide and protein level.

DIA MS/MS spectra were searched using the DIA-NN (version 1.7.15) software with a UniportKB human database (UP000005640). The precursor mass tolerance was set to 10 ppm, trypsin was set to the protease and one missed cleavage allowed. The maximum number of variable modifications was set to 3. Precursor mass range was from 350 to 1250 m/z, while the fragmentation ion mass range was from 100 to 2000 m/z. The false discovery rate (FDR) was set at 1% at both the peptide and protein level.

For differential expression analysis, proteins with more than 50% missing values were removed. The distribution of protein expression was tested for normality across all samples, then t-test was applied for those with normal distribution, while Wilcoxon ranked sum test was performed for those failed to pass the normality test.

Multiple reaction monitoring (MRM) quantitation of plasma proteins

Concentrations of target proteins in the plasma were measured using MRM method on a QTRAP 6500 mass spectrometer (Sciex, USA). The instrument parameters of the MRM assay were optimized for each synthetic

peptide by directly infusing the peptides into the mass spectrometer. The top two high-intensity product ions of each peptide precursor ion were selected based on the optimal collision energy (CE) values and collision cell exit potential (CXP). All optimized data were collected and compared to theoretical spectra, and high-intensity y-ions were used for subsequent MRM assays.

The peptides were separated using an LC-40D X3 (Shimadzu, Japan) liquid chromatographic system. The mobile phase A was 0.1% formic acid in distilled water, and the mobile phase B was 0.1% formic acid in 100% acetonitrile. Peptides were reconstituted in mobile phase A, 15 μL of each sample was loaded into the sample loop. A gradient consisting of 13% B for 1.5 min, 13–48% buffer B for 4.5 min, 48–98% for 0.5 min, 98% for 1.5 min, 98–13% B for 0.1 min, and 13% B for 1.9 min was used. The MS detection was carried out in positive mode with the following parameters: electrospray voltage of 5500 V, curtain gas at 40 psi, ion source gas 1 (GS1) at 55 psi, ion source gas 2 (GS2) at 55 psi, and temperature at 500 °C. Quantitation were performed using the scheduled MRM mode. The time of MRM detection window was 180 s, and the cycle time was 1.0 s.

The mass spectrometer was controlled by the Analyst software (Sciex, USA), and the raw data were analyzed by Sciex OS software (Sciex, USA). Calibration curves for each peptide were generated using synthetic peptides and stable-isotope labeled internal standards with a linear regression coefficient of determination (R^2) ≥ 0.99.

GO and KEGG pathway analysis

GO and KEGG enrichment analysis were performed using Metascape (https://metascape.org/), David bioinfo rmatics resource v6.8 (https://david.ncifcrf.gov/).

Statistical analysis

R (version 4.1.1) was used for all the statistical analysis, Prior to analysis, rigorous preprocessing steps were applied to the data. First, the data was normalized by column sum. Next, outliers were removed to mitigate the impact of extreme values on subsequent analyses. Proteins with less than 50% of missing data across all samples were imputed using a random forest imputation method to minimize information loss. Principal Component Analysis (PCA) was then applied to reduce the dimensionality of the dataset and explore the major sources of variation.

To assess the assumptions required for parametric tests, we tested for normality and homoscedasticity. Depending on the sample size and distribution characteristics, we employed the Shapiro-Wilk test to assess the normality of data distribution and Levene's test to examine homogeneity of variance. For samples that met the assumptions of normality and homoscedasticity, parametric analyses were performed. Conversely, non-parametric analyses were conducted for samples that did not satisfy these assumptions. To compare two groups, we utilized both parametric t-tests and non-parametric Mann-Whitney U tests, depending on the data distribution and underlying assumptions. All tests were two sided, and p values < 0.05 were considered statistically significant, and fold change > 1.2 or < 0.83 were considered as up- or down-regulated, respectively.

Feature selection and logistic regression

For the peptides with P value < 0.05 and fold change of 1.2 between malignant and benign groups, we considered these peptides showing consistent expression trend at protein level. 12 peptides were selected as the target lists, and random combinations of 3, 4, up to 12 from the 12 features were considered, using AUC values as a selection criterion. Finally, we selected a panel of 6-peptide markers which had the AUC among the top and balancing sensitivity and specificity. The expression values of the six peptides were used as features in a logistic regression model to classify subjects as having either benign or malignant nodules.

Survival analysis

For Kaplan-Meier survival analysis, log-rank test was applied to analyze differences in survival time [15]. The influence of gene expression on survival time was evaluated by the Cox proportional hazard model. Samples were stratified according to transcription levels: samples with gene expression higher than 50% were considered as the high-expression cohort, while samples with gene expression lower than 50% were considered as the low-expression cohort. The lung adenocarcinoma RNAseq dataset from TCGA (Dataset ID: TCGA.LUAD. sampleMap/HiSeqV2) was used, and clinical information was downloaded from the UCSC genome browser (https://xenabrowser.net/).

Results

Study design

The design of this study is shown in Fig. 1. We enrolled three cohorts of patients from two hospitals, with each patient pathologically confirmed as having either benign or malignant pulmonary nodule. Cohort 1 (80 patients) was used for discovery study whereas cohort 2 (46 patients) and cohort 3 (59 patients) were used for validation studies as two independent cohorts. Plasma samples from cohort 1 were used for differential protein expression analysis to compare patients between malignant versus benign nodules. From the list of proteins with altered expression, candidate protein biomarkers were selected and logistic regression classification was applied to identify panels of proteins whose expression can differentiate

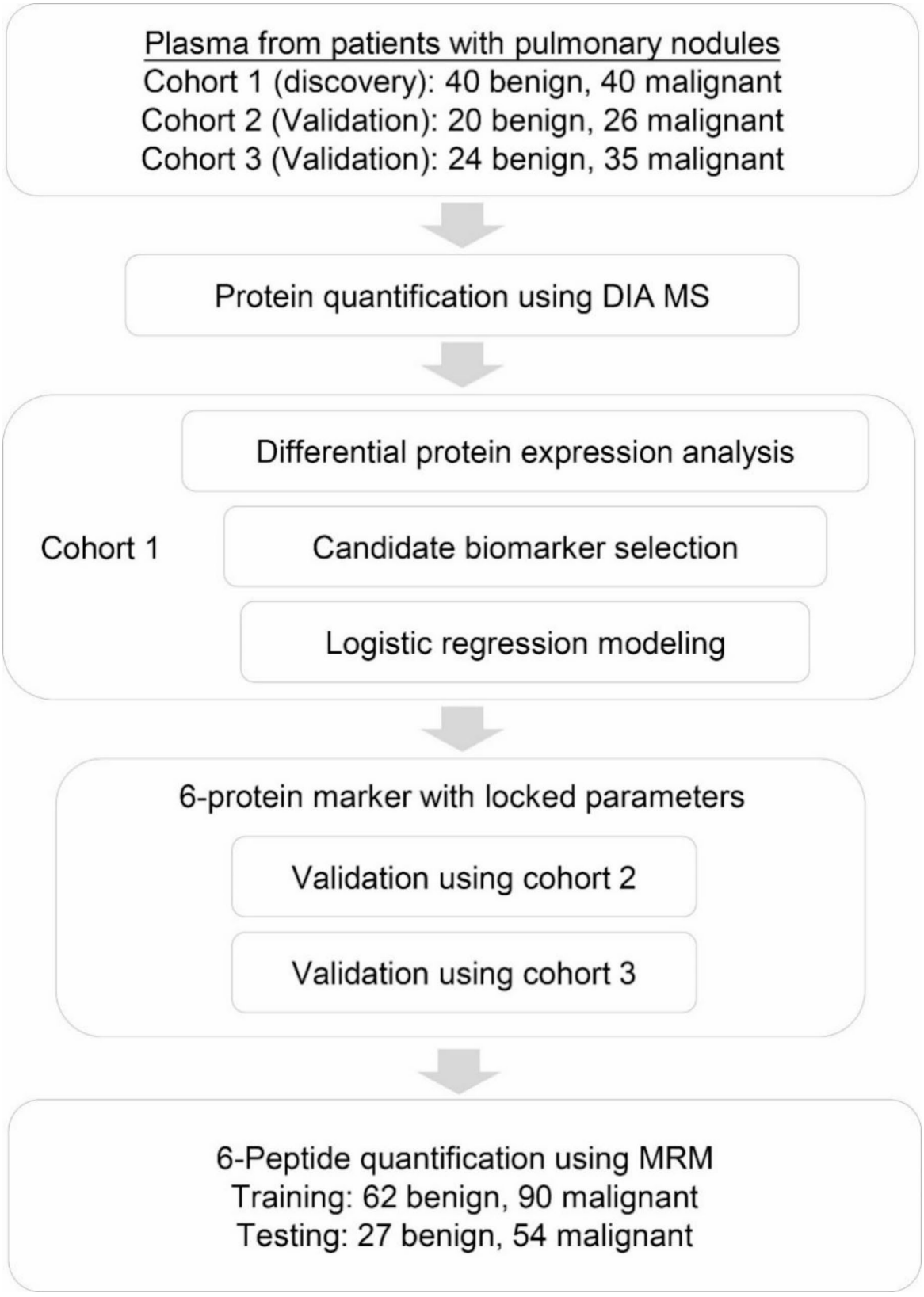

Fig. 1 Study design. Flow chart showing the design this study

nodule types between benign and malignant. These biomarker panels were further validated by two independent cohorts from two hospitals. In the assay development phase, an MRM method was applied to a subset of 155 plasma samples from the 185 subjects in the three DIA cohorts and 78 newly collected samples to measure the plasma concentration of selected protein biomarkers using the signature peptides. The demographic data for all the patients enrolled in this study is shown in Table 1.

Quantitative mass spectrometric analysis of plasma proteome from patients with benign and malignant pulmonary nodules

We applied data-independent acquisition mass spectrometry technology (DIA-MS) to quantify the plasma proteome. To ensure the high quality of our data acquisition process, we interspersed quality control (QC) samples during mass spectrometry data acquisition. The QC sample is the mixture of a small portion from each plasma sample from all patients in the discovery stage. The correlation coefficients of all the QC samples were over 0.99 (Fig. S1A), indicating high consistency and reproducibility of our experimental procedure. Signal intensity of proteins spans six orders of magnitude (Fig. S1B), indicating that even after depletion of high-abundance proteins the dynamic range of protein abundance remains high.

In total, we quantified 451 proteins from 80 plasma samples in cohort 1 (Fig. S1C, Table S1). Using a statistical significance cutoff of 0.05 and fold change cutoff of 1.2, we identified 15 up-regulated and 5 down-regulated proteins, which were shown in a volcano plot (Fig. 2A). Heat map of the differentially expressed proteins showed that while these proteins display subtle expression patterns, variabilities among each of the two groups appeared more evident (Fig. 2B). This was also supported by the principal component analysis showing that the two groups, benign (BE) and malignant (MA) could not be separated by these proteins, and that PC1 only explained 7.42% of the variability (Fig. 2C). Gene ontology analysis and KEGG pathway analysis showed that the most enriched molecular function was involved in cholesterol metabolism, and lipid transport and lipoproteins were among the enriched biological pathways. Not surprisingly, exosomes were among the most enriched cellular components (Fig. 2D). The proteins representing these pathways included apolipoprotein A1 (APOA1), apolipoprotein A2 (APOA2), apolipoprotein D (APOD), apolipoprotein A4 (APOA4), to name a few.

Discovery of a plasma protein biomarker panel to distinguish benign from malignant nodules

To identify biomarkers that can accurately differentiate benign from malignant nodules, we focused on proteotypic peptides as surrogates of six protein markers comprised of APOA4, CD14, PFN1, APOB, PLA2G7, and IGFBP2 (Fig. 3). The peptide sequences were listed in Table 2. The expression at protein level was also presented, among these proteins, APOA4, CD14, and PFN1 showed significantly differential expression, while APOB, PLA2G7, and IGFBP2 showed a P value greater than 0.05 (Fig. S2). Nevertheless, using these six proteins as a panel and cohort 1 as the discovery set, PCA showed improved separation between benign and malignant nodules, and the first principal component explained 30% of the variability (Fig. 4A). We then built a logistic regression classification model for the six-protein panel, and resulted in an AUC of the receiver operator characteristic (ROC) curve of 0.87 and 0.91, in the training and testing set, respectively (Fig. 4B). The distribution of sensitivity, specificity, PPV, and NPV values as the function of the threshold used to calculate logistic regression odds ratios were shown in Fig. 4C, which indicated that a threshold value of 0.164 balanced all four values. Using this threshold value, the sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated from the confusion matrix (Fig. 4D), and resulting in 1.0, 0.40, 0.625, and 1.0, respectively.

Validation of the plasma protein biomarker panel in independent patient cohorts

We used Cohort 2 and 3 as the independent validation cohorts to assess the classification accuracy of the panel. While the plasma samples from cohort 2 was independently collected from the same hospital, plasma samples from cohort 3 was from a different hospital. The same DIA-MS method was applied to quantify plasma proteins. Although the PCA generated from the intensities of the six-protein panel measured from these two patient cohorts showed no obvious separation, the first component explained about 30% of the variation for both cohorts (Fig. 5, A-B). Using the logistic regression parameters and the threshold value to classify benign from malignant nodules, the AUC of the ROC curve remained 0.82 and 0.81, respectively (Fig. 5C). The biomarker panel showed a sensitivity, specificity, PPV, and NPV of 0.962, 0.35, 0.658, and 0.875 respectively in cohort 2 (Fig. 5D, upper panel), and of 0.914, 0.542, 0.744, and 0.813 respectively in cohort 3 (Fig. 5D, lower panel).

The plasma protein biomarker panel is capable of detecting malignant pulmonary nodules with high accuracy

As DIA is a relative quantification method and MRM is able to measure the absolute concentration of biomarkers with high throughput, we developed an MRM assay to measure the plasma concentration of the 6 proteins by measuring the signature peptides using heavy arginine/

Table 1 Demographic data and disease characteristics of the participants in the three cohorts

Characteristics		Cohort 1 (discovery) $n=80$			Cohort 2 (validation) $n=46$			Cohort 3 (validation) $n=59$		
	level	Benign	Malignant	P	Benign	Malignant	P	Benign	Malignant	P
Number		40	40		19	27		24	35	
Sex (%)	Female	19 (47.5)	24 (60.0)	0.37	9 (47.4)	12 (44.4)	1	9 (37.5)	24 (68.6)	0.032
	Male	21 (52.5)	16 (40.0)		10 (52.6)	15 (55.6)		15 (62.5)	11 (31.4)	
Age (median [IQR])		56.50 [51.75, 63.25]	57.00 [51.50, 68.25]	0.61	57.00 [52.00, 59.00]	58.00 [55.00, 64.00]	0.336	54.00 [48.75, 57.00]	52.00 [48.00, 63.00]	0.763
Size (median [IQR])		13.50 [10.78, 16.88]	15.50 [10.00, 20.15]	0.427	13.00 [10.00, 15.50]	15.00 [11.20, 20.85]	0.131	12.00 [7.50, 21.00]	12.00 [9.00, 17.00]	0.94
Histopathology (%)	Adenocarcinoma		31 (77.5)	0.006		22 (81.5)	0.022		29 (82.9)	0.004
	Squamous cell carcinomas		0 (0.0)			5 (18.5)			2 (5.7)	
	Adenocarcinoma in situ		7 (17.5)			0 (0.0)			3 (8.6)	
	Other categories		2 (5.0)						1 (2.9)	
	Granuloma	5 (12.5)			1 (5.3)			4 (16.7)		
	Hamartoma	7 (17.5)			1 (5.3)			1 (4.2)		
	Hyperplasia	2 (5.0)			1 (5.3)			2 (8.3)		
	Inflammation	1 (2.5)			0 (0.0)			4 (16.7)		
	Mixed	0 (0.0)			0 (0.0)			0 (0.0)		
	Pneumonia	21 (52.5)			12 (63.2)			7 (29.2)		
	Tuberculosis	2 (5.0)			0 (0.0)			4 (16.7)		
	Other	2 (5.0)			4 (21.1)			2 (8.3)		
Smoking (%)	Current	12 (30.0)	6 (15.0)	0.18	5 (26.3)	8 (29.6)	1	10 (41.7)	9 (25.7)	0.496
	Former	0 (0.0)	0 (0.0)		0 (0.0)	0 (0.0)		1 (4.2)	2 (5.7)	
	Never	28 (70.0)	34 (85.0)		14 (73.7)	19 (70.4)		13 (54.2)	24 (68.6)	

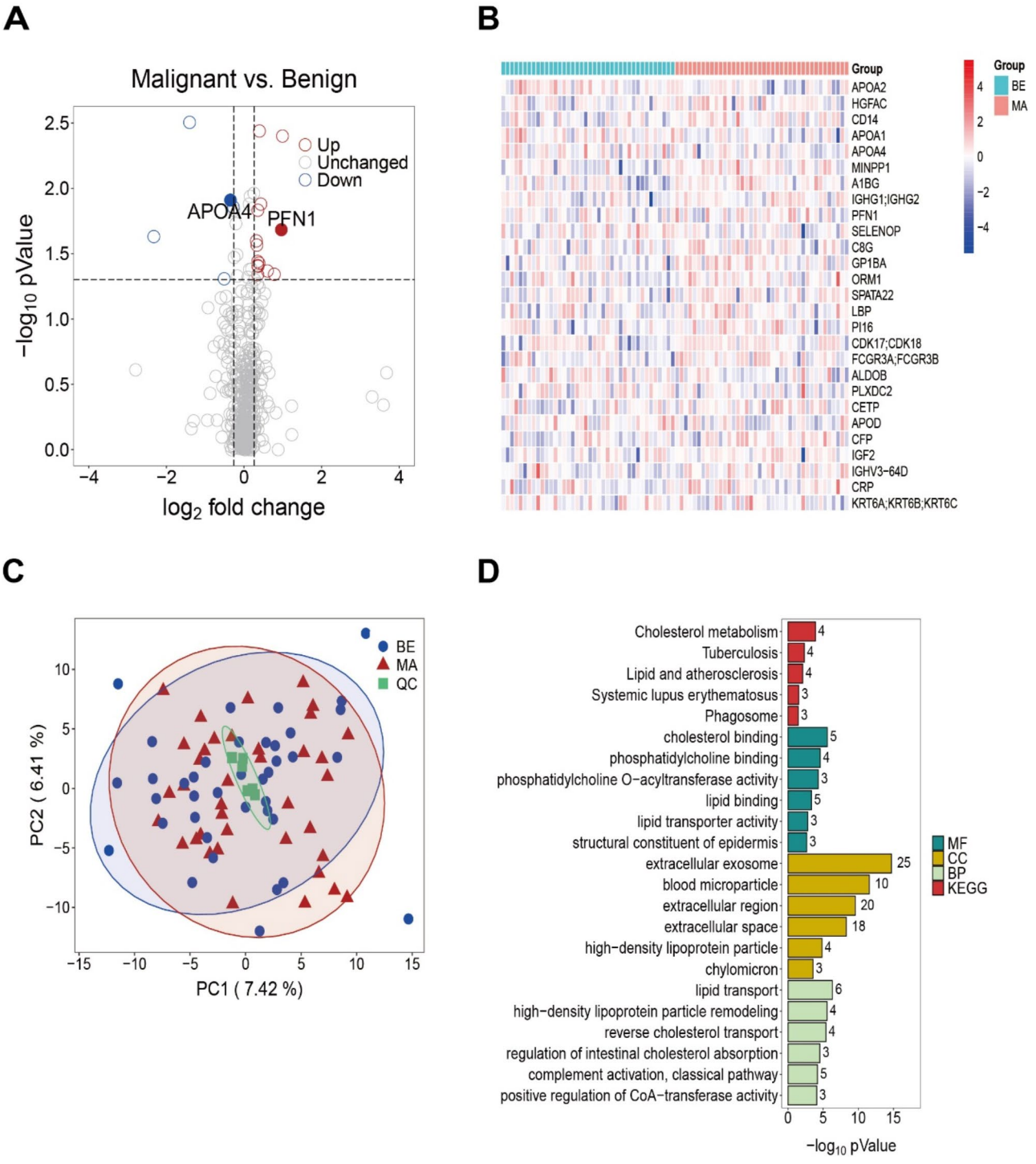

Fig. 2 Quantitative proteomic analysis of plasma samples from patients diagnosed with pulmonary nodule. (**A**) Volcano plot showing differentially expressed proteins in blue (down) or red (up) circles. X-axis shows log2-fold change of plasma proteins between malignant ($N=40$) and benign ($N=40$) nodule patient groups, and y-axis shows log10 of statistical significance values. (**B**) Heat map of 26 differentially expressed proteins between malignant (MA) patients and benign (BE) subjects. Intensities of proteins were log2-transformed. Different color in protein names indicates different biological processes derived from these proteins. (**C**) Principal component analysis of plasma samples from cohort 1 using the plasma proteome expression data. (**D**) Gene Ontology (GO) analysis of differentially expressed proteins between patients and healthy controls

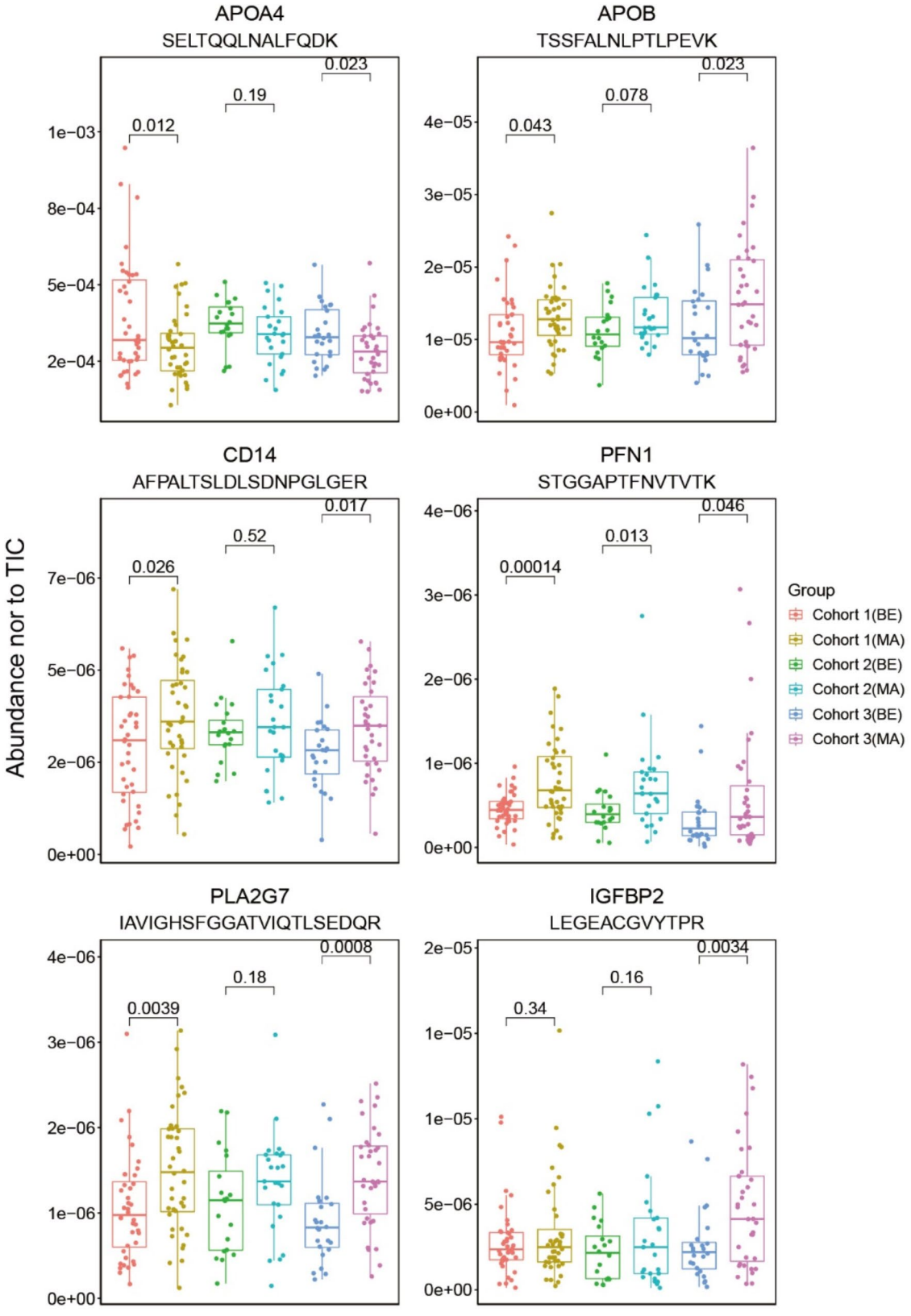

Fig. 3 Peptide intensity plot at the peptide level of feature proteins selected for logistic regression analysis. Boxplot showing differential expression of represented peptides from six proteins between patient groups confirmed as either malignant (MA) or benign (BE) nodules. Cohort 1: discovery stage, cohort 2: validation 1, cohort 3: validation 2. Note that patients in cohort 2 and 3 are from two different hospitals

Table 2 Selected peptide feature sequences

Protein name	Peptide sequence	Precursor charge
APOA4	SELTQQLNALFQDK	2
APOB	TSSFALNLPTLPEVK	3
CD14	AFPALTSLDLSDNPGLGER	3
IGFBP2	LEGEACGVYTPR	2
PFN1	STGGAPTFNVTVTK	2
PLA2G7	IAVIGHSFGGATVIQTLSEDQR	3

lysine-labeled peptides as internal standards combined with external calibration using unlabeled peptides (Table S2). We focused on proteotypic peptides as surrogates of six protein biomarker panel comprised of APOA4, CD14, PFN1, APOB, PLA2G7, and IGFBP2. The peptide sequences and transition ion information were listed in Table S3. Because the peptide representing PFN1 is enriched in serine and threonine residues that could be phosphorylated, we monitored two synthetic peptides using the MRM assay. The peptide STGGAPTFNVTVTK which was identified by DIA showed strong signal (Fig. S3A), whereas an alternative peptide DSPSVWAAVPGK showed poor signal (Fig. S3B). Therefore, we selected the former peptide for quantification. The above six peptides showed superior product ion peaks in our LC-MS/MS system, with the covariance (CV) of seven repeated measurements of a same sample mixture below 10% (Fig. S3C). Among these peptides, APOA4, IGFBP2 and PFN1 showed significantly differential expression (Fig. 6A). A subset of 62 benign and 90 malignant nodule subjects were used as the training cohort, and 27 benign and 54 malignant nodule subjects were used as the testing cohort (Fig. 1). Using concentrations of the six peptides to build a logistic model, the classifier achieved an average AUC of the ROC curve of 0.758 (95% CI 0.683–0.834) in the training dataset and 0.751 (95% CI 0.634–0.868) in the testing dataset (Fig. 6B). Using the logistic regression parameters and threshold value to classify benign from malignant nodules, the biomarker panel showed a sensitivity, specificity, PPV, and NPV of 0.922, 0.419, 0.697 and 0.788 respectively in the training dataset (Fig. 6C, upper panel), and of 0.926,0.370,0.746 and 0.714 respectively in the testing dataset (Fig. 6C, lower panel).

Survival analysis of biomarker panel proteins

We further utilized The Cancer Genome Atlas (TCGA) database, and performed survival analysis to find whether there was any correlation between gene expression of the biomarker proteins and the survival of lung cancer patients. Kaplan-Meier analysis and log-rank test showed that increased expression of PFN1 gene significantly correlated with poor overall survival (OS) and disease-free survival (DFS) (Fig. S4, A-B). In contrast, none of the other five proteins in the six-protein panel showed significance (data not shown). Similarly, APOA1 and ALDOB, two significantly expressed proteins but not in the six-protein panel, also showed no statistical significance (Fig. S4, C-D).

Discussion

Based on plasma protein expression profiles and iterative feature selection, we identified a six-protein panel consisting of APOA4, APOB, CD14, PFN1, PLA2G7, and IGFBP2 to classifying pulmonary nodule patients between benign and malignant nodules. While these proteins have not been directly associated with pulmonary nodule previously, some have been connected to lung cancer. APOA4 was found over expressed in the serum and the respiratory epithelium of atypical adenomatous hyperplasia patients [16]. $CD14^+$ cells play an important role in tumor microenvironment, and a recent study shows that levels of $CD14^+$ cells negatively correlate with overall survival of lung cancer patients [17], and $CD14^+$ macrophages may exert a broader function in other cancers as well. As an actin-binding protein, profilin (PFN1) plays an important role in actin dynamics and has been reported to induce tumor metastasis in non-small cell lung cancer through promoting microvesicle secretion [18]. Regardless, PFN1 were also identified as plasma biomarkers for various disease ranging from infection disease, nervous system disorder to cardiovascular disorders. [19–21]. IGFBP2 has been shown to promote cancer progression in multiple cancers through various signaling pathways downstream of insulin signaling [2–24], and its expression in serum is considered as a biomarker for lung cancer [25, 26]. Thus, these results documented in literature provide additional level of support for the protein markers discovered in our study.

Comparing to known protein biomarkers used in clinic to manage pulmonary nodule patients, the sensitivity, specificity, NPV and AUC of the panel of proteins discovered in our study are generally improved. In a similar study, although protein biomarkers were combined with clinical information to construct an integrated classifier to classify benign and malignant nodules, the AUC of the classifier was 0.76, only slight improvement than the classical model developed by Mayo Clinic using clinical characteristics alone [12]. Our study only utilized protein biomarkers, and reached a satisfactorily AUC of over 0.81 in all three data sets. Nevertheless, the specificity and PPV of this model were not as good as in discovery data set. To rule out the possibility of overfitting due to limited sample size in the discovery data set, we collected two additional sets of independent samples to validate our results. Both validation data sets resulted an AUC of the ROC curve close to that of the discovery data set, which is a better performance comparing to similar studies. Although the calculated NPV was not as optimal as previously reported studies [12, 27], when considering

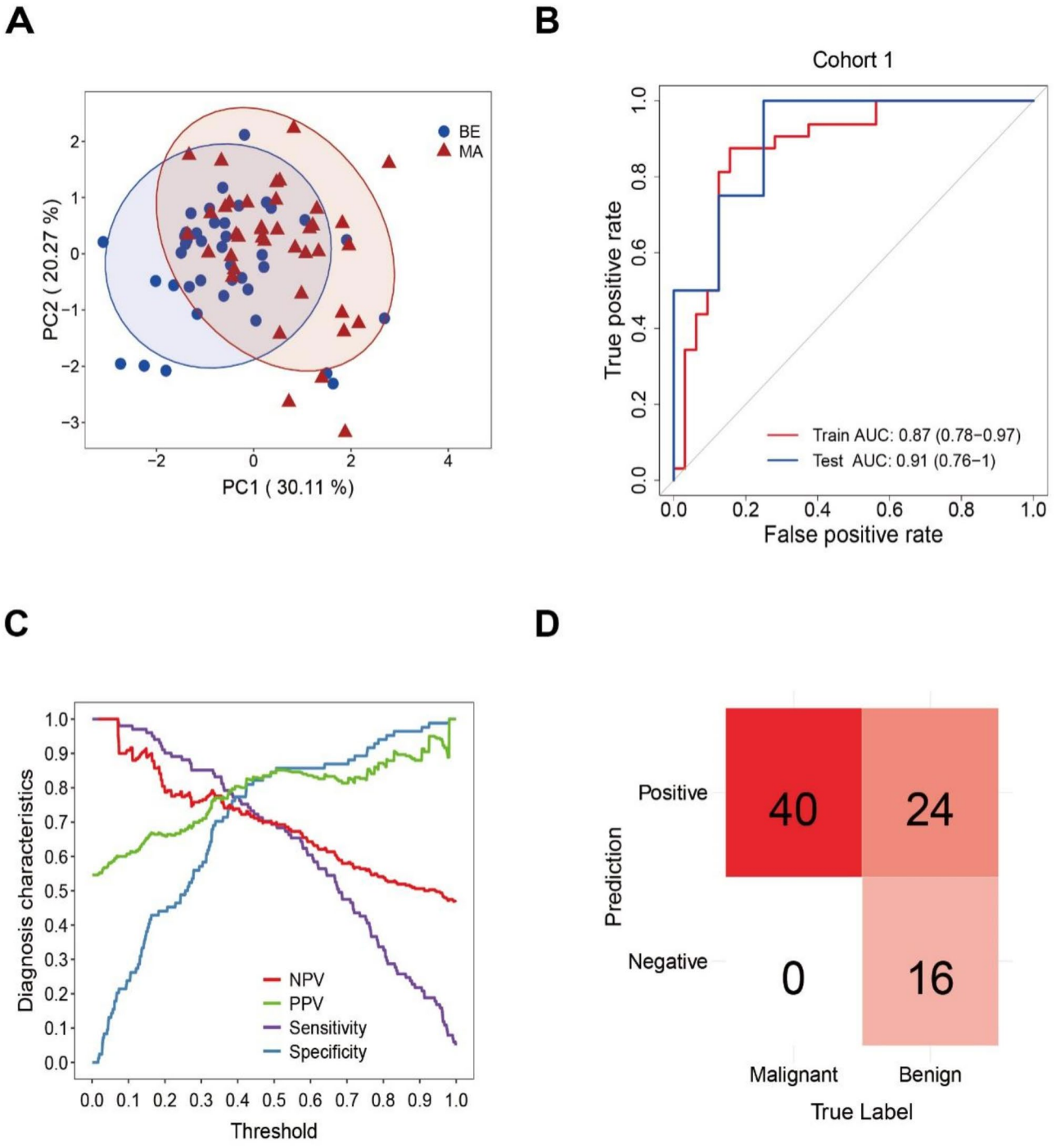

Fig. 4 Logistic regression classification of benign and malignant nodules using cohort 1 as discovery data set. **(A)** Principal component analysis of plasma samples using the 6 candidate proteins (APOA4, APOB, CD14, PFN1, PLA2G7, and IGFBP2). **(B)** ROC curves of a six-protein logistic regression classifier (APOA4, APOB, CD14, PFN1, PLA2G7, and IGFBP2) for distinguishing benign and malignant nodules. **(C)** Sensitivity, specificity, PPN, and NPV value distributions over the range of threshold values from 0 to 1. **(D)** Confusion matrix showing the classification results in cohort 1

the prevalence of pulmonary nodules into the calculation, the NPV can reach as high as nearly 1.0, but at the cost of low PPV (Fig. 4B). Thus, a balance is needed for the test under development to be clinically useful.

Distinguishing benign from malignant small pulmonary nodules with an intermediate risk has been a long-lasting clinical challenge to physicians. A balance needs to be reached between utilizing aggressive diagnostic methods to detect malignant nodule and sparing patients

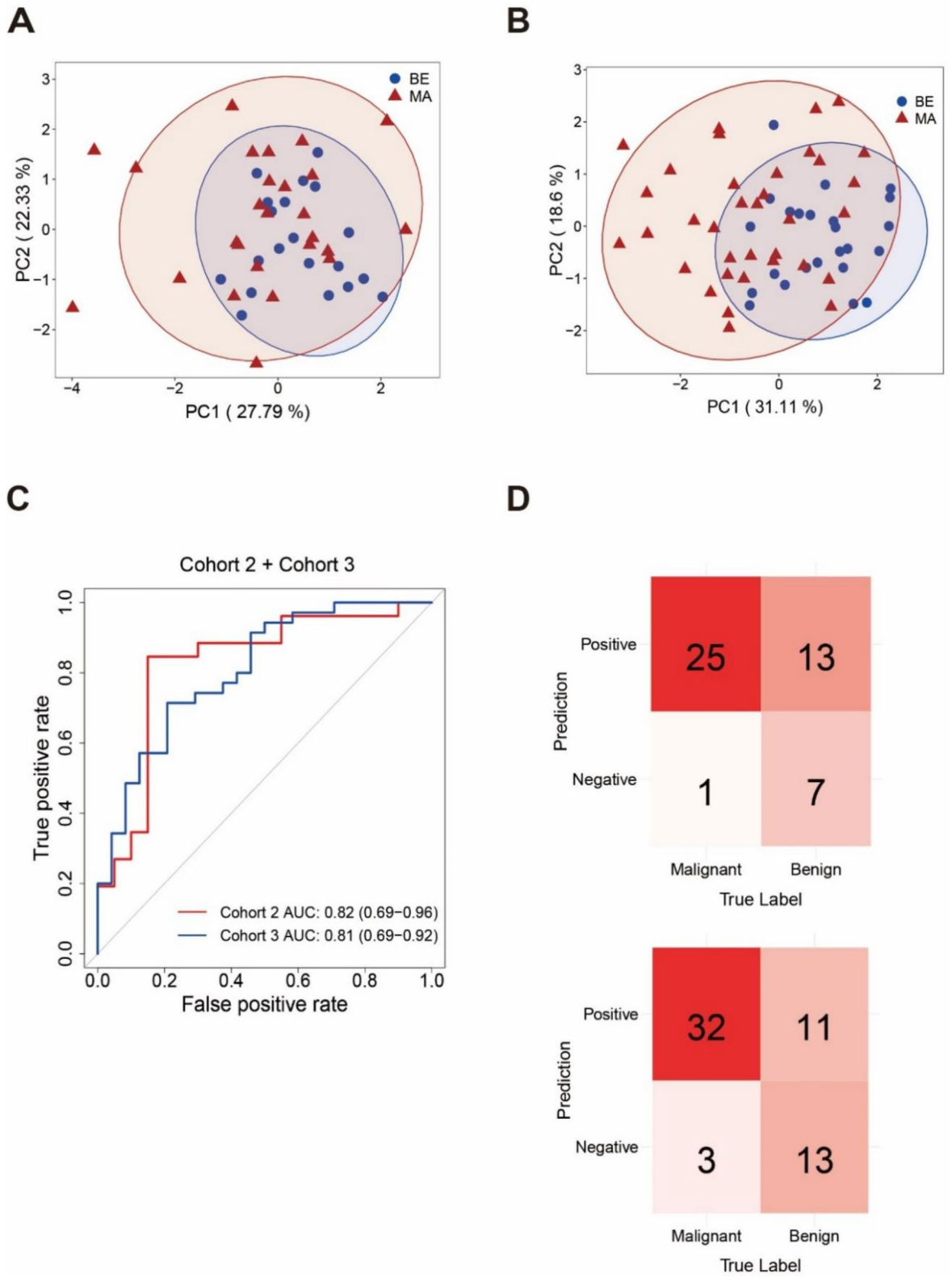

Fig. 5 Logistic regression classification of benign and malignant nodules using cohort 2 and cohort 3 as independent validation data sets. **(A)** Principal component analysis of plasma samples from cohort 2 using the 6 candidate proteins (APOA4, APOB, CD14, PFN1, PLA2G7, and IGFBP2). **(B)** Principal component analysis of plasma samples from cohort 3 using the 6 candidate proteins. **(C)** ROC curves of the six-protein logistic regression classifier for distinguishing benign and malignant nodules in cohort 2 and 3. **(D)** Confusion matrix showing the classification error in cohort 2 and 3

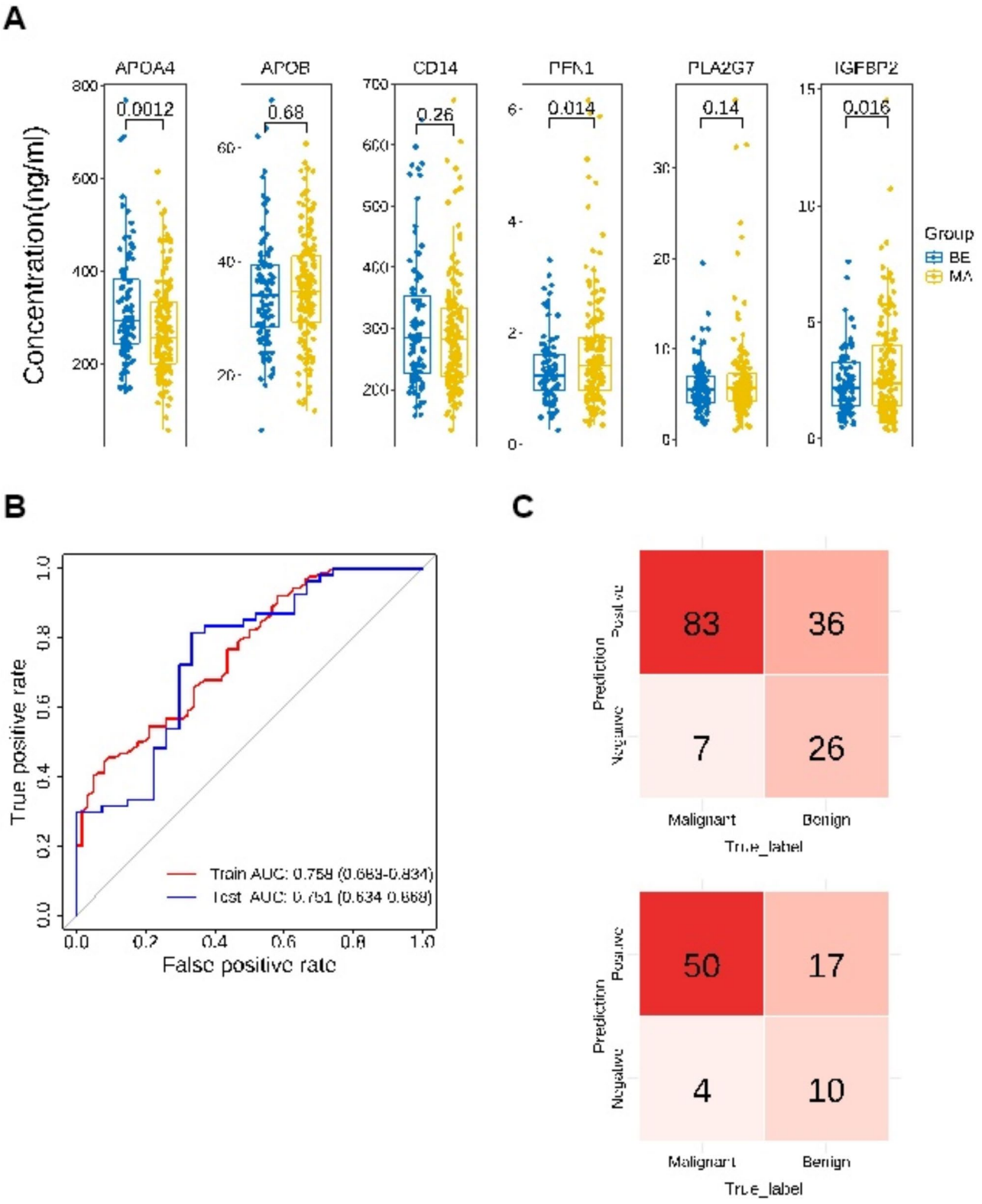

Fig. 6 MRM quantification and logistic regression classification of benign and malignant nodule subjects. **(A)** Peptide intensity plot at the peptide level of feature proteins selected for logistic regression. Boxplot showing differential expression of represented peptides from six proteins between patient groups confirmed as either malignant (MA) or benign (BE) nodules. **(B)** ROC curves of the six-peptide logistic regression classifier to distinguishing benign and malignant nodules. **(C)** Confusion matrix showing the classification results of the MRM assay

from invasive procedures for those with benign nodules. Improving the sensitivity and specificity of a blood-based non-invasive test, when applying to intermediate-risk patients can deliver high NPV, thus prevents patients from having to go through unnecessary invasive procedures. A two-year follow up study of the PANOPTIC test concluded that applying the non-invasive test to nodule patients with a pretest probability of cancer of < 50% resulted in a 40% reduction in invasive procedures accompanied by 3% misclassified malignant nodules [28]. A prospective observational study found that patients with pretest probability of cancer of < 50% tested with the integrated classifier were 74% less likely to undergo an invasive procedure [29]. Because our study is at an early validation phase, large number of subjects are needed for further study to assess the clinical utility. Based on our improved sensitivity and specificity, we predict that our 6-protein panel should perform in par with the clinical utility offered from the PANOPTIC method.

The strength of this study lies in that two independent data sets were used to validate the initial study, and that patients from two different hospitals were recruited. Future development of the test can focus on dramatically improved NPV, which can divert nearly 60% of patients with benign nodule to CT surveillance. Because the test is non-invasive, thus avoiding invasive biopsy procedures and reducing patient burdens. The limitations of the study are the relatively small sample size. Further refinement and validation of this panel of biomarker proteins with increased sample size is warranted.

Conclusion

Our study provides retrospective plasma proteomic profiles of patients diagnosed with benign and malignant pulmonary nodules. Our results provide new panel of protein biomarkers for distinguishing benign and malignant pulmonary nodules that worth further development into clinically useful blood tests.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12014-025-09532-w.

Supplementary Material 1: Figure S1. Assessment of the quantitative proteomic data quality. A Correlation analysis of the quality control (QC) samples. B Distribution of protein abundance of all quantified proteins. C Distribution of number of quantified proteins in each sample in cohort 1

Supplementary Material 2: Figure S2. Boxplot of abundance values at the protein level of feature peptides selected for logistic regression analysis in cohort 1. Boxplot showing differential expression of six proteins between patient groups confirmed as either malignant (MA) or benign (BE) nodules. Protein expression data from three cohorts of samples are shown

Supplementary Material 3: Figure S3. Quality of the MRM data. (A-B) Extracted ion chromatograms of peptides STGGAPTFNVTVTK (A) and DSPSVWAAVPGK of PFN1 (B), along with their heavy peptide internal standards in two random testing samples. (C) Variable coefficient (CV) values in QC samples for the 6-peptide biomarkers.

Supplementary Material 4: Figure S4. Survival analysis of feature proteins used for logistic regression. (A) Overall survival and (B) disease-free survival of nodule patients stratified by expression level of PFN1. (C) Overall survival of nodule patients stratified by expression level of APOA1. (D) Overall survival of nodule patients stratified by expression level of ALDOB

Supplementary Material 5: Table S1: 80PNp_DIA_Protein raw data. Proteomic quantification of the plasma proteins in cohort 1 using DIA

Supplementary Material 6: Table S2: 233PNp_6Peptide_MRM. Plasma concentration of the six selected biomarker proteins using targeted MRM method. The unit of concentration is ng/ml

Supplementary Material 7: Table S3: MRM transition information. The information of transition ions and ion source parameter values of the MRM assay

Acknowledgements
The authors thank all individuals who participated in this study.

Author contributions
B.J., T.W., F.G., L.L., and J.D. conceived the idea and directed the study. T.W. and X.D. performed the experiments and analyzed the mass spectrometry data. L.P. performed the statistical analysis. J.Y. and B.G. contributed to some experimental data. B.J., J.D. and B.G. contributed to patient diagnosis and clinical sample collection. L.L. wrote the manuscript.

Funding
This work was supported by the Key Research and Development Project (No. 222102310069).

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
All the studies mentioned in this article were approved by the Ethics Committee, and written informed consent was obtained by all the participant.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Oncology, the First Affiliated Hospital of Zhengzhou University, Zhengzhou 450052, Henan, China
[2]Durbrain Medical Laboratory, Hangzhou 310000, Zhejiang, China
[3]Department of Clinical Laboratory, Chongqing University Cancer Hospital, Chongqing 400030, China
[4]Department of Radiology, the First Affiliated Hospital of Zhengzhou University, No.1 East Jianshe Road, Zhengzhou 450052, Henan, China

Received: 13 June 2024 / Accepted: 27 February 2025
Published online: 07 April 2025

References

1. Mazzone PJ, Lam L. Evaluating the patient with a pulmonary nodule: A review. JAMA. 2022;327(3):264–73. https://doi.org/10.1001/jama.2021.24287.
2. Oudkerk M, Liu S, Heuvelmans MA, Walter JE, Field JK. Lung cancer LDCT screening and mortality reduction - evidence, pitfalls and future perspectives. Nat Rev Clin Oncol. 2021;18(3):135–51. https://doi.org/10.1038/s41571-020-00432-6.
3. Pinsky PF, Gierada DS, Black W, Munden R, Nath H, Aberle D, Kazerooni E. Performance of lung-RADS in the National lung screening trial: a retrospective

assessment. Ann Intern Med. 2015;162(7):485–91. https://doi.org/10.7326/m14-2086.
4. Zeng D, Wang C, Mu C, Su M, Mao J, Huang J, Jiang J. Cell-free DNA from Bronchoalveolar lavage fluid (BALF): a new liquid biopsy medium for identifying lung cancer. Ann Transl Med. 2021;9(13):1080. https://doi.org/10.21037/atm-21-2579.
5. Zhang X, Yu Z, Xu Y, Chao Y, Hu Q, Li C, Zhang X. Utility of cell-free DNA from bronchial washing fluid in diagnosis and genomic determination for radiology-suspected pulmonary nodules. Br J Cancer. 2022;127(12):2154–65. https://doi.org/10.1038/s41416-022-01969-2.
6. Li H, Li M, Guo H, Lin G, Huang Q, Qiu M. Integrative analyses of Circulating mRNA and LncRNA expression profile in plasma of lung Cancer patients. Front Oncol. 2022;12:843054. https://doi.org/10.3389/fonc.2022.843054.
7. Ma S, Yu X, Jin X, Qiu F, Chen X, Wang R, Cao C. The usefulness of liquid-based cytology of Bronchoalveolar lavage fluid combined with bronchial brush specimens in lung cancer diagnosis. J Int Med Res. 2022;50(11):3000605221132708. https://doi.org/10.1177/03000605221132708.
8. Liang W, Chen Z, Li C, Liu J, Tao J, Liu X, He J. Accurate diagnosis of pulmonary nodules using a noninvasive DNA methylation test. J Clin Invest. 2021;131(10). https://doi.org/10.1172/jci145973.
9. Schotten LM, Darwiche K, Seweryn M, Yildiz V, Kneuertz PJ, Eberhardt WEE, Oezkan F. DNA methylation of PTGER4 in peripheral blood plasma helps to distinguish between lung cancer, benign pulmonary nodules and chronic obstructive pulmonary disease patients. Eur J Cancer. 2021;147:142–50. https://doi.org/10.1016/j.ejca.2021.01.032.
10. Tian T, Lu J, Zhao W, Wang Z, Xu H, Ding Y, Shen H. Associations of systemic inflammation markers with identification of pulmonary nodule and incident lung cancer in Chinese population. Cancer Med. 2022;11(12):2482–91. https://doi.org/10.1002/cam4.4606.
11. Li XJ, Hayward C, Fong PY, Dominguez M, Hunsucker SW, Lee LW, Kearney P. A blood-based proteomic classifier for the molecular characterization of pulmonary nodules. Sci Transl Med. 2013;5(207):207ra142. https://doi.org/10.1126/scitranslmed.3007013.
12. Silvestri GA, Tanner NT, Kearney P, Vachani A, Massion PP, Porter A, Mazzone PJ. Assessment of plasma proteomics biomarker's ability to distinguish benign from malignant lung nodules: results of the PANOPTIC (Pulmonary nodule plasma proteomic Classifier) trial. Chest. 2018;154(3):491–500. https://doi.org/10.1016/j.chest.2018.02.012.
13. Vachani A, Pass HI, Rom WN, Midthun DE, Edell ES, Laviolette M, Massion PP. Validation of a multiprotein plasma classifier to identify benign lung nodules. J Thorac Oncol. 2015;10(4):629–37. https://doi.org/10.1097/jto.0000000000000447.
14. Tanner NT, Springmeyer SC, Porter A, Jett JR, Mazzone P, Vachani A, Silvestri GA. Assessment of integrated classifier's ability to distinguish benign from malignant lung nodules: extended analyses and 2-Year Follow-Up results of the PANOPTIC (Pulmonary nodule plasma proteomic Classifier) trial. Chest. 2021;159(3):1283–7. https://doi.org/10.1016/j.chest.2020.10.069.
15. Goel MK, Khanna P, Kishore J. Understanding survival analysis: Kaplan-Meier estimate. Int J Ayurveda Res. 2010;1(4):274–8. https://doi.org/10.4103/0974-7788.76794.
16. Borlak J, Länger F, Chatterji B. Serum proteome mapping of EGF Transgenic mice reveal mechanistic biomarkers of lung cancer precursor lesions with clinical significance for human adenocarcinomas. Biochim Biophys Acta Mol Basis Dis. 2018;1864(10):3122–44. https://doi.org/10.1016/j.bbadis.2018.06.019.
17. Schenk EL, Boland JM, Withers SG, Bulur PA, Dietz AB. Tumor microenvironment CD14(+) cells correlate with poor overall survival in patients with Early-Stage lung adenocarcinoma. Cancers (Basel). 2022;14(18). https://doi.org/10.3390/cancers14184501.
18. Wang Y, Lu Y, Wan R, Wang Y, Zhang C, Li M, Hu C. Profilin 1 induces tumor metastasis by promoting microvesicle secretion through the ROCK 1/p-MLC pathway in Non-Small cell lung Cancer. Front Pharmacol. 2022;13:890891. https://doi.org/10.3389/fphar.2022.890891.
19. Che Y, Ren J, Zhao H, Yang Y, Chen Z. Orosomucoid 2 as a biomarker of carotid artery atherosclerosis plaque vulnerability through its generation of reactive oxygen species and lipid accumulation in vascular smooth muscle cells. Biochem Biophys Res Commun. 2024;705:149736. https://doi.org/10.1016/j.bbrc.2024.149736.
20. Fish-Low CY, Than LTL, Ling KH, Sekawi Z. The potential of eight plasma proteins as biomarkers in redefining leptospirosis diagnosis. J Proteome Res. 2024;23(9):4027–42. https://doi.org/10.1021/acs.jproteome.4c00376.
21. Shen HP, Dong X, Li ZB, Wu JZ, Zheng CM, Hu XJ, Li JC. Protein profiles and novel molecular biomarkers of schizophrenia based on 4D-DIA proteomics. J Proteome Res. 2024;23(7):2376–85. https://doi.org/10.1021/acs.jproteome.4c00040.
22. Liu D, Xia AD, Wu LP, Li S, Zhang K, Chen D. IGF2BP2 promotes gastric cancer progression by regulating the IGF1R-RhoA-ROCK signaling pathway. Cell Signal. 2022;94:110313. https://doi.org/10.1016/j.cellsig.2022.110313.
23. Lu H, Wang L, Gao W, Meng J, Dai B, Wu S, Fang B. IGFBP2/FAK pathway is causally associated with dasatinib resistance in non-small cell lung cancer cells. Mol Cancer Ther. 2013;12(12):2864–73. https://doi.org/10.1158/1535-7163.Mct-13-0233.
24. Yao X, Wang Y, Duan Y, Zhang Q, Li P, Jin R, Zhou X. IGFBP2 promotes salivary adenoid cystic carcinoma metastasis by activating the NF-κB/ZEB1 signaling pathway. Cancer Lett. 2018;432:38–46. https://doi.org/10.1016/j.canlet.2018.06.008.
25. Dong J, Zeng Y, Zhang P, Li C, Chen Y, Li Y, Wang K. Serum IGFBP2 level is a new candidate biomarker of severe malnutrition in advanced lung Cancer. Nutr Cancer. 2020;72(5):858–63. https://doi.org/10.1080/01635581.2019.1656755.
26. Guo C, Lu H, Gao W, Wang L, Lu K, Wu S, Fang B. Insulin-like growth factor binding protein-2 level is increased in blood of lung cancer patients and associated with poor survival. PLoS ONE. 2013;8(9):e74973. https://doi.org/10.1371/journal.pone.0074973.
27. Vachani A, Hammoud Z, Springmeyer S, Cohen N, Nguyen D, Williamson C, Kearney P. Clinical utility of a plasma protein classifier for indeterminate lung nodules. Lung. 2015;193(6):1023–7. https://doi.org/10.1007/s00408-015-9800-0.
28. Kheir F, Uribe JP, Cedeno J, Munera G, Patel H, Abdelghani R, Majid A. Impact of an integrated classifier using biomarkers, clinical and imaging factors on clinical decisions making for lung nodules. J Thorac Dis. 2023;15(7):3557–67. https://doi.org/10.21037/jtd-23-42.
29. Pritchett MA, Sigal B, Bowling MR, Kurman JS, Pitcher T, Springmeyer SC. Assessing a biomarker's ability to reduce invasive procedures in patients with benign lung nodules: results from the ORACLE study. PLoS ONE. 2023;18(7):e0287409. https://doi.org/10.1371/journal.pone.0287409.

Publisher's note